

CohortDiagnostics: phenotype evaluation across a network of observational data sources using population-level characterization

Gowtham A. Rao^{1,2}, Azza Shoaibi^{1,2}, Rupa Makadia^{1,2}, Jill Hardin^{1,2}, Joel Swerdel^{1,2}, James Weaver^{1,2}, Erica A Voss^{1,2}, Mitchell M. Conover^{1,2}, Stephen Fortin^{1,2}, Anthony G. Sena^{1,2}, Chris Knoll^{1,2}, Nigel Hughes^{1,2}, James P. Gilbert^{1,2}, Clair Blacketer^{1,2}, Alan Andryc^{1,2}, Frank DeFalco^{1,2}, Anthony Molinaro^{1,2}, Jenna Repts^{1,2}, Martijn J Schuemie^{1,2,3}, Patrick B Ryan^{1,2,4}

1. Observational Health Data Analytics, Janssen Research and Development, LLC, Titusville, NJ, USA
2. OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY
3. Department of Biostatistics, University of California, Los Angeles, CA, USA
4. Department of Biomedical Informatics, Columbia University, New York, USA

Corresponding Author

Gowtham Rao

Janssen Research and Development, LLC

1125 Trenton-Harbourton Road

Titusville, NJ, 08560, USA

Telephone: +1 609 7306691

Email: GRao9@ITS.JNJ.com

ORCID ID: 0000-0002-4949-7236

Key words: Phenotype algorithm; phenotype evaluation; phenotype validation; measurement error; OMOP Common Data Model

Word count: 3,824/4,000 excluding title page, abstract, references, figures, and tables.

ABSTRACT

Objective:

This paper introduces a novel framework for evaluating phenotype algorithms (PAs) using the open-source tool, Cohort Diagnostics.

Materials and Methods:

The method is based on several diagnostic criteria to evaluate a patient cohort returned by a PA.

Diagnostics include estimates of incidence rate, index date entry code breakdown, and prevalence of all observed clinical events prior to, on, and after index date. We test our framework by evaluating one PA for systemic lupus erythematosus (SLE) and two PAs for Alzheimer's disease (AD) across 10 different observational data sources.

Results:

By utilizing CohortDiagnostics, we found that the population-level characteristics of individuals in the cohort of SLE closely matched the disease's anticipated clinical profile. Specifically, the incidence rate of SLE was consistently higher in occurrence among females. Moreover, expected clinical events like laboratory tests, treatments, and repeated diagnoses were also observed. For AD, although one PA identified considerably fewer patients, absence of notable differences in clinical characteristics between the two cohorts suggested similar specificity.

Discussion:

We provide a practical and data-driven approach to evaluate PAs, using two clinical diseases as examples, across a network of OMOP data sources. Cohort Diagnostics can ensure the subjects identified by a specific PA align with those intended for inclusion in a research study.

Conclusion:

Diagnostics based on large-scale population-level characterization can offer insights into the misclassification errors of PAs.

BACKGROUND AND SIGNIFICANCE

Phenotype algorithms (PA) are computerized queries used to identify specific clinical events on health data sources such as electronic health records or administrative claims.[1-4] However, the reliability of evidence generated from observational studies may be threatened by misclassification errors.[5] A reproducible framework is needed to systematically evaluate PA's for the detection, quantification, and reduction of such misclassification errors.

Misclassification errors can be assessed using metrics such as sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). These metrics typically depend on comparison to a gold standard reference classifier, such as a comprehensive disease registry or medical record reviews.

Unfortunately, disease registries are not always available, and even when they are, they often cover only a limited range of conditions and might be incomplete.[6] Medical record reviews, while valuable, are resource-intensive, time-consuming, prone to interobserver bias, and unfeasible in large deidentified data sources.[7, 8] Furthermore, most medical record reviews provide only PPV information.

Recent advances have led to the introduction of scalable alternatives like CALIBER and PheValuator.[9, 10] Although these novel methods report on the existence and magnitude of measurement errors, they do not identify the sources of these errors or suggest modifications to the PA to enhance its performance.

This manuscript proposes a framework to address these gaps and supplements existing methods for PA evaluation.

OBJECTIVE

In this work, we introduce a new framework to assess potential misclassification errors in PAs using population-level characterization. This framework has been integrated into CohortDiagnostics, an open-

source software that is able to run on person level health data in Observational Medical Outcomes Partnership (OMOP) common data model format.[11] To illustrate its effectiveness, we apply this methodology to two distinct health conditions represented as computable phenotypes: Systemic Lupus Erythematosus (SLE) and Alzheimer’s Disease (AD).

MATERIALS AND METHODS

Overview:

We use a data-driven approach to evaluate PAs. This method is based on a set of summary statistics (characterization of the cohort) that serve as diagnostic indicators. Each of these ‘diagnostics’ provides insights on potential misclassification errors.

To clarify, when a PA is run against a data source, the result is a ‘cohort’. A cohort is a set of individuals who satisfy all the criteria specified in the PA for a duration of time represented by `cohort_start_date` and `cohort_end_date`. ‘CohortDiagnostics’ is an open-source software tool that generates and visualizes summary statistics called diagnostics. These diagnostics include estimates of incidence rate; the breakdown of entry event codes on the index date (ie, cohort entry); the distribution of type of visits prior to, on, and after the index date; and the prevalence of all observed clinical events prior to, on, and after the index date.

Table 1 lists the entire set of diagnostics that are available in CohortDiagnostics and provides a guide on how to use it.

Table 1: Diagnostics and Guide on using diagnostics to infer misclassification error.

| Cohort Definition Diagnostics |
|---|
| Review the codes by vocabulary that are part of the PA, such as International Classification of Diseases (ICD), SNOMED-CT |
| Check if semantically appropriate codes have been selected as part of the PA |
| Examine if any codes in the resolved code set are semantically inconsistent with the clinical description of the target phenotype |
| Identify any missing or orphaned codes based on PHOEBE (PHEntype Observed Entity Baseline Endorser)[12, 13] |
| |

| Cohort Count Diagnostics |
|--|
| Check if PAs produce a cohort with zero counts in one or more data sources |
| Evaluate if the same PA produces lower than anticipated cohort counts in one data source, compared to other data sources |
| Look for substantial differences in counts for similar PAs when applied to the same data source |
| If the PA allows a person to enter the cohort at multiple distinct temporal periods, examine the ratio of subject count to event count |
| Identify any inclusion rules that have no subjects or very few subjects satisfying in a data source |
| Find any inclusion rules that drastically reduce the cohort counts in a data source |
| Evaluate how the impact of inclusion rules differs across data sources |
| |
| Incidence Rate Diagnostics |
| Check if the incidence rate is monotonous and stable when stratified by calendar year |
| Look for any sudden or abrupt change in pattern, and investigate possible explanations. Note: strata with low denominator count needs to be removed from consideration to avoid statistical instability. |
| Evaluate the consistency of the incidence rate across demographic sub-groups |
| Examine if the incidence rate stratified by age*sex*calendar year follows an explainable pattern |
| Compare incidence rates with those reported from external sources |
| Check for any strata that appear lower compared to the other strata, e.g., < 100 subjects |
| |
| Time Distribution Diagnostics |
| Check if the people in the cohort are observed in the data source for less than expected duration for any of the three time distributions (Time in days between a person's observation start date in the data source and their cohort start date; time in days between a person's cohort start date and cohort end date; Time in days between a person's cohort start date and the last date of continuous observation in the data source) |
| Look for persons with low or even 0 days observed in the three time distributions. |
| Evaluate if the time distribution follows a uniform or skewed distribution |
| Compare the time distributions across data sources |
| |
| Index Event Breakdown Diagnostics |
| Identify the main concept(s) that are driving entry into cohort |
| Evaluate the consistency of the concepts driving entry of subjects into the cohort across data sources |
| Check if cohort entry is predominantly or unexpectedly due to a certain code |
| |
| Visit Context Diagnostics |
| Determine the most common type of visit that a person in the cohort experiences around cohort entry |
| Check if the visit type associated with the cohort entry event matches expectations |
| Evaluate the utilization rates of Emergency department or Inpatient hospitalization shortly before, during or after index |
| |
| Cohort Overlap Diagnostics |
| Determine the proportion of subjects present in both PAs compared to only one, when overlapping two similar PAs for the same phenotype. |
| |
| Characterization Diagnostics |
| Check for concepts that should be present but aren't (e.g., proportions of common treatments, known diagnostic procedures, known risk factors or coexisting conditions). |
| Identify concepts that shouldn't be seen but are (e.g., Contradictory condition). |
| Look for other concepts that may suggest presence of disease that may invalidate the phenotype |
| Examine concepts in the time prior that suggest the outcome started earlier (e.g., Specific treatment, cooccurring conditions, specific diagnostic work, a complication or an exacerbation of the condition, the concepts in the |

| |
|--------------------------|
| concept set expression)? |
|--------------------------|

| |
|--|
| Compare cohort characteristics for the same PAs when observed over multiple data sources |
|--|

Incidence rates:

Incidence rates are calculated for all permutations of 10-year age groups, sex, and calendar year strata.

The rates are determined by dividing the number of individuals entering a stratum for the first time, termed ‘incident cohort entry’, by the sum of person-years contributed by all eligible individuals in data source who could potentially enter the strata for the first time, referred to as ‘eligible for incident cohort entry’.

Typically, incidence rates stratified by calendar year are anticipated to be present in a continuous, unvarying monotonous temporal pattern with no abrupt shifts. Should any interruptions in this pattern be observed, it could suggest alterations in either clinical practices or data capture processes, leading to potential inconsistencies in PA performance. Moreover, these incidence rate findings can be cross-referenced with expected epidemiological trends documented in the existing literature for additional validation.

Index event breakdown

Index event breakdown shows the count of cohort entries where a specific code in the PA’s entry event criteria, coincided with the index date of cohort entry. In other words, these are codes that likely triggered the cohort entry. The frequency of these codes allows us to assess their individual contributions to the cohort.

If most individuals are entering the cohort based on a limited number of the total specified codes in the PA, this could potentially point towards specificity errors. A higher occurrence of codes that might be semantically narrower compared to the clinical definition of the phenotype may suggest sensitivity errors. Additionally, any variations in the rank order of codes among different data sources might indicate measurement heterogeneity.

Visit Context

This diagnostic presents the count of individuals who experienced different types of healthcare visits (outpatient, inpatient, emergency department) in relation to the index date, as follows: 1) 'Before' represents visits that concluded within 30 days prior to the index date. 2) 'During' accounts for visits that began before and extended up to or beyond the index date. 3) 'Simultaneous' covers visits that initiated on the index date. 4) 'After' includes visits that commenced within 30 days post the index date.

We anticipate that certain types of visits will be more common for specific patient phenotypes. For instance, severe acute conditions requiring intensive care will probably result in inpatient visits. A preponderance of unanticipated visit types might suggest a specificity error.

Cohort Overlap

The cohort overlap diagnostics conducts pairwise comparison of cohorts from two PAs, reporting the individuals identified by either one or both PAs, as well as only one PA. This diagnostics in CohortDiagnostics is visualized using a Venn diagram or a table. Examining the overlap between two different PAs representing the same disease can provide insights into the potential sensitivity loss associated with one algorithm compared to the other.

Cohort Characterization

Cohort characterization diagnostics provides an overview of the cohort using descriptive statistics on demographic factors, condition, drug exposures, measurements, and occurrences of procedure codes. For each selected data source, CohortDiagnostics displays the prevalence of all observed clinical events (denoted by codes) at different time periods relative to the index date. Default time windows include a) 365 to 31 days prior to the index date, b) 30 to 1 day before the index date, c) on the index date, d) 1 to 30 days post the index date, and e) 31 to 365 days post the index date.

Clinical events are represented through one or more codes. The prevalence is given for each code individually, and some are grouped using a vocabulary hierarchy. This diagnostic feature allows us to

simulate, at the cohort level, the process by which clinicians establish and confirm clinical diagnoses. We expect individuals diagnosed with a certain disease to exhibit its signs and symptoms on or before the index date. Similarly, we anticipate diagnostic tests related to the disease to occur on or before its onset, followed by relevant treatment occurrences on or after onset. A lack of such expected characteristics might point to misclassification errors. To enable comparative analyses across multiple PAs, the tool carries out pairwise comparisons of all observed characteristics for each assessed PA. The results, including proportions or means and the standardized (mean) difference for each covariate, are presented in tables and scatter plots.

Application

Our evaluation of PAs focuses on two distinct scenarios: First, a researcher may examine a single PA on its own merit, by looking for possible misclassification errors across one or more data sources. This evaluation could guide the researcher in adjusting the PA in successive iterations to reduce possible misclassification errors. Second, a researcher may compare the diagnostic performance of two or more PAs that represent the same clinical concept in the same data source. This would help the researcher infer which PA might have lower misclassification errors, allowing them to choose from the two the PA that offers the best performance.

To illustrate these two scenarios, we implemented our proposed framework on two clinical concepts of interest - SLE and AD. Before evaluating the PA, we ensure we understand the known clinical profile of persons we are attempting to capture in the data source. This is done by writing a clinical description for medical condition/disease, with elements like overview, presentation, diagnostic evaluation, therapy plan, risk factors, and prognosis. The authored clinical description serves as a tool that enables documentation of the shared understanding among researchers of the target clinical idea. It also provides justification for the phenotype development design choices and expected clinical attributes to look for during phenotype evaluation.

Phenotypes

System Lupus Erythematosus

SLE is an autoimmune disease with a wide range of severity characterized by periods of exacerbation and relative quiescence and occurs predominantly among women of child-bearing age (15 to 44 years).

Based on SLE clinical description, we developed a PA that allows patients to enter the cohort on the earliest of either a diagnosis code, treatment for (ie, hydroxychloroquine, steroids, biologics, or immunosuppressants) or signs and symptoms related to SLE (ie, Inflammatory dermatosis, rash, joint or back pain, endocarditis), as long as there was at least one diagnosis code for SLE within 0 to 90 days

from the entry date. All patients were required to have at least 365 days of continuous observation prior to the index date. The full PA for SLE including condition and drug codes and temporal logic is in Appendix 1.

Alzheimer's Disease

AD is an age associated progressive neurodegenerative disorder and the most common cause of dementia.[14] For AD, we constructed 2 PAs. The first AD PA (referred to as the simple PA) allows patients to enter the cohort on first occurrence of an AD diagnosis. The second AD cohort is a more restrictive and is derived from the work by Imfeld et. al.,[15] requiring one of 3 inclusion criteria: 1) the first occurrence of AD diagnosis as the entry event criterion, and any of the following inclusion criteria in relation to entry date: a) a prescription on or after for AD drug, b) a second AD diagnosis any time after, c) a prior dementia test, d) a prior, simultaneous, or subsequent dementia symptom, or e) if the first occurrence was diagnosed in an inpatient setting; or having the 2) first occurrence of dementia followed by at least 2 prescriptions for AD drugs; or 3) prescription for AD drugs followed by a diagnosis of AD. Individuals were excluded if they were under 18 years of age at cohort start date, were subsequently diagnosed with diseases that, when present, make the diagnosis of AD less likely (eg, Vascular dementia, Lewy Body disease, Pick's disease), or had an occurrence of a stroke diagnosis within 2 years before index date.

Data

The data sources used in the evaluation are described in Appendix 2. We included 6 claims based data; JMDC, Merative™ MarketScan® Commercial Claims and Encounters Database (CCA), Merative™ MarketScan® Medicare Supplemental and Coordination of Benefits Database (MDCR), Merative™ MarketScan® Multi-State Medicaid Database (MDCD), IQVIA® Adjudicated Health Plan Claims Data (Pharmetrics Plus), Optum's Clinformatics® Data Mart - Socio-Economic Status (Optum SES) and 4 electronic medical record (EHR) data; IQVIA® LPD in Australia (LPDAU), IQVIA® Disease Analyzer France (France DA), IQVIA® Disease Analyzer Germany (German DA), Optum® de-identified Electronic Health

Record dataset (Optum EHR). These data sources have been standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).[16, 17] Extract, transform, and load (ETL) specifications for all data sources except LPDAU, France DA, German DA, and Pharmetrics Plus are available at ETL-LambdaBuilder.[18] The standardized data are assessed using a rigorous data quality process to evaluate conformance, completeness, and plausibility of the data.[19]

Data analysis (CohortDiagnostics software)

CohortDiagnostics is open-source software application written in the R programming language that implements the described theoretical framework.[20] Given a set of instantiated cohorts, a set of cohort definition details, and a connection to a remote database with person level data converted to the OMOP CDM [16, 17] (version 5.3+), CohortDiagnostics produces a set of aggregate summary statistics called Diagnostics. The output contains no patient-level data and has additional privacy protection using minimum cell count thresholds.[21] All output conforms to the prespecified CohortDiagnostics results data model and is formatted as unencrypted comma separated value (.csv) files (an intentional design decision to allow an investigator to audit compliance with privacy governance). The output .csv files, from one or more data sources, may then be combined and the results reviewed using an interactive R Shiny web application called DiagnosticsExplorer. The software and user documentation are available on OHDSI Github repository called CohortDiagnostics.[22]

RESULTS

Table 2 summarizes the number of patients who met the definitions for SLE and the 2 AD PAs in each data source. Below we provide brief overviews of the key insights informed by the evaluation process.

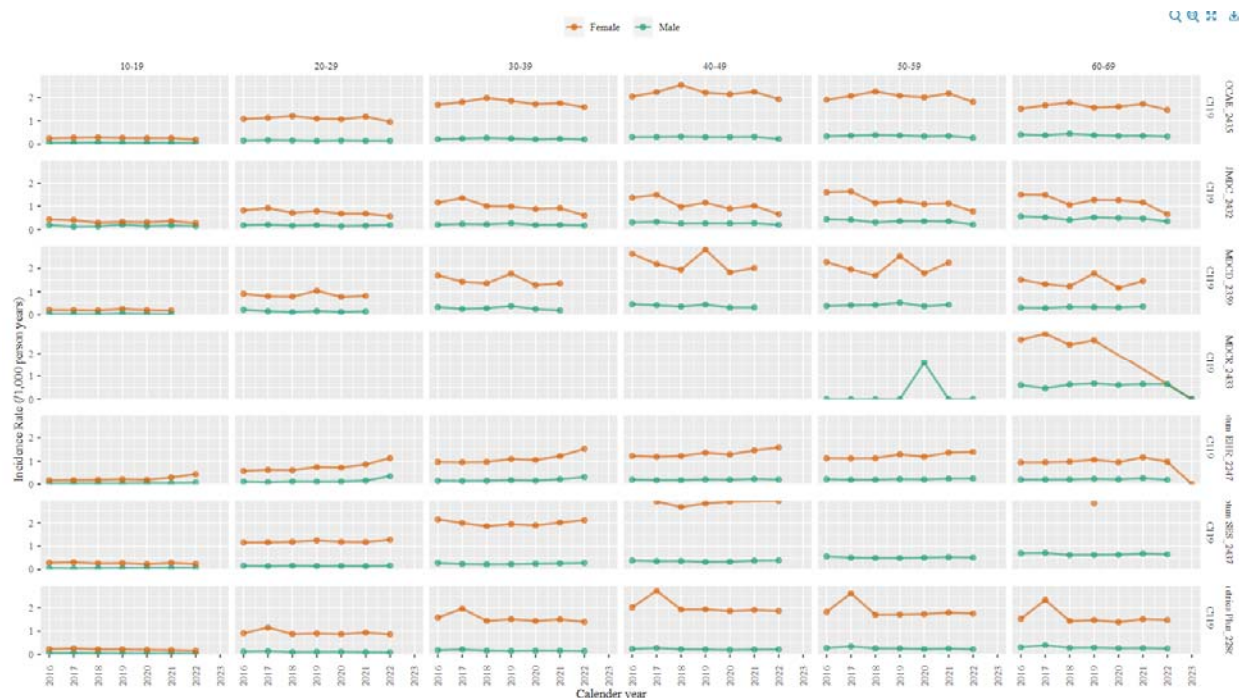
The full output of CohortDiagnostics is available in the interactive website.[23]

Systemic Lupus Erythematosus

Insights from incidence rate plots:

Figure 1 illustrates the pattern of incidence rate of SLE in each data source, stratified by age, gender and calendar year. Except for 2 general practitioner data sources France Disease Analyzer (France DA) and Germany Disease Analyzer (Germany DA), we observed high concordance across data sources for an incidence range from 30 to 50 per 100,000 person-years. Incidence rate estimation variation because of database heterogeneity in can be substantial, and such variation is not necessarily an evidence of measurement error.[24] However, observing concordance among data sources provides some reassurance that the PA measurement error is not causing substantial incidence rate heterogeneity.

Figure 1: Incidence rate of Systemic Lupus Erythematosus stratified by age decile, gender, and calendar year



As expected, females have approximately 5-fold greater incidence of SLE compared with males.

However, the rates increase by age, and peak around 40 to 50 years, which is slightly older than previously reported typical age of SLE onset of 15-44 years.[25] This may imply sensitivity error among younger age patients (eg, younger women may receive treatment for SLE like symptoms without a

diagnosis) or index date misclassification (eg, older patients may already have had the disease but its onset was not recorded in the data source).

Insights from index event breakdown:

Across data sources, a substantial proportion of individuals enter the SLE cohort based on SLE symptoms or treatment. This indicates that many patients receive treatment for SLE, before their diagnosis is coded and recorded for administrative or clinical purposes. That is, a diagnosis date is observed in the data source, but this date lags the date persons could be presumed to have the disease (represented by date treatment or symptom onset). This represents index date misclassification error.

Lastly, all the events (appearing as codes) observed on the index date are related to SLE, which suggests the absence of specificity error or false positives.

Insights from cohort characterization

Figure 2 is a screen shot from the CohortDiagnostics tool showing the most prevalent conditions and drug exposures observed in the Optum® EHR data source among the SLE cohort on the index date. SLE treatments such as prednisone, hydroxychloroquine, methotrexate and cyclophosphamide were observed on or shortly after the index date. Some individuals started these drugs in the period 365 to 30 day prior to index date, indicating potential index date misclassification. Consistent with the clinical description of SLE which stated that follow-up visits were expected, we observed SLE diagnosis codes occurring post index (30-50%). Laboratory tests such as urinalysis and antinuclear antibody were also observed (eg, in 7 to 10% in Optum® EHR on index date) and these tests clustered temporally around the index date. Observing expected baseline and post index characteristics and clinical events suggests that the patients returned by the SLE PA are likely true cases and that misclassification may be limited.

Figure 2: Characterization output from CohortDiagnostics tool showing the most prevalent conditions and drug exposures on or around index date

| Covariate Name | Analysis Name | Concept Id | Database Name | T (0d to 0d) | T (1d to 30d) | T (-30d to -1d) |
|---|---------------------|------------|---|--------------|---------------|-----------------|
| Systemic lupus erythematosus | ConditionOccurrence | 257628 | Optum® de-identified Electronic Health Record | 61.2% | 29.8% | |
| Hydroxychloroquine | DrugEraStart | 1777087 | Optum® de-identified Electronic Health Record | 17.1% | 5.0% | |
| Essential hypertension | ConditionOccurrence | 320128 | Optum® de-identified Electronic Health Record | 17.0% | 13.2% | 3.8% |
| Acetaminophen | DrugEraStart | 1125215 | Optum® de-identified Electronic Health Record | 14.4% | 8.9% | 4.4% |
| Prednisone | DrugEraStart | 1551099 | Optum® de-identified Electronic Health Record | 14.0% | 5.4% | |
| Lupus erythematosus | ConditionOccurrence | 255891 | Optum® de-identified Electronic Health Record | 10.1% | 3.7% | |
| Ondansetron | DrugEraStart | 1000960 | Optum® de-identified Electronic Health Record | 8.8% | 7.3% | 2.7% |
| Sodium chloride | DrugEraStart | 987823 | Optum® de-identified Electronic Health Record | 7.8% | 7.3% | 2.4% |
| Aspirin | DrugEraStart | 1112807 | Optum® de-identified Electronic Health Record | 7.7% | 3.3% | 1.9% |
| Hyperlipidemia | ConditionOccurrence | 432867 | Optum® de-identified Electronic Health Record | 7.4% | 5.9% | 1.8% |
| Albuterol | DrugEraStart | 1154343 | Optum® de-identified Electronic Health Record | 6.8% | 4.1% | 2.0% |
| Cholecalciferol | DrugEraStart | 19095164 | Optum® de-identified Electronic Health Record | 6.6% | 3.0% | 1.3% |
| Hydrocodone | DrugEraStart | 1174888 | Optum® de-identified Electronic Health Record | 6.6% | 5.0% | 2.3% |
| Rheumatoid arthritis | ConditionOccurrence | 80809 | Optum® de-identified Electronic Health Record | 6.0% | 3.8% | |
| Potassium chloride | DrugEraStart | 19049105 | Optum® de-identified Electronic Health Record | 5.6% | 6.6% | 1.6% |
| Anemia | ConditionOccurrence | 439777 | Optum® de-identified Electronic Health Record | 5.5% | 4.9% | |
| Levothyroxine | DrugEraStart | 1501700 | Optum® de-identified Electronic Health Record | 5.5% | 1.7% | 1.4% |
| Type 2 diabetes mellitus without complication | ConditionOccurrence | 4193704 | Optum® de-identified Electronic Health Record | 5.5% | 4.4% | 1.4% |
| Omeprazole | DrugEraStart | 923645 | Optum® de-identified Electronic Health Record | 5.1% | 2.3% | 1.3% |
| Disorder of muscle | ConditionOccurrence | 137275 | Optum® de-identified Electronic Health Record | 4.9% | 3.0% | |
| Chest pain | ConditionOccurrence | 77670 | Optum® de-identified Electronic Health Record | 4.8% | 4.6% | 1.4% |
| Acquired hypothyroidism | ConditionOccurrence | 138384 | Optum® de-identified Electronic Health Record | 4.6% | 3.0% | |
| Gastroesophageal reflux disease | ConditionOccurrence | 318800 | Optum® de-identified Electronic Health Record | 4.6% | 3.2% | |
| Gabapentin | DrugEraStart | 797399 | Optum® de-identified Electronic Health Record | 4.4% | 1.9% | 1.1% |
| Morphine | DrugEraStart | 1110410 | Optum® de-identified Electronic Health Record | 4.3% | 3.9% | 1.3% |
| Oxycodone | DrugEraStart | 1124957 | Optum® de-identified Electronic Health Record | 4.3% | 4.1% | 1.4% |
| Ibuprofen | DrugEraStart | 1177480 | Optum® de-identified Electronic Health Record | 4.3% | 2.9% | 1.5% |

Alzheimer Disease

Insights from cohort overlap:

In all data sources, the Imfeld et.al. PA returned fewer patients (19% to 81%) compared with the simpler AD PA (See Table 1).[15] In the cohort overlap, among individuals who were present in either cohort, the proportion of individuals present in both cohorts ranged between 18% and 45%. Further, 35% to 81% of individuals were identified only by the simple PA; and 0% to 20% were identified only by the Imfeld et al PA. We can infer that the simpler PA is likely to have higher sensitivity compared with the Imfeld et al PA.

Table 2: Cohort counts for the phenotype algorithms for Systemic Lupus Erythematosus and Alzheimer Disease

| Data Source | <u>SLE</u> | <u>Alzheimer's disease</u> | | |
|------------------------------|-------------------|-----------------------------------|---------------------|-------------------------------------|
| | Count | Count: Simple | Count: Imfeld, 2013 | Relative difference (Imfeld/simple) |
| <i>CCAE_2435</i> | 435,810 | 38,413 | 24,073 | 37.3% |
| <i>France DA_2354</i> | 223 | 3,804 | 718 | 81.1% |
| <i>German DA_2352</i> | 10,776 | 106,663 | 26,963 | 74.7% |
| <i>JMDC_2432</i> | 31,600 | 9,064 | 3,554 | 60.8% |
| <i>LPDAU_2353</i> | 673 | 1,052 | 373 | 64.5% |
| <i>MDCD_2359</i> | 99,165 | 349,543 | 111,224 | 68.2% |
| <i>MDCR_2433</i> | 53,697 | 479,742 | 290,711 | 39.4% |
| <i>Optum EHR_2247</i> | 260,614 | 540,074 | 435,949 | 19.3% |
| <i>Optum SES_2437</i> | 348,541 | 840,314 | 430,062 | 48.8% |
| <i>Pharmetrics Plus_2286</i> | 375,000 | 288,092 | 137,527 | 52.3% |

SLE = Systemic Lupus Erythematosus

Imfeld, P., et al, Seizures in patients with Alzheimer's disease or vascular dementia: a population-based nested case-control analysis. *Epilepsia*, 2013. 54(4): p. 700-7

Insights from visit context

The distributions of the visit type around the index date among the 2 cohorts were comparable in most data sources with less than 10% of the individuals in either cohorts identified during or at the start of an inpatient visit. This suggests that neither AD PAs were likely to capture more severe cases of AD.

Insights from cohort characterization

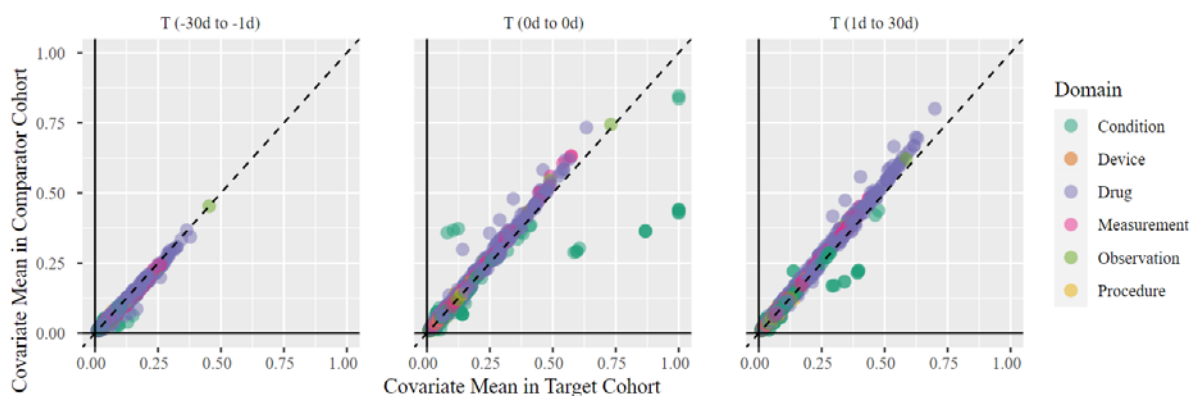
Table 3 reports a selected set of characteristics from the 2 AD cohorts at baseline ie, from 365 days before the index date up to and including the index date in the Optum® EHR data source (Data from all other data sources are available in the CohortDiagnostics shiny app). The covariates are defined using the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) vocabulary hierarchy grouping. We observe that even though the 2 cohorts were defined using different PAs and have considerably different number of patients with less than 50% overlap, the distributions of the main baseline characteristics were comparable.

Table 3: Selected baseline characteristics among patients with Alzheimers Disease by Phenotype Algorithm.

| Characteristic | Simple (n = 540,074) | Imfeld et al (n = 435,949) | Characteristic | Simple (n = 540,074) | Imfeld et al (n = 435,949) |
|----------------------------------|----------------------------|-------------------------------------|--|----------------------------|-------------------------------------|
| Age group | | | Cardiovascular disease | | |
| 55 - 59 | 1% | 1% | Atrial fibrillation | 14% | 13% |
| 60 - 64 | 2% | 3% | Cerebrovascular disease | 8% | 5% |
| 65 - 69 | 5% | 5% | Coronary arteriosclerosis | 15% | 14% |
| 70 - 74 | 11% | 10% | Heart disease | 36% | 33% |
| 75 - 79 | 32% | 33% | Heart failure | 12% | 10% |
| 80 - 84 | 37% | 36% | Ischemic heart disease | 7% | 6% |
| 85 - 89 | 11% | 10% | Peripheral vascular disease | 5% | 4% |
| Gender | | | Pulmonary embolism | 1% | 1% |
| FEMALE | 64% | 63% | Venous thrombosis | 1% | 1% |
| MALE | 36% | 37% | Neoplasms | | |
| Race | | | Hematologic neoplasm | 1% | 1% |
| Black or African American | 8% | 8% | Malignant neoplastic disease | 6% | 6% |
| White | 79% | 84% | Malignant tumor of breast | 1% | 2% |
| General | | | Primary malignant neoplasm of prostate | 1% | 1% |
| Acute respiratory disease | 7% | 6% | | | |
| Chronic liver disease | | 1% | | | |
| Chronic obstructive lung disease | 9% | 9% | | | |
| Dementia | 100% | 84% | | | |
| Depressive disorder | 16% | 15% | | | |
| Diabetes mellitus | 18% | 18% | | | |
| Gastroesophageal reflux disease | 12% | 12% | | | |
| Gastrointestinal hemorrhage | 2% | 2% | | | |
| Hyperlipidemia | 30% | 30% | | | |
| Hypertensive disorder | 47% | 44% | | | |
| Obesity | 3% | 4% | | | |
| Osteoarthritis | 13% | 13% | | | |
| Pneumonia | 6% | 6% | | | |
| Renal impairment | 18% | 16% | | | |
| Rheumatoid arthritis | 1% | 1% | | | |
| Schizophrenia | 1% | | | | |
| Urinary tract infectious disease | 13% | 11% | | | |
| Visual system disorder | 8% | 7% | | | |

Figure 3 is a screenshot from CohortDiagnostics that illustrates the covariate balance between the 2 AD PAs in Optum® EHR on 3 different time periods around the index date. Overall, we observed that most features near the diagonal, indicating comparable cohort characteristics distribution between the 2 cohorts. However, some covariates are off the diagonal with a larger a Standardized Mean Difference. For example, during 30 days to 1 days before index, we observed higher prevalence of vascular dementia, and other late effects of cerebrovascular accidents in the simple PA compared with the Imfeld et al PA. This suggests that the Imfeld et al PA is less likely to misclassify cerebrovascular accident events as AD. We also observed that the simpler PA had higher utilization of drugs commonly used in AD such as donepezil and memantine in the same immediate period prior to index date. Conversely, the Imfeld et al PA demonstrated higher utilization of these drugs on index. Both PAs had similar utilization after index. This suggests that the simple PA is subject to higher index date misclassification compared with the complex PA.

Figure 3: Covariate balance between the 2 Alzheimer's Disease Phenotype Algorithms



When we compared covariates constructed from codes that were not part of either AD PA entry event, we observed considerable cohort similarity. This suggests that the 2 PAs identified patients with the similar clinical profiles despite incomplete cohort overlap. Overall, the descriptive data of these PAs for AD revealed that, while the Imfeld et. al. identified fewer patients (raising concerns about its sensitivity),

we did not observe a higher prevalence of clinical characteristics that strongly suggest a higher specificity when compared with the other simple PA.

DISCUSSION

We have developed and integrated an empirical methodology for the evaluation of PAs into a new tool designed for the OMOP Common Data Model called CohortDiagnostics. We have demonstrated this evaluation framework on one SLE and 2 AD PAs. Our evaluation framework categorizes errors into three types: sensitivity errors, specificity errors and index date misclassification errors, providing a consistent means of assessment. This approach allows for the identification and assessment of these errors in any PA by reviewing population-level characterization.

In our application, we conclude that the SLE PA demonstrates acceptable operating characteristics and is suitable for use across the data sources assessed, even though there is potential index event misclassification. On the other hand, we found that the Imfeld et al PA may have lower sensitivity than the simple PA. The simple PA has index event misclassification and a potential specificity error explained by the observed cerebrovascular accidents events.

We have shown that this empirical and scalable framework for PA evaluation offers insights into misclassification errors. It not only detects the existence of these errors but provides an understanding of their direction and magnitude. We demonstrate that it can provide reasons for the origin of such errors, enabling researchers to refine their PAs iteratively. This method can work together with traditional case-level retrospective medical record adjudication or innovative approaches like PheValuator, which quantify estimates of measurement error.[10, 26] When paired with validation analyses for quantifying measurement errors, our population-level characterization leads to a comprehensive understanding of a PA's performance.

Our software tool, CohortDiagnostics, performs extensive diagnostics across multiple data sources for one or more PAs. It presents results in a privacy-compliant format. It is designed to perform phenotype evaluation across an observational database network. This feature allows a coordinator site to distribute a self-contained phenotype evaluation study package to each contributing data partner site, which can then independently execute it. After the execution, each site may share aggregate summary statistics back to the coordinator site, complying with local data governance and privacy policies. These site-level summary statistics can then be aggregated into one integrated viewer for collaborative review. This aggregated data can be used by a team of experts to discuss the merits of the PAs under evaluation and to understand associated misclassification errors. This framework has been recently implemented in numerous observational network studies and collaborations.[27-30]

The network-based phenotype evaluation process reinforces confidence in a PA. It allows for the evaluation of the consistency of diagnostics across different data sources, geographical locations, and time periods. Consistent trends in misclassification errors increase our confidence that our PAs have reliable operating characteristics, rather than representing an artifact from a specific data source. Such findings are crucial as they support the conclusion that a PA is applicable across various data sources. Moreover, evaluating a PA across a network offers valuable insights into different clinical settings, practices, and data capture processes. We are optimistic that this framework will encourage the use of more robust and externally valid PAs.

CohortDiagnostics also informs code selection during phenotype development. Selecting the right set of code to represent a clinical idea of interest is known to be challenging and inconsistent.[31] While code selection should be guided by clinical judgment, the empirical impact of these judgments can be readily evaluated through our tool. This evaluation can measure the effect of alternative codes on the PA performance by assessing the impact on counts and characteristics.

Despite its strengths, our approach has some limitations. It cannot numerically quantify measurement errors and should be used in conjunction with other methods that include a gold standard, such as PheValuator, or other validation methods.[32] Furthermore, analyzing descriptive results to gain insights on misclassification errors can be subjective and time-consuming. More methodological research is required to formalize a scalable, reproducible process and establish empirically driven. Finally, this approach is based on the assumption that the evaluation data sources have been standardized to the OMOP CDM and have undergone data quality review and it is fit for research use.[19]

CONCLUSION

In this paper, we introduce a framework for phenotype evaluation, that is intended to be done prior to observational research. It helps ensure that the individuals identified by the PA are consistent with the profiles of the patients we intend to study. Utilization of this framework enhances researchers' confidence in the validity of their study outcomes. The framework has been integrated into the CohortDiagnostics software. We have shown how this open-source software can enable collaborative research within a broad research community and can scale to multiple PAs, over multiple data sources that can be repeated over multiple time periods enabling creation of a repository of such evaluations.[3, 4]

DECLARATIONS

Acknowledgments

None

Ethics approval

NA

Availability of data and material

The data that support the findings of this study are available to license from Merative, Optum, IQVIA and JMDC.

Conflicts of interest / Competing Interests

All authors are employees of Janssen Research & Development, LLC, and shareholders of Johnson & Johnson (J&J) stock.

This study was sponsored by Janssen Research & Development, LLC.

Author Contributions

All coauthors contributed to the conceptualization, drafting, editing, and approving the manuscript.

REFERENCES

1. Overby, C.L., et al., *A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury*. Journal of the American Medical Informatics Association, 2013. **20**(e2): p. e243-e252.
2. Hripcsak, G. and D.J. Albers, *High-fidelity phenotyping: richness and freedom from bias*. Journal of the American Medical Informatics Association, 2017. **25**(3): p. 289-294.
3. Weaver, J., et al. *Best Practices for Creating the Standardized Content of an Entry in the OHDSI Phenotype Library*. in *5th OHDSI Annual Symposium*. 2019.
4. Chapman, M., et al., *Desiderata for the development of next-generation electronic health record phenotype libraries*. GigaScience, 2021. **10**(9): p. giab059.
5. Kuha, J., C. Skinner, and J. Palmgren, *Misclassification Error*, in *Encyclopedia of Biostatistics*. 2005.
6. Boggon, R., et al., *Cancer recording and mortality in the General Practice Research Database and linked cancer registries*. Pharmacoepidemiology and Drug Safety, 2013. **22**(2): p. 168-175.
7. Vassar, M. and M. Holzmann, *The retrospective chart review: important methodological considerations*. Journal of educational evaluation for health professions, 2013. **10**: p. 12-12.
8. Worster, A. and T. Haines, *Advanced statistics: understanding medical record review (MRR) studies*. Acad Emerg Med, 2004. **11**(2): p. 187-92.
9. Denaxas, S., et al., *UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER*. Journal of the American Medical Informatics Association, 2019. **26**(12): p. 1545-1559.
10. Swerdel, J.N., G. Hripcsak, and P.B. Ryan, *PheValuator: development and evaluation of a phenotype algorithm evaluator*. Journal of biomedical informatics, 2019. **97**: p. 103258.
11. Overhage, J.M., et al., *Validation of a common data model for active safety surveillance research*. J Am Med Inform Assoc, 2012. **19**(1): p. 54-60.
12. Ostropolets, A., et al., *PHOEBE 2.0: selecting the right concept sets for the right patients using lexical, semantic, and data-driven recommendations*.
13. Ostropolets A, R.P., Hripcsak G. *Phenotyping in distributed data networks: selecting the right codes for the right patients*. in *AMIA Annual Symposium Proceedings*. 2022.
14. Jameson, J.L., et al., *Harrison's Manual of Medicine*. 2020: McGraw-Hill.
15. Imfeld, P., et al., *Seizures in patients with Alzheimer's disease or vascular dementia: a population-based nested case-control analysis*. Epilepsia, 2013. **54**(4): p. 700-7.
16. Voss, E.A., 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): p. 553-564.
17. Hripcsak, G., et al., *Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers*. Stud Health Technol Inform, 2015. **216**: p. 574-8.
18. Blacketer, C. *ETL-LambdaBuilder*. 2023; Available from: <https://ohdsi.github.io/ETL-LambdaBuilder/>.
19. Blacketer, C., et al., *Increasing trust in real-world evidence through evaluation of observational data quality*. Journal of the American Medical Informatics Association, 2021. **28**(10): p. 2251-2257.
20. Team, R.C., *R: A language and environment for statistical computing*. 2013.
21. Samarati, P. and L. Sweeney, *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. 1998.

22. Rao, G. *Cohort Diagnostics: An R package*. 2023 [cited 2023; Available from: <https://ohdsi.github.io/CohortDiagnostics>].
23. Rao, G. *Evaluation of Systemic Lupus Erythematosus and Alzheimers Disease using Cohort Diagnostics*. 2023; Available from: <https://data.ohdsi.org/CohortDiagnostics/>.
24. Li, X., et al., *Characterizing the incidence of adverse events of special interest for COVID-19 vaccines across eight countries: a multinational network cohort study*. Medrxiv, 2021.
25. Jiménez, S., et al., *The epidemiology of systemic lupus erythematosus*. Clinical reviews in allergy & immunology, 2003. **25**: p. 3-11.
26. Swerdel, J.N., D. Ramcharran, and J. Hardin, *Using a data-driven approach for the development and evaluation of phenotype algorithms for systemic lupus erythematosus*. PLOS ONE, 2023. **18**(2): p. e0281929.
27. Shoaibi, A., et al., *Phenotype algorithms for the identification and characterization of vaccine-induced thrombotic thrombocytopenia in real world data: a multinational network cohort study*. Drug Safety, 2022. **45**(6): p. 685-698.
28. Kostka, K., et al., *Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS*. Clinical Epidemiology, 2022: p. 369-384.
29. Desai, P.M., et al., *Phenotype Development and Evaluation of Heart Failure: A Case Study in using Patient Level Prediction to Improve Phenotype Validity*.
30. Herrera, R., et al., *Epidemiology of vasomotor symptoms (VMS) in menopausal women (EpiVaSym): a multi-country, large-scale OHDSI network analytic study*.
31. Ostropolets, A., et al., *Reproducible variability: assessing investigator discordance across 9 research teams attempting to reproduce the same observational study*. Journal of the American Medical Informatics Association, 2023. **30**(5): p. 859-868.
32. Swerdel, J.N., et al., *PheValuator 2.0: Methodological improvements for the PheValuator approach to semi-automated phenotype algorithm evaluation*. Journal of Biomedical Informatics, 2022. **135**: p. 104177.

Appendix 1

Systemic lupus erythematosus indexed on signs, symptoms, treatment, or diagnosis (FP)

Human Readable Cohort Definition

Cohort Entry Events

People enter the cohort when observing any of the following:

1. condition occurrences of ‘SLE or signs and symptoms suggestive of SLE’; having at least 1 condition occurrence of ‘Systemic lupus erythematosus (SLE)’ for the first time in the person’s history, starting between 0 days before and 90 days after ‘SLE or signs and symptoms suggestive of SLE’ start date.
2. drug exposures of ‘SLE treatments’; having at least 1 condition occurrence of ‘Systemic lupus erythematosus (SLE)’ for the first time in the person’s history, starting between 0 days before and 90 days after ‘SLE treatments’ start date.

Limit cohort entry events to the earliest event per person.

Cohort Exit

The person exits the cohort at the end of continuous observation.

Cohort Eras

Entry events will be combined into cohort eras if they are within 0 days of each other.

Concept Sets:

Systemic lupus erythematosus (SLE)

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants | Mapped |
|------------|---|-----------|------------|----------|-------------|--------|
| 255891 | Lupus erythematosus | 200936003 | SNOMED | NO | YES | NO |
| 4300204 | Systemic lupus erythematosus-associated | 402865003 | SNOMED | NO | YES | NO |

| | | | | | | |
|----------|--|-----------------|--------|----|-----|----|
| | antiphospholipid syndrome | | | | | |
| 4319305 | Rash of systemic lupus erythematosus | 95332009 | SNOMED | NO | YES | NO |
| 4145240 | Renal tubulo-interstitial disorder in systemic lupus erythematosus | 307755009 | SNOMED | NO | YES | NO |
| 37016279 | Glomerular disease due to systemic lupus erythematosus | 308751000119106 | SNOMED | NO | YES | NO |
| 46273369 | Endocarditis due to systemic lupus erythematosus | 72181000119109 | SNOMED | NO | YES | NO |

SLE treatments

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants | Mapped |
|------------|-----------------------|---------|------------|----------|-------------|--------|
| 1777087 | hydroxychloroquine | 5521 | RxNorm | NO | YES | NO |
| 1551099 | prednisone | 8640 | RxNorm | NO | YES | NO |
| 1506270 | methylprednisolone | 6902 | RxNorm | NO | YES | NO |
| 1550557 | prednisolone | 8638 | RxNorm | NO | YES | NO |
| 1305058 | methotrexate | 6851 | RxNorm | NO | YES | NO |
| 19014878 | azathioprine | 1256 | RxNorm | NO | YES | NO |
| 40236987 | belimumab | 1092437 | RxNorm | NO | YES | NO |
| 1101898 | leflunomide | 27169 | RxNorm | NO | YES | NO |
| 19003999 | mycophenolate mofetil | 68149 | RxNorm | NO | YES | NO |

SLE or signs and symptoms suggestive of SLE

| Concept | Concept Name | Code | Vocabulary | Excluded | Descendants | Mapped |
|---------|--------------|------|------------|----------|-------------|--------|
|---------|--------------|------|------------|----------|-------------|--------|

| ID | | | | | | |
|----------|--|-----------------|--------|----|-----|----|
| 255891 | Lupus erythematosus | 200936003 | SNOMED | NO | YES | NO |
| 4300204 | Systemic lupus erythematosus-associated antiphospholipid syndrome | 402865003 | SNOMED | NO | YES | NO |
| 4319305 | Rash of systemic lupus erythematosus | 95332009 | SNOMED | NO | YES | NO |
| 4145240 | Renal tubulo-interstitial disorder in systemic lupus erythematosus | 307755009 | SNOMED | NO | YES | NO |
| 37016279 | Glomerular disease due to systemic lupus erythematosus | 308751000119106 | SNOMED | NO | YES | NO |
| 46273369 | Endocarditis due to systemic lupus erythematosus | 72181000119109 | SNOMED | NO | YES | NO |
| 439777 | Anemia | 271737000 | SNOMED | NO | NO | NO |
| 140214 | Eruption | 271807003 | SNOMED | NO | NO | NO |
| 45766714 | Inflammatory dermatosis | 703938007 | SNOMED | NO | NO | NO |
| 74125 | Inflammatory polyarthropathy | 417373000 | SNOMED | NO | NO | NO |
| 77074 | Joint pain | 57676002 | SNOMED | NO | NO | NO |
| 194133 | Low back pain | 279039007 | SNOMED | NO | NO | NO |
| 4272240 | Malaise | 367391008 | SNOMED | NO | NO | NO |
| 78517 | Multiple joint pain | 35678005 | SNOMED | NO | NO | NO |
| 138525 | Pain in limb | 90834002 | SNOMED | NO | NO | NO |
| 80809 | Rheumatoid arthritis | 69896004 | SNOMED | NO | NO | NO |

Alzheimer's disease (Simple)

Human Readable Cohort Definition

Cohort Entry Events

People enter the cohort when observing any of the following:

1. condition occurrence of 'Alzheimer's disease' for the first time in the person's history.

Limit cohort entry events to the earliest event per person.

Cohort Exit

The person exits the cohort at the end of continuous observation.

Cohort Eras

Entry events will be combined into cohort eras if they are within 0 days of each other.

Concept Sets:

Alzheimer's disease

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants | Mapped |
|------------|---------------------|----------|------------|----------|-------------|--------|
| 378419 | Alzheimer's disease | 26929004 | SNOMED | NO | YES | NO |

Alzheimer's disease (based on Imfeld, 2013)

Human Readable Cohort Definition

Cohort Entry Events

People with continuous observation of 365 days before event may enter the cohort when observing any of the following:

1. condition occurrence of 'Alzheimer Disease ' for the first time in the person's history; with any of the following criteria:
 1. having at least 1 drug exposure of 'Prescription for an Alzheimers disease drug', starting between 0 days before and all days after 'Alzheimer Disease ' start date.
 2. having at least 1 condition occurrence of 'Alzheimer Disease ', starting 1 days after 'Alzheimer Disease ' start date.
 3. having at least 1 procedure occurrence of 'Specific dementia test ', starting anytime on or before 'Alzheimer Disease ' start date.
 4. having at least 1 measurement of 'Specific dementia test ', starting anytime on or before 'Alzheimer Disease ' start date.
 5. having at least 1 observation of 'Specific dementia test ', starting anytime on or before 'Alzheimer Disease ' start date.
 6. having at least 1 condition occurrence of 'Dementia symptoms '.
 7. having at least 1 visit occurrence of 'Inpatient or ER visit', starting anytime on or before 'Alzheimer Disease ' start date and ending between 0 days before and all days after 'Alzheimer Disease ' start date.
2. condition occurrence of 'Dementia' for the first time in the person's history; having at least 2 drug exposures of 'Prescription for an Alzheimers disease drug', starting between 0 days before and all days after 'Dementia' start date.
3. drug exposure of 'Prescription for an Alzheimers disease drug' for the first time in the person's history; having at least 1 condition occurrence of 'Alzheimer Disease ', starting between 0 days before and all days after 'Prescription for an Alzheimers disease drug' start date.

Limit cohort entry events to the earliest event per person.

Inclusion Criteria

1. No occurrence of any other specific dementia diagnosis (e.g., VD, Pick's disease, or Lewy body dementia [LBD]) after the Alzheimer's disease diagnosis date

1. Entry events having at most 0 condition occurrences of 'Other specific dementia diagnosis (e.g., VD, Pick's disease, or Lewy body dementia [LBD])', starting 1 days after cohort entry start date.

2. No occurrence of Stroke diagnosis within 2 years prior to the Alzheimer's disease diagnosis date

Entry events having at most 0 condition occurrences of 'Stroke (ischemic or hemorrhagic)', starting between 730 days before and 0 days before cohort entry start date.

3. ≥ 18 years old

Entry events with the following event criteria: who are ≥ 18 years old.

Cohort Exit

The person exits the cohort at the end of continuous observation.

Cohort Eras

Entry events will be combined into cohort eras if they are within 0 days of each other.

Concept Sets:

Alzheimer Disease

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants |
|------------|---------------------|----------|------------|----------|-------------|
| 378419 | Alzheimer's disease | 26929004 | SNOMED | NO | YES |

Dementia

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants |
|------------|--|----------------|------------|----------|-------------|
| 37312036 | Aggression due to dementia | 788861009 | SNOMED | NO | YES |
| 37312035 | Agitation due to dementia | 788862002 | SNOMED | NO | YES |
| 4041685 | Amyotrophic lateral sclerosis with dementia | 230258005 | SNOMED | NO | YES |
| 37312031 | Anxiety due to dementia | 788866004 | SNOMED | NO | YES |
| 37312030 | Apathetic behavior due to dementia | 788867008 | SNOMED | NO | YES |
| 35608576 | Behavioral and psychological symptoms of dementia | 10171000132106 | SNOMED | NO | YES |
| 4092747 | Cerebral degeneration presenting primarily with dementia | 279982005 | SNOMED | NO | YES |
| 4182210 | Dementia | 52448006 | SNOMED | NO | YES |
| 37116464 | Dementia caused by heavy metal exposure | 733184002 | SNOMED | YES | NO |

| | | | | | |
|----------|---|-----------|--------|-----|-----|
| 37017549 | Dementia co-occurrent with human immunodeficiency virus infection | 713844000 | SNOMED | YES | NO |
| 4244346 | Dialysis dementia | 9345005 | SNOMED | YES | NO |
| 37311665 | Disinhibited behavior due to dementia | 789170003 | SNOMED | NO | YES |
| 4043378 | Frontotemporal dementia | 230270009 | SNOMED | NO | YES |
| 45765480 | Frontotemporal dementia with parkinsonism-17 | 702429008 | SNOMED | NO | YES |
| 377788 | General paresis - neurosyphilis | 51928006 | SNOMED | YES | NO |
| 45765477 | GRN-related frontotemporal dementia | 702426001 | SNOMED | NO | YES |
| 4059191 | H/O: dementia | 161465002 | SNOMED | NO | YES |
| 372610 | Postconcussion syndrome | 40425004 | SNOMED | YES | NO |
| 37017247 | Presenile dementia co-occurrent with human immunodeficiency virus infection | 713488003 | SNOMED | YES | NO |
| 37311890 | Psychological symptom due to dementia | 789011007 | SNOMED | NO | YES |
| 37312577 | Wandering due to dementia | 789062005 | SNOMED | NO | YES |

Prescription for an Alzheimers disease drug

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants |
|------------|--------------|--------|------------|----------|-------------|
| 715997 | donepezil | 135447 | RxNorm | NO | YES |
| 757627 | galantamine | 4637 | RxNorm | NO | YES |
| 701322 | memantine | 6719 | RxNorm | NO | YES |
| 733523 | rivastigmine | 183379 | RxNorm | NO | YES |
| 836654 | tacrine | 10318 | RxNorm | NO | YES |

Specific dementia test

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants |
|------------|--|-----------|------------|----------|-------------|
| 4169175 | Mini-mental state examination | 273617000 | SNOMED | NO | YES |
| 40491929 | Mini-mental state examination score | 447316007 | SNOMED | NO | YES |
| 40490379 | Assessment using mini-mental state examination | 446971008 | SNOMED | NO | YES |
| 4167593 | Abbreviated Mental Test | 273255001 | SNOMED | NO | YES |
| 4013636 | Magnetic resonance imaging | 113091000 | SNOMED | NO | YES |
| 4125350 | CT of head | 303653007 | SNOMED | NO | YES |

| | | | | | |
|---------|--|-----------|--------|----|-----|
| 4019823 | Single photon emission computerized tomography | 105371005 | SNOMED | NO | YES |
|---------|--|-----------|--------|----|-----|

Dementia symptoms

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants |
|------------|---|-----------|------------|----------|-------------|
| 4304008 | Memory impairment | 386807006 | SNOMED | NO | YES |
| 440424 | Aphasia | 87486003 | SNOMED | NO | YES |
| 132342 | Apraxia | 68345001 | SNOMED | NO | YES |
| 4173136 | Agnosia | 42341009 | SNOMED | NO | YES |
| 4024716 | Aphasia, agnosia, dyslexia AND/OR apraxia | 106169008 | SNOMED | NO | YES |

Other specific dementia diagnosis (e.g., VD, Pick's disease, or Lewy body dementia [LBD])

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants |
|------------|--------------------------------|----------------|------------|----------|-------------|
| 443605 | Vascular dementia | 429998004 | SNOMED | NO | YES |
| 44782710 | Dementia due to Pick's disease | 21921000119103 | SNOMED | NO | YES |
| 380701 | Diffuse Lewy body disease | 80098002 | SNOMED | NO | YES |

Stroke (ischemic or hemorrhagic)

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants |
|------------|-------------------------------------|-----------------|------------|----------|-------------|
| 372924 | Cerebral artery occlusion | 20059004 | SNOMED | NO | NO |
| 375557 | Cerebral embolism | 75543006 | SNOMED | NO | NO |
| 376713 | Cerebral hemorrhage | 274100004 | SNOMED | NO | YES |
| 443454 | Cerebral infarction | 432504007 | SNOMED | NO | YES |
| 441874 | Cerebral thrombosis | 71444005 | SNOMED | NO | NO |
| 439847 | Intracranial hemorrhage | 1386000 | SNOMED | NO | YES |
| 379778 | Multi-infarct dementia | 56267009 | SNOMED | YES | YES |
| 43530727 | Spontaneous cerebral hemorrhage | 291571000119106 | SNOMED | NO | NO |
| 42538062 | Spontaneous intracranial hemorrhage | 738779002 | SNOMED | NO | NO |
| 4148906 | Spontaneous subarachnoid hemorrhage | 270907008 | SNOMED | NO | NO |
| 432923 | Subarachnoid hemorrhage | 21454007 | SNOMED | NO | NO |

Inpatient or ER visit

| Concept ID | Concept Name | Code | Vocabulary | Excluded | Descendants |
|------------|------------------------------------|------|------------|----------|-------------|
| 262 | Emergency Room and Inpatient Visit | ERIP | Visit | NO | YES |

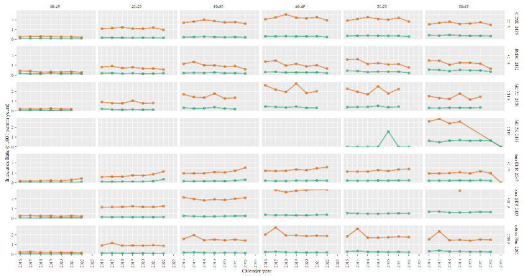
| | | | | | |
|------|-----------------|----|-------|----|-----|
| 9201 | Inpatient Visit | IP | Visit | NO | YES |
|------|-----------------|----|-------|----|-----|

Appendix 2

Database Summary

| Database ID | Database Name | Database Description | Source Country | API Reference | Open Access (Y/N) | API Type (Y/N) | API Status (Y/N) | API Version | API License | API URL | API Type (Y/N) | API Status (Y/N) | API Version | API License | API URL | API Type (Y/N) | API Status (Y/N) | API Version | API License | API URL | API Type (Y/N) | API Status (Y/N) | API Version | API License | API URL |
|-------------|---------------|----------------------------------|----------------|---------------------|-------------------|----------------|------------------|-------------|-------------|---------------------|----------------|------------------|-------------|-------------|---------------------|----------------|------------------|-------------|-------------|---------------------|----------------|------------------|-------------|-------------|---------------------|
| 1 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |
| 2 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |
| 3 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |
| 4 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |
| 5 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |
| 6 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |
| 7 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |
| 8 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |
| 9 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |
| 10 | COVID-19 | Global COVID-19 cases and deaths | USA | https://covid19.com | Y | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com | Y | Y | 1.0.0 | CC-BY-NC | https://covid19.com |

14788 149



| Product Name | Category | SKU | Description | Price | Stock | Weight |
|--------------------------------------|--------------|--------|--------------------------------------|---------|---------|---------|
| Systemic Insecticide | Conditioners | 100123 | Systemic Insecticide for Home Use | \$12.99 | 1000 | 1.0kg |
| Hydrophilic Emulsion | Detergents | 101456 | Hydrophilic Emulsion for Laundry | \$8.99 | 500 | 0.5kg |
| Essential Amino Acids | Conditioners | 102789 | Essential Amino Acids for Hair | \$15.99 | 200 | 0.2kg |
| Anti-static Agent | Detergents | 103012 | Anti-static Agent for Fabrics | \$6.99 | 300 | 0.3kg |
| Protein Shampoo | Detergents | 104345 | Protein Shampoo for Damaged Hair | \$9.99 | 400 | 0.4kg |
| Fluoropolymer Coating | Conditioners | 105678 | Fluoropolymer Coating for Textiles | \$18.99 | 150 | 0.15kg |
| Conditioner | Detergents | 106901 | Conditioner for Sensitive Skin | \$7.99 | 600 | 0.6kg |
| Softener | Detergents | 108234 | Softener for Linens | \$5.99 | 700 | 0.7kg |
| Agar | Detergents | 109567 | Agar for Laboratory Use | \$11.99 | 100 | 0.1kg |
| Hydrolyzed Protein | Conditioners | 110890 | Hydrolyzed Protein for Hair | \$14.99 | 180 | 0.18kg |
| Aluminum Chloride | Detergents | 112123 | Aluminum Chloride for Deodorant | \$9.99 | 350 | 0.35kg |
| Methylcellulose | Detergents | 113456 | Methylcellulose for Thickening | \$6.99 | 250 | 0.25kg |
| Informational Website | Conditioners | 114789 | Informational Website for Customers | \$0.99 | 1000000 | 0.001kg |
| Potassium Chloride | Detergents | 116012 | Potassium Chloride for Fertilizer | \$10.99 | 120 | 0.12kg |
| Ammonia | Conditioners | 117345 | Ammonia for Cleaning | \$4.99 | 800 | 0.8kg |
| Synthetic Fragrance | Detergents | 118678 | Synthetic Fragrance for Perfume | \$13.99 | 160 | 0.16kg |
| Lightfast Dye for Textiles | Conditioners | 119901 | Lightfast Dye for Textiles | \$16.99 | 140 | 0.14kg |
| Biodegradable | Detergents | 121234 | Biodegradable Detergent | \$8.99 | 450 | 0.45kg |
| Essence of Rose | Conditioners | 122567 | Essence of Rose for Skincare | \$19.99 | 120 | 0.12kg |
| Chlorine | Conditioners | 123890 | Chlorine for Disinfection | \$4.99 | 900 | 0.9kg |
| Argemone Seed Oil | Conditioners | 125123 | Argemone Seed Oil for Skincare | \$17.99 | 110 | 0.11kg |
| Microencapsulated Active Ingredients | Conditioners | 126456 | Microencapsulated Active Ingredients | \$20.99 | 100 | 0.1kg |
| Calcium Chloride | Detergents | 127789 | Calcium Chloride for Water Softening | \$7.99 | 280 | 0.28kg |
| Monomers | Detergents | 129012 | Monomers for Polymerization | \$11.99 | 170 | 0.17kg |
| Resin | Detergents | 130345 | Resin for Adhesives | \$9.99 | 220 | 0.22kg |
| Dispersion | Detergents | 131678 | Dispersion for Coatings | \$13.99 | 190 | 0.19kg |