

UC Davis

UC Davis Previously Published Works

Title

Variation in Phenotype Definitions in Observational Clinical Research: A Review of Three Conditions.

Permalink

<https://escholarship.org/uc/item/5f61179s>

Journal

AMIA Annual Symposium Proceedings, 2024

ISSN

1531-605X

Authors

Shoaibi, Azza

Ostropolets, Anna

Weaver, James

et al.

Publication Date

2024

Peer reviewed

Variation in Phenotype Definitions in Observational Clinical Research: A Review of Three Conditions

Azza Shoaibi¹, Anna Ostropelets¹, James Weaver¹, Gowtham A. Rao¹, Asieh Golozar², Joel Swerdel¹, Vojtech Huser², Atif Adam³, Thamir Alshammari⁴, Nnabuchi Anikpezie⁵, William A Baumgartner Jr⁶, Milou Brand⁷, Ruochong Fan⁸, Andrew S. Kanter⁹, Dave Kern¹, Septi Melisa¹⁰, Evan Minty¹¹, Jessica Mo¹², Bolu Oluwalade¹³, Lisa M. Schilling⁶, Hayden Spence¹⁴, Jacqueline C. Stocking¹⁵, Linying Zhang⁸, Patrick Ryan^{1,9}

1 Janssen Research & Development, Titusville, NJ, USA; **2** Odysseus Data Services, Cambridge, MA, USA; **3** IQVIA Ltd., Boston, MA, USA; **4** King Saud University, Riyadh, Saudi Arabia; **5** Department of Population Health, University of Mississippi Medical Center, Jackson, MS, USA; **6** University of Colorado School of Medicine, Aurora, CO, USA; **7** Real World Solutions, IQVIA, Cambridge, MA, USA; **8** Washington University in St. Louis, St. Louis, MO, USA; **9** Department of Biomedical Informatics, Columbia University, New York, USA; **10** International PhD Program in Biotech and Healthcare Management, Taipei Medical University, Taipei, Taiwan; **11** O'Brien Institute for Public Health, University of Calgary, Canada; **12** Johns Hopkins University School of Medicine, Baltimore, MD; **13** The Children's Hospital of Philadelphia, PA, USA; **14** Aptive Resources LLC, Alexandria, VA, USA; **15** Department of Internal Medicine, Davis School of Medicine, University of California, California, USA

Abstract

Our study examined the heterogeneity of phenotype algorithms (PA) in literature on Alzheimer's disease (AD), major depressive disorder (MDD), and pulmonary arterial hypertension (PAH), focusing on the impact of PA differences on patient overlap and incidence rate variability across conditions in six observational databases. We reviewed 49 replicated PAs (13 for AD, 23 for MDD, and 13 for PAH) and found significant heterogeneity. These varied PAs identified distinct patient cohorts, resulting in significant incidence rate heterogeneity. Despite some papers reporting primary condition codes and inclusion. comprehensive documentation ensuring reproducibility was often lacking, underscoring the need for more transparent and robust research practices.

Introduction

The choice of phenotype algorithms (PA) to identify patients of interest for a study can affect the findings of an observational study. Although there is consensus among researchers of the need to minimize measurement error in phenotyping¹, there are discrepancies in how researchers describe health events across studies, what approaches they use to develop and evaluate these PAs, and how they report PAs in publications^{2, 3}.

Justifying the differences between algorithms for the same phenotype is challenging. An independent reviewer may find published PAs to be subjective or ambiguous even when code lists are provided. This subjectivity may be intrinsic because of research intent ambiguity (i.e., incidence reporting vs. clinical decision support) or it may be contextual (i.e., a specialist in a field may have context-specific knowledge about a clinical definition that is neither accessible nor apparent to the reviewer). Further, PA discrepancies may arise because of interpretation subjectivity of other researchers who may not be privy to the primary intent.

The objective of this paper is to investigate PA heterogeneity in published literature on three conditions: Alzheimer's disease (AD), major depressive disorder (MDD), and pulmonary arterial hypertension (PAH). Our review focused on

assessing the completeness and clarity with which PAs were reported in recent publications. Additionally, we quantified the impact of PA differences on 1) patient overlap across PAs for the same condition and 2) incidence rate heterogeneity across PAs for the same condition. These analyses were conducted over six observational data sources standardized to a common data model (CDM).

Methods

Each February, the Observational Health Data Sciences and Informatics (OHDSI) community hosts 'Phenotype Phebruary', a month-long collaborative effort focused on advancing the science of phenotype development and evaluation. In 2024, we employed a community-developed, standardized software-enabled workflow to evaluate and compare published PAs on AD, MDD and PAH.

PA identification

For study identification, we selected observational studies reporting PAs for any of the three conditions from a standardized PubMed search (see <https://forums.ohdsi.org/t/ohdsi-phenotype-phebruary-2024/20940/6>). Studies published in the last 4 years (2020-2024) using retrospective data from claims and electronic health records (EHR) were considered. We excluded publications that did not have free full text access, did not report diagnostic codes, or were part of a clinical trial, systematic review, or disease-specific registry study. We also excluded probabilistic PAs. In the second step, we abstracted PA information using a standardized template to obtain all variables related to the study setting and PA specification (author, study citation, year, country, database, code sets, inclusion criteria with relevant code set, other requirements, exclusion criteria with relevant code set, age restrictions, patient characteristics, validation metrics if available).

PA replication

In the replication step, we used ATLAS, a rule-based phenotype algorithm builder, to replicate PAs with terminologies, temporal relationships, and logic^{4,5}. At least two investigators led and confirmed accuracy of each PA replication, and all implementations were publicly reviewed by OHDSI collaborators at virtual meetings.

PA transparency reporting

we performed a content review of the publications to evaluate transparency in PA reporting. We graded the descriptions for transparency based on the following items: 1) Presence of a clinical (case) definition of the condition of interest; 2) Presence of a description of PA entry event criteria with any exclusion/inclusion criteria, including temporal logic; 3) Presence of one or more code lists for use in entry/inclusion/exclusion criteria; 4) Presence of any validation study metrics (i.e., PPV, NPV, sensitivity, specificity) for the PA; 5) the PA represented in any computer language (executable code).

PA characterization

We generated and visualized comprehensive, population-level descriptive summary statistics for each PA instantiation across the six observational data sources using the CohortDiagnostics^{6,7} (CD) R package.

Data sources

We analyzed four US administrative claims databases, one Japanese administrative claims database, one US electronic health record database, and one German primary care database (See Table 1), representing routinely collected clinical data stored in a network of databases standardized in structure and content to the Observational Medical Outcomes Partnership CDM⁸. This standardization allowed us to conduct a distributed analysis across the data sources with strictly consistent analytic code without the need for merging or sharing patient-level data, thus preserving patient privacy compliant with data governance regulations.

Table 1. Data sources

Source name	Abbreviation	Country	Type	Content
Optum® de-identified Clinformatics® Data Mart	Clinformatics®	US	Claims	Private payer, primarily <65 years
Merative™ MarketScan® Commercial Database	CCAE	US	Claims	Employer-based private payer; primarily <65 years
Merative™ MarketScan® Medicare Supplemental Beneficiaries	MDCR	US	Claims	Retirees; primarily >65 years
Merative™ MarketScan® Multi-state Medicaid	MDCD	US	Claims	Medicaid enrollees; all ages
IQVIA Disease Analyzer Germany	DA Germany	Germany	EHR	Primary ambulatory care; all ages
Japan Medical Data Center	JMDC	Japan	Claims	Private payer; 18–65 years
Optum® de-identified Electronic Health Record Dataset	Optum® EHR	US	EHR	US health systems; all ages

Heterogeneity assessment

To estimate the heterogeneity across PAs for each condition of interest, we reported the following metrics for each database: 1) PA pairwise patient overlap as a measure of PA consistency. This is defined as the proportion of patients identified by both PAs in pair, compared to total number identified by either PA. 2) Background incidence rate (IR) of the conditions of interest among the database at-risk population, defined as patients observed on January 1st of 2017, 2018 or 2019 (index dates) provided they had at least one-year prior observation and no history of the clinical condition. We defined time-at-risk for observing incident events as time from index date until the earliest of the event of interest or end of database observation. The IR was calculated by dividing the number of incident events observed during the at-risk period by total person-time at-risk. We reported background IRs as new events per 1000 person-years; stratified by age (all ages, <5, 5-17, 18-34, 35-54, 55-64, 65-74, 75-84 >=85 years), sex, and age-by-sex. 3) We used the background rate ratio (RR) to estimate variation. We defined background RR as the highest IR divided by the lowest IR for each condition-specific set of PAs within a data source.

Overall, the methods used in this study were based on open-source tools to facilitate collaboration among OHDSI researchers across a network of data sources without sharing patient-level data, while adhering to standardized analytics. Specifically, CohortDiagnostics is an open-source R package that generates group-level summary statistics for any cohort created using the OMOP Common Data Model.

Results

For AD, our literature search returned 13 studies from which we identified 7 PAs for AD and 6 for Alzheimer's disease and related disorders (ADRD). For MDD, our search returned 24 studies from which we identified 23 PAs for replication. One study with an MDD PA did not include a code set and was unable to be replicated). Finally, our search returned 14 PAH studies from which we identified 13 PAs for replication. One study defined PA using a machine learning algorithm that consisted of data elements unavailable in our data sources. Comprehensive characterization results for cohorts instantiated from all PAs across all data sources are reported in the following condition-specific web-based applications:

- http://results.ohdsi.org/app/16_PhenotypePhebAlzh
- https://results.ohdsi.org/app/17_PhenotypePhebMdd
- http://results.ohdsi.org/app/18_PhenotypePhebPah

In addition to characterization results, the above applications also include overlap results. The full citations for all papers reviewed are available at: Full list of reviewed studies by condition.xlsx.

PA transparency reporting

Across all 51 manuscripts reviewed, none reported all 5 PA transparency items. In our review of AD and MDD, studies rarely stated a clinical case definition that their PA aimed to represent. Specifically, only one AD/ADRD study

and six MDD studies provided or referenced a clear clinical case definition. This lack of clarity persisted regardless of whether the condition was the focus of the study population or an outcome of interest. The absence of explicit case definitions contributed to confusion. For instance, our analysis revealed that the term "Alzheimer's" was used in two distinct ways: some studies used it within the broader category of ADRD, which includes various forms of dementia, while others used a narrower definition that excludes conditions like vascular or unspecified dementia. However, this distinction was often implied rather than explicitly stated. Conversely, nearly all the studies on pulmonary arterial hypertension (PAH) provided a clear clinical description of PAH.

The presence of the other four transparency items was generally consistent across the three conditions. A considerable proportion of the studies reported the entry criteria, albeit implicitly, alongside a list of exclusion criteria. Additionally, most of the manuscripts reported the diagnosis code list used to define the condition of interest (ICD-10/9 codes), which in most instances were the implicit entry criterion. However, very few studies reported the codes used to define other commonly used inclusion/exclusion criteria such as drug exposures, procedure occurrences, or provider specialty. Although the criteria for including or excluding patients were frequently reported, descriptions often lacked clear temporal logic, making the PA difficult to interpret or reproduce. Only 14 studies reported validation metrics associated with their PA. While several studies mentioned previous research that had used or developed the PA applied in their work, PA modification was common, and hypotheses or conjecture on how such changes might influence PA performance was seldom provided.

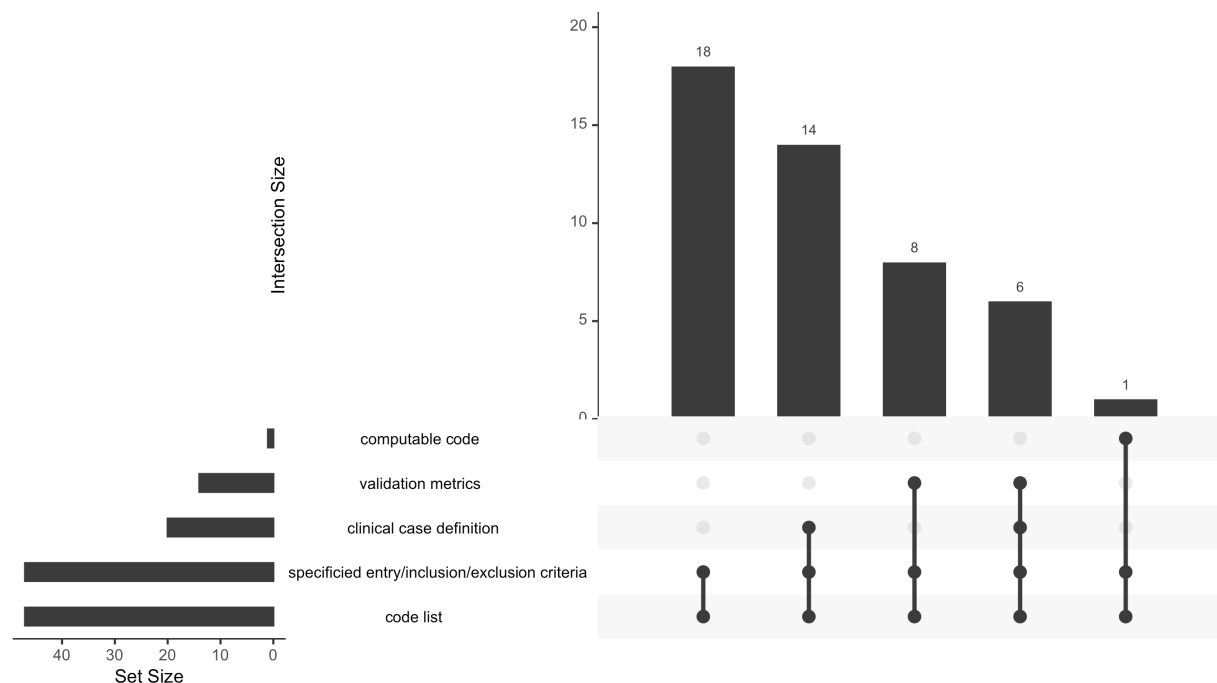


Figure 1. Upset plot visualizing the number of papers with intersecting transparency reporting items.

PA overlap

Among the seven PA identified for AD, three shared identical criteria, while the remaining exhibited variations in either the logic, the codes used, or both. Regarding ADRD, two out of the six identified PA were identical. For MDD, three out of the 23 PA shared identical parameters. Finally, two out of the ten PA identified for PAH were found to be identical.

Figure 2 illustrates the pairwise overlap observed across the five distinct AD PAs in Clinformatics®. Every bar represents the total patients identified by two (either) PAs (the one indicated across columns and the one indicated across rows)- labeled from A to E. The dark purple section of the bar represents the proportion of patients identified by the PA indicated on the column heading only, the light purple indicates the proportion of patients identified by the

PA indicated on the row only and the gray section represent the proportion captured by both PAs. The larger the gray section, the larger the overlap between two given PAs. For example, in Figure 2 below, PA C (AD per CCW 30) and PA B (AD per Chen) had the largest overlap with roughly 82% of patients identified by either algorithm were identified by both PAs; only 18% were identified by PA C only and no subjects were identified by PA B only. Conversely, when reviewing overlap between E and A, only 17% of the patients were identified by both PAs and the majority (83%) were identified by A only. As illustrated in figure 2, there is substantial overlap variation across pairwise comparisons, ranging from 0.1 to 0.85 among AD PAs for example. We observed a similar pattern in all databases and for ADRD, MDD and PAH. In most pairwise comparisons, certain patients were identified by one PA but not the other, indicating limited overlap that implies that each PA captures distinct individuals. In a few instances, a noticeable pattern emerged where one PA predominantly identified patients also identified by other PAs.

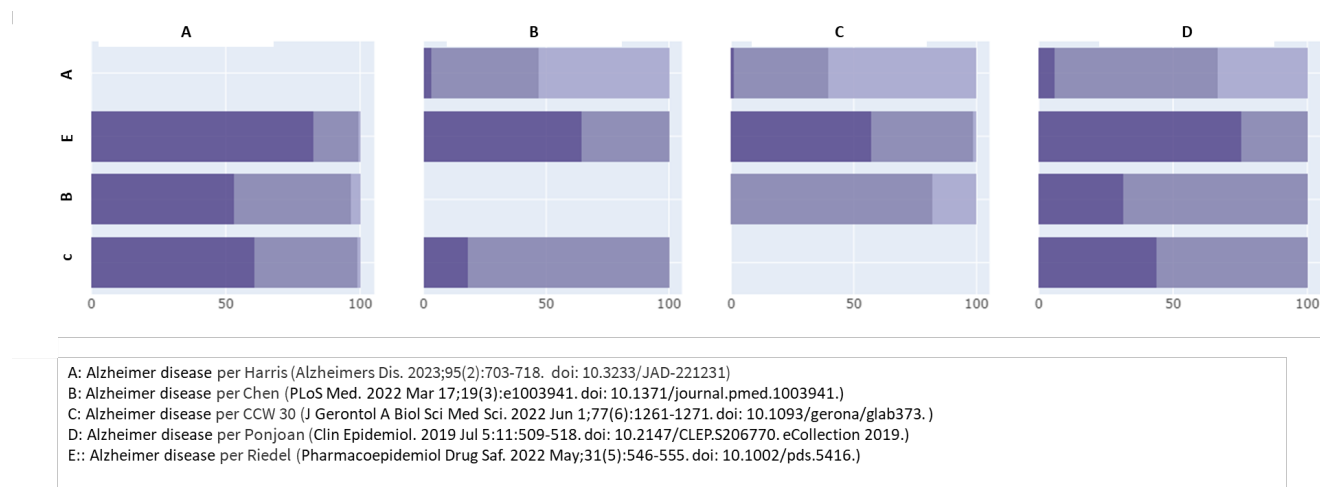


Figure 2. Pairwise cohort overlap of 5 selected PAs for Alzheimer's disease. The dark and light purple show the subjects identified by only one PA in each pair and the mid-intensity grey shows the patients identified by both PAs.

Incidence rates

We observed substantial IR heterogeneity across PAs within each condition. Table 2 reports PAH IRs among patients aged 55-64 years using different PAs across databases. Table 2 also is a heatmap displaying IR variation PA (rows) and by database (columns).

Table 2. Incidence rates per 1000 person-years of selected PAH PAs among persons aged 55-64 years by database. The rate ratio compares the largest to lowest IR across data sources (row) and across PAs (column)

Phenotype Algorithm	JMDC	Optum® EHR	Clinformatics®	CCAE	MDCD	Rate Ratio
PAH per Dafour et al	0.0193	0.3071	0.2658	0.1668	0.1411	15.91
PAH per Anand et al	0.0176	0.0378	0.1186	0.0046	0.0057	25.78
PAH per Didden et al	0.1073	0.2059	0.697	0.4116	1.1639	10.85
PAH per Karnes et al	0.0176	0.0378	0.1186	0.0046	0.0057	25.78
PAH per Sargent T et al	0.0149	0.1855	0.2197	0.1448	0.1042	14.74
PAH per Shah et al	0.0062	0.083	0.0764	0.0539	0.0426	13.39
PAH per Tsang Y et al	0.0114	0.1668	0.1707	0.1132	0.0833	14.97
PAH per Vasudeva et al	0.1258	0.251	0.7859	0.4618	1.2653	10.06
PAH per Schikowski et al	0.0158	0.0585	0.0764	0.0391	0.0577	4.84
PAH per Gauthier-Loiselle et al	0.0114	0.0638	0.0798	0.0352	0.0757	7.00
Rate Ratio	20.29	8.12	10.29	100.39	221.98	

Key: JMDC: Japan Medical Data Center; Optum® EHR: Optum® de-identified Electronic Health Record Dataset; Clinformatics®: Optum® de-identified Clinformatics® Data Mart; CCAE: Merative™ MarketScan® Commercial Database; MDCD: Merative™ MarketScan® Multi-state Medicaid

Overall, the variation of the IR was substantial, for each PA across databases and within each database across PAs. In some cases, *the IR heterogeneity of PAs within a database exceeded the IR heterogeneity of a single PA across databases*. This is apparent by the rate ratios by data source (columns), where we observed extreme values. For example, in CCAE the relative difference between PAs with highest and lowest IR was two orders of magnitude (RR=100.39). This trend was consistent across conditions and age groups. Table 3 displays the IR heterogeneity across PAs by database and condition. Each cell reports the RR for each set of condition-specific PAs within a database. Additionally, the IR range within each database is provided. The findings indicate that IR heterogeneity was modest for (ADRD) yet substantial for MDD and PAH.

Table 3: PA IR heterogeneity by database and condition measured by background RR.

Database	ADRD	AD	MDD	PAH
DA Germany	3.9 (0.747-2.897)	4.7 (0.182-0.859)	54.7 (0.222-12.148)	102.6 (0.001-0.082)
JMDC	1.6 (0.324-0.520)	19.9 (0.007-0.147)	401.3 (0.024-9.551)	9.8 (0.006-0.055)
Optum® EHR	1.3 (2.624-3.513)	8.1 (0.300-2.420)	42.9 (0.573-24.587)	8.2 (0.024-0.192)
Clinformatics®	1.5 (6.592-10.041)	7.5 (0.834-6.226)	21.4 (1.907-40.854)	21.3 (0.041-0.877)
CCAE	2.3 (0.338-0.772)	11.5 (0.022-0.253)	24.2 (1.403-33.918)	82.2 (0.002-0.164)
MDCD	1.8 (2.308-4.182)	12.8 (0.169-2.167)	33.7 (1.139-38.351)	121.7 (0.003-0.377)
MDCR	1.5 (18.195-27.469)	5.6 (2.971-16.588)	71.0 (0.580-41.233)	340.2 (0.006-2.211)

Key - ADRD: Alzheimer's disease and related disorders, AD: Alzheimer's disease, MDD: major depressive disorder, PAH: pulmonary arterial hypertension, JMDC: Japan Medical Data Center; Optum® EHR: Optum® de-identified Electronic Health Record Dataset; Clinformatics®: Optum® de-identified Clinformatics® Data Mart; CCAE: Merative™ MarketScan® Commercial Database; MDCD: Merative™ MarketScan® Multi-state Medicaid

Discussion and Conclusions

We examined 51 recent observational studies that included 49 PAs that we were able to replicate for three conditions. We evaluated PA heterogeneity in terms of reporting transparency, impact on patient identification overlap, and impact on background incidence. We instantiated the PAs across a network of six observational healthcare databases and found that heterogeneous PAs for the same condition identified distinct cohorts and produced considerable background IR heterogeneity. In some cases, the IR heterogeneity of PAs within a database exceeded the IR heterogeneity of a single PA across databases. Further, our analysis indicated substantial heterogeneity in PAs across the three conditions, even among studies using identical databases to address similar research questions.

While multiple PAs were defined for each of the three conditions, we observed some within-condition consistency possibly attributable to similar intent by the authors. Overall, AD and ADRD heterogeneity was not as drastic as that observed in MDD and PAH. Some heterogeneity was expected because clinical descriptions of many medical conditions evolve over time⁹. Further, medical terminologies may incorporate clinical description evolution with significant delay. This delay is a result of a careful balancing between terminology stability (a pragmatic necessity for successful clinical adoption) and timely response to the evolving pathophysiological science of a given condition. It is a challenge for terminology standards development organizations to judge exactly when a new consensus has been reached in assessing disease or symptom boundaries. In fact, given underlying terminology evolution, we should expect published PAs to respond accordingly by leveraging this evolution. To illustrate the point, the ability to reliably study some conditions was vastly enhanced by the adoption of new versions of ICD terminology.

Sources of inconsistency across PAs were multifaceted. Across the three conditions, code list and temporal logic variation was evident. Additionally, code choice discrepancies sometimes occurred across two vocabularies (ICD9 and ICD10CM) within the same paper. Inclusion criteria were variably employed across the three conditions, such as inpatient diagnosis, primary diagnosis, subsequent diagnosis, specialist diagnosis, or treatment requirement. Unique to PAH, procedure-based criteria were also included. While a notable proportion of studies required a subsequent (presumably confirmatory) diagnosis, there was variation in the timeframe specified to observe the second diagnosis. Limited patients overlap and heterogeneous background IRs observed across PAs could be attributable to any, some combination of, or all these factors. Our findings suggest that any PA heterogeneity, whether in code lists or temporal logic, should be transparently reported and empirically evaluated in part of phenotype development for a research study before final design decisions are made and study execution. Further, in adherence to open-science principles, these study artifacts should be publicly registered in a study protocol a priori.

Guidelines for conducting and reporting observational research emphasize the importance of clearly defining phenotypes to specify the targeted population and ensure reproducibility, promoting transparent, open, and robust research¹⁰. Hence, we selected five elements for review. Despite the notable diversity observed in the applied PA, manuscript reporting on PA seems to follow a consistent trend. Although many papers included a paragraph outlining codes for the primary condition of interest and additional inclusion criteria, they often fell short of fulfilling these objectives. Firstly, a precise depiction of the population of interest necessitates explicitly defining the clinical description or the clinical case definition or referencing an external description from a reputable source such as national guidelines. Failure to do so risks conflating study populations. For instance, although all retrieved papers focused on Alzheimer's disease as the target population, some researchers aimed to study Alzheimer's disease while others targeted Alzheimer's-related dementias. Distinguishing between these two conditions requires careful examination of the papers. Similarly, some papers described a broader category of depression rather than specifically addressing major depressive disorder, which can only be discerned by analyzing the code sets and identifying other types of depression codes included. This poses a significant challenge to transparent and robust research, as the crucial component—the population—may not be clearly understood by the audience, leading to potential misinterpretation of the results. Second, reproducibility potential is overall unsatisfactory. In our previous research¹¹, which aligns with other previous studies, we discovered that a good quality text description is not enough for reproducibility as the researchers interpret the information in different ways leading to a drastic difference in patient sample and composition. This paper shows that the problem of reproducibility is even more severe. A large portion of the papers we examined failed to explicitly define entry event, time frame or event code sets, which were described implicitly (such as F32.xx for depression). Notably, one paper we reviewed only said that they used “Read codes” without providing the actual codes and was subsequently excluded from the review. Also, some of the papers stored their code sets in the supplements not accessible at the time of the review, compromising reproducibility. We therefore argue for explicit documentation and reporting of the computable code used to create phenotypes to enable reproducible research.

Our study has several limitations. First, the three phenotypes we used represented only three disease definitions. Thus, our findings may not generalize on phenotypes for other conditions or for symptoms or adverse events. Second, we replicated the definitions as much as the provided information allowed us to. Given that only 1 study provided the fully computable PA, and many descriptions lacked details to allow accurate reproducibility, our replications might not represent the exact definitions used in the reviewed studies. Third, we followed a structured literature review to identify relevant studies, but we did not conduct a full systematic review of the literature. Finally, our findings are limited to the data sources used. Different findings may emerge when other sources of data are used.

References

1. Hripesak George, Albers David J. High-fidelity phenotyping: richness and freedom from bias. *Journal of the American Medical Informatics Association*. 2017;25(3):289-294.
2. Xu Jie, Rasmussen Luke V, Shaw Pamela L, et al. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. *Journal of the American Medical Informatics Association*. 2015;22(6):1251-1260.
3. Wei Wei-Qi, Rowley Robb, Wood Angela, et al. Improving reporting standards for phenotyping algorithm in biomedical research: 5 fundamental dimensions. [ocae005] *Journal of the American Medical Informatics Association*. 2024.
4. Hripesak G., Shang N., Peissig P. L., et al. Facilitating phenotype transfer using a common data model [in eng]. *J Biomed Inform*. 2019;96(suppl 31325501) Aug:103253 PMC6697565.
5. <https://ohdsi.github.io/TheBookOfOhdsi/OhdsiAnalyticsTools.html#atlas>.
6. <https://ohdsi.github.io/CohortDiagnostics/>. Available at. Accessed.
7. Rao Gowtham A., Shoaibi Azza, Makadia Rupa, et al. CohortDiagnostics: phenotype evaluation across a network of observational data sources using population-level characterization. *medRxiv*. 2023:2023.2006.2028.23291982.
8. Voss Erica A, Makadia Rupa, Matcho Amy, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association*. 2015;22(3):553-564.
9. Manderscheid R. W., Ryff C. D., Freeman E. J., et al. Evolving definitions of mental illness and wellness [in eng]. *Prev Chronic Dis*. 2010;7(1)(suppl 20040234) Jan:A19 PMC2811514.

10. Wang S. V., Verpillat P., Rassen J. A., et al. Transparency and Reproducibility of Observational Cohort Studies Using Large Healthcare Databases [in eng]. *Clin Pharmacol Ther.* 2016;99(3)(suppl 26690726) Mar:325-332 PMC7833029.
11. Ostropolets A., Albogami Y., Conover M., et al. Reproducible variability: assessing investigator discordance across 9 research teams attempting to reproduce the same observational study [in eng]. *J Am Med Inform Assoc.* 2023;30(5)(suppl 36826399) Apr 19:859-868 PMC10114120.