Whirlwind Introduction to Open-Source Analytic Tools

Anthony Sena, Adam Black & Martijn Schuemie

6 July 2025

OHDSI

Observational Health Data Sciences and Informatics

Disclaimer

- Anthony and Martijn are employees and shareholders of Johnson & Johnson

Goal for today

The goal for today is to provide a high-level overview of the Health-Analytics Data to Evidence Suite (HADES) and related tools.

We do not have the time to teach you how to use all these tools today. But all the packages have extensive documentation, and there is the Book of OHDSI (2nd edition is coming soon!).

HADES design principles

Why convert to the Common Data Model?

- Transforming data to the OMOP CDM is a large investment
- The benefits come from being able to use the same tools and analytics across many databases

OMOP CDM Database Support

- Amazon Redshift,
- Azure DataBricks
- Azure Synapse
- DuckDB
- Google BigQuery
- InterSystems IRIS
- Microsoft SQL Server
- Oracle
- PostgreSQL
- Snowflake
- SQLite

What is HADES?

HADES
HEALTH ANALYTICS DATA-TO-EVIDENCE SUITE

Collection of open-source R packages used in almost every OHDSI study

- Run directly against data in the OMOP Common Data Model (CDM)
- Cutting edge analytics and visualizations

GitHub https://ohdsi.github.io/Hades/

Package	Version	Maintainer(s)	Availability
Achilles	v1.7.2	Frank DeFalco	CRAN
Andromeda	v0.6.7	Martijn Schuemie	CRAN
BigKnn	v1.0.2	Martijn Schuemie	GitHub
BrokenAdaptiveRidge	v1.0.0	Marc Suchard	CRAN
Capr	v2.0.8	Martin Lavallee	GitHub
Characterization	v2.0.1	Jenna Reps	GitHub
CirceR	v1.3.3	Chris Knoll	CRAN
CohortDiagnostics	v3.3.0	Jamie Gilbert	GitHub
CohortExplorer	v0.1.0	Gowtham Rao	CRAN
CohortGenerator	v0.11.2	Anthony Sena	GitHub
CohortIncidence	v4.0.0	Chris Knoll	GitHub
CohortMethod	v5.4.0	Martijn Schuemie	GitHub
Cyclops	v3.4.1	Marc Suchard	CRAN
DatabaseConnector	v6.3.2	Martijn Schuemie	CRAN
DataQualityDashboard	v2.6.1	Katy Sadowksi	GitHub
DeepPatientLevelPrediction	v2.1.0	Egill Fridgeirsson	GitHub
EmpiricalCalibration	v3.1.3	Martijn Schuemie	CRAN
EnsemblePatientLevelPrediction	v1.0.2	Jenna Reps	GitHub
Eunomia	v2.0.0	Frank DeFalco	CRAN
EvidenceSynthesis	v0.5.0	Martijn Schuemie	CRAN
FeatureExtraction	v3.7.1	Ger Inberg	CRAN
Hydra	v0.4.0	Anthony Sena	GitHub
IterativeHardThresholding	v1.0.2	Marc Suchard	CRAN
Keeper	v0.2.0	Anna Ostropolets	GitHub
MethodEvaluation	v2.3.0	Martijn Schuemie	GitHub
OhdsiSharing	v0.2.2	Lee Evans	GitHub
OhdsiShinyModules	v3.0.2	Jenna Reps	GitHub
ParallelLogger	v3.3.1	Martijn Schuemie	CRAN
PatientLevelPrediction	v6.3.9	Egill Friogeirsson & Jenna Reps	GitHub
PhenotypeLibrary	v3.34.0	Gowtham Rao	GitHub
PheValuator	v2.2.11	Joel Swerdel	GitHub
ResultModelManager	v0.5.11	Jamie Gilbert	GitHub
ROhdsiWebApi	v1.3.3	Gowtham Rao	GitHub
SelfControlledCaseSeries	v5.3.0	Martijn Schuemie	GitHub
SelfControlledCohort	v1.6.0	Jamie Gilbert	GitHub
ShinyAppBuilder	v3.1.0	Jenna Reps	GitHub
SqlRender	v1.18.1	Martijn Schuemie	CRAN

The open-source tools that empower OHDSI research are not only available to the community, but they are DEVELOPED by the community. We thank the many developers and maintainers who empower our research initiatives around the world!

HADES design principles

- Open-source software
  - Code validation & unit tests
- Driven by methods research
  - State-of-the-art methods
  - Empirically shown to work
  - Encourage use of best practices (e.g. large-scale propensity scores)
- Allow for evaluation of assumptions
  - E.g. by using negative controls
- Large-scale analytics
  - E.g. many exposures and outcomes in one study
- Platform independent

HADES R packages

41 packages so far. Some important ones

- Characterization
  - Characterization
  - CohortIncidence
  - TreatmentPatterns
- Estimation
  - CohortMethod
  - SelfControlledCaseSeries
  - EvidenceSynthesis
- Prediction
  - PatientLevelPrediction
- Cohorts
  - CohortGenerator
  - CohortDiagnostics
  - Capr
- Support
  - ROhdsiWebApi
  - DatabaseConnector + SqlRender*
  - Cyclops
  - ParallelLogger
  - FeatureExtraction
  - Andromeda
  - Strategus*

DatabaseConnector + SqlRender

- **Goal:** write SQL code once, run on all supported platforms
- Emulate temp tables for platforms that don't have them
- Translates `OhdsiSql` into the desired SQL dialect
  - OhdsiSql is a subset of SQL Server SQL
  - Separate syntax for parameterization of SQL

OhdsiSql	SqlServer	PostgreSQL
`SELECT TOP 10 person_id FROM @cdm.person;`	`SELECT TOP 10 person_id FROM my_cdm.person;`	`SELECT person_id FROM my_cdm.person LIMIT 10;`

Strategus

- **Goal**: Design and coordinate execution of HADES packages for studies.
- Design principles for Strategus:
  - All study design choices are captured in an **analysis specification**: an open, machine-readable document (JSON)
  - R execution environment fully specified using **renv** providing transparency and reproducibility
  - Patient level data **never** leaves the site
  - All aggregate results are in an open format (csv) to allow inspection
  - Supports HADES packages for managing and viewing results across sites via Rshiny (OhdsiShinyAppBuilder/OhdsiShinyModules)

Distributed Research Network

- Multiple sites with data
  - Hospital EHRs (Electronic Health Records)
  - Administrative Claims
- Patient-level data cannot be shared
- Each site uses the Common Data Model (CDM)

Distributed Research Network

- A site can lead a study

Distributed Research Network

- A site can lead a study
- Analysis code is developed locally

Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants

Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants
- Results are generated (aggregated statistics)

Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants
- Results are generated (aggregated statistics)
- Results are sent back to lead site

Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants
- Results are generated (aggregated statistics)
- Results are sent back to lead site
- Evidence is synthesized

Strategus for study execution

Tour of HADES Analytics

OHDSI

Observational Health Data Sciences and Informatics

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with **<insert your favorite disease>**, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have **<insert your favorite disease>**, which treatments were patients exposed to amongst **<list of treatments for disease>** and in which sequence?
	Outcome incidence	Amongst patients who are new users of **<insert your favorite drug>**, how many patients experienced **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?
Population-level effect estimation	**Safety surveillance**	Does exposure to **<insert your favorite drug>** increase the risk of experiencing **<insert an adverse event>** within **<time horizon following exposure start>**?
	Comparative effectiveness	Does exposure to **<insert your favorite drug>** have a different risk of experiencing **<insert any outcome (safety or benefit) >** within **<time horizon following exposure start>**, relative to **<insert your comparator treatment>**?
Patient level prediction	**Disease onset and progression**	For a given patient who is diagnosed with **<insert your favorite disease>**, what is the probability that they will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?
	Treatment response	For a given patient who is a new user of **<insert your favorite chronically-used drug>**, what is the probability that they will **<insert desired effect>** in **<time window>**?
	Treatment safety	For a given patient who is a new user of **<insert your favorite drug>**, what is the probability that they will experience **<insert adverse event>** within **<time horizon following exposure>**?

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with **<insert your favorite disease>**, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have **<insert your favorite disease>**, which treatments were patients exposed to amongst **<list of treatments for disease>** and in which sequence?
	Outcome incidence	Amongst patients who are new users of **<insert your favorite drug>**, how many patients experienced **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?
Popula effect e		**dverse**
		ny <insert your comparator treatment>?
Patient level prediction	**Disease onset and progression**	For a given patient who is diagnosed with **<insert your favorite disease>**, what is the probability that they will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?
	Treatment response	For a given patient who is a new user of **<insert your favorite chronically-used drug>**, what is the probability that they will **<insert desired effect>** in **<time window>**?
	Treatment safety	For a given patient who is a new user of **<insert your favorite drug>**, what is the probability that they will experience **<insert adverse event>** within **<time horizon following exposure>**?

Standardizing the question itself helps clarify what your objective is

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with **<insert your favorite disease>**, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have **<insert your favorite disease>**, which treatments were patients exposed to amongst **<list of treatments for disease>** and in which sequence?
	Outcome incidence	Amongst patients who are new users of **<insert your favorite drug>**, how many patients experienced **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?
Population-level effect estimation	**Safety surveillance**	Does exposure to **<insert your favorite drug>** increase the risk of experiencing **<insert an adverse event>** within **<time horizon following exposure start>**?
	Comparative effectiveness	Does exposure to **<insert your favorite drug>** have a different risk of experiencing **<insert any outcome (safety or benefit) >** within **<time horizon following exposure start>**, relative to **<insert your comparator treatment>**?
Patient level prediction	**Disease onset and progression**	For a given patient who is diagnosed with **<insert your favorite disease>**, what is the probability that they will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?
	Treatment response	For a given patient who is a new user of **<insert your favorite chronically-used drug>**, what is the probability that they will **<insert desired effect>** in **<time window>**?
	Treatment safety	For a given patient who is a new user of **<insert your favorite drug>**, what is the probability that they will experience **<insert adverse event>** within **<time horizon following exposure>**?

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with **<insert your favorite disease>**, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have **<insert your favorite disease>**, which treatments were patients exposed to amongst **<list of treatments for disease>** and in which sequence?
	Outcome incidence	Amongst patients who are new users of **<insert your favorite drug>**, how many patients experienced **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?
Popula effect e		**dverse**
		ny <insert
Patient level prediction	**progression**	**ability** that they will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?
	Treatment response	For a given patient who is a new user of **<insert your favorite chronically-used drug>**, what is the probability that they will **<insert desired effect>** in **<time window>**?
	Treatment safety	For a given patient who is a new user of **<insert your favorite drug>**, what is the probability that they will experience **<insert adverse event>** within **<time horizon following exposure>**?

Amongst patients who are new users of **GLP-1s**, how many patients experienced **Acute Myocardial Infarction** within **drug exposure**?

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with **<insert your favorite disease>**, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have **<insert your favorite disease>**, which treatments were patients exposed to amongst **<list of treatments for disease>** and in which sequence?
	Outcome incidence	Amongst patients who are new users of **<insert your favorite drug>**, how many patients experienced **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?
Population-level effect estimation	**Safety surveillance**	Does exposure to **<insert your favorite drug>** increase the risk of experiencing **<insert an adverse event>** within **<time horizon following exposure start>**?
	Comparative	Does exposure to **<insert your favorite drug>** have a different risk of experiencing **<insert any**
Patient level prediction		
	Treatment safety	For a given patient who is a new user of **<insert your favorite drug>**, what is the probability that they will experience **<insert adverse event>** within **<time horizon following exposure>**?

For a given patient who is a new user of **GLP-1s**, what is the probability that they will experience **an AMI** while **exposed to the drug**?

Analytic use case	Type	Structure
Clinical chara	Disease Natural History	Amongst patients who are diagnosed with <insert your favorite disease>, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have <insert your favorite disease>, which treatments were patients exposed to amongst <list of treatments for disease>, and in which sequence?
		time
Population level effect estimation		n adverse
	Comparative effectiveness	Does exposure to <insert your favorite drug> have a different risk of experiencing <insert any outcome (safety or benefit) > within <time horizon following exposure start>, relative to <insert your comparator treatment>?
Patient level prediction	Disease onset and progression	For a given patient who is diagnosed with <insert your favorite disease>, what is the probability that they will go on to have <another disease or related complication> within <time horizon from diagnosis>?
	Treatment response	For a given patient who is a new user of <insert your favorite chronically-used drug>, what is the probability that they will <insert desired effect> in <time window>?
	Treatment safety	For a given patient who is a new user of <insert your favorite drug>, what is the probability that they will experience <insert adverse event> within <time horizon following exposure>?

Does exposure to **GLP-1s** have a different risk of experiencing **AMI** while **exposed to drug**, relative to **DPP-4s**?

One question, three answers

"Recent studies have demonstrated that conducting hundreds or thousands of sensitivity analyses to evaluate the influence of design and analysis choices ... is becoming feasible and could be one of the future direction to appraise the robustness of pharmacoepidemiological results"

[5]Univ. Grenoble Alpes,

Correspondence
Charles Khouri, Pharma
Pharmacovigilance, CH
Email: ckhouri@chu-grenoble.fr

HADES Cohort packages

Cohort construction and evaluation

Capr
Develop and manipulate complex cohort definitions in R
Learn more…

CirceR
An R wrapper for Circe, a library for creating cohort definitions, expressing them as JSON, SQL, or Markdown.
Learn more…

CohortGenerator
Instantiating cohorts in a database based on a set of cohort definitions.
Learn more…

PhenotypeLibrary
The OHDSI Phenotype Library: a collection of community-maintained pre-definined cohorts.
Learn more…

CohortDiagnostics
Generate a wide set of diagnostics to evaluate cohort definitions against databases in the CDM.
Learn more…

PheValuator
Semi-automated evaluation of cohorts, producing metrics suchs as sensitivity, specificity, and positive and negative preditive value.
Learn more…

CohortExplorer
Visually explore all individual-level data of patients in a cohort
Learn more…

Keeper
Knowledge-Enhanced Electronic Profile Review
Learn more…

OHDSI Cohort definitions

OHDSI Cohort definitions are

- the foundation of HADES analytics
- transparent
- reproducible
- transportable
- reusable

OHDSI Cohort definitions

A Cohort is a set of persons who satisfy one or more inclusion criteria for a duration of time.

$$\text{Cohort} = \{ (\text{person}_1, \text{day}_1), (\text{person}_1, \text{day}_2), (\text{person}_1\ \text{day}_3), \ldots (\text{person}_i, \text{day}_j) \}$$

Consequences of this definition:

- One person may belong to multiple cohorts
- One person may belong to the same cohort at multiple different time periods
- One person may not belong to the same cohort multiple times during the same period of time
- One cohort may have zero or more members
- A codeset is NOT a cohort… logic for how to use the codeset in criteria is required

Anatomy of a cohort definition

1. What initial event(s) define cohort entry?

3. What defines a person's cohort exit?

2. What inclusion criteria are applied to the initial events?

Building Cohorts with Atlas

ATLAS

English

- Home
- Data Sources
- Search
- Concept Sets
- Cohort Definitions
- Characterizations
- Cohort Pathways
- Incidence Rates
- Profiles
- Estimation
- Prediction
- Reusables
- Jobs
- Configuration
- Feedback

Cohort #1791399

created by anonymous on 2024-11-21 15:11, modified by anonymous on 2024-11-21 15:35

GLP1 exposure_new_user_dm

Definition | Concept Sets | Generation | Samples | Reporting | Export | Versions | Messages 2

Enter a cohort definition description here

Cohort Entry Events

Events having any of the following criteria:

+ Add Initial Event...

a drug exposure of GLP1 ingredients_dm

+ Add attribute...

Delete Criteria

✕ for the first time in the person's history

with continuous observation of at least 365 days before and 0 days after event index date

Limit initial events to: earliest event per person.

Restrict initial events

Inclusion Criteria

New inclusion criteria

Limit qualifying events to: earliest event per person.

Cohort Exit

Event Persistence:

Event will persist until: end of a continuous drug exposure

Continuous Exposure Persistence:

Specify a concept set that contains one or more drugs. A drug era will be derived from all drug exposure events for any of the drugs within the concept set, using the specified persistence window as a maximum allowable gap in days between successive exposure events and adding a specified surveillance window to the final exposure event. If no exposure event end date is provided, then an exposure event end date is inferred to be event start date + days supply in cases when days supply is available or event start date + 1 day otherwise. This event persistence assures that the cohort end date will be no greater than the drug era end date.

Concept set containing the drug(s) of interest: GLP1 ingredients_dm

- Persistence window: allow for a maximum of 30 days between exposure records when inferring the era of persistence exposure
- Surveillance window: add 0 days to the end of the era of persistence exposure as an additional period of surveillance prior to cohort exit.
- Use days supply and exposure end date for exposure duration. *change*

Censoring Events:

Exit Cohort based on the following criteria:

+ Add Censoring Event...

Apache 2.0
open source software

provided by
OHDSI
join the journey

Building cohorts with Capr

```
library(Capr)

GLP1 <- cs(descendants(93143,44506754,40170911,1583722,45774435,44816332))

GPL1CohortDefinition <- cohort(
  entry = entry(drugExposure(GLP1, firstOccurrence()),
                continuousObservation(priorDays = 365L),
                primaryCriteriaLimit = "First"),
  exit = drugExit(GLP1, persistenceWindow = 30L)
)
```

Building lots of cohorts with Capr

```
library(Capr)
# cohort template function
DrugEraCohortDefinition <- function(drugConceptSet) {
  cohort(
    entry = entry(drugExposure(drugConceptSet, firstOccurrence()),
                  continuousObservation(priorDays = 365L),
                  primaryCriteriaLimit = "First"),
    exit = drugExit(drugConceptSet, persistenceWindow = 30L)
  )
}

GLP1 <- cs(descendants(93143,44506754,40170911,1583722,45774435,44816332))

DrugEraCohortDefinition(drugConceptSet = GLP1)
DrugEraCohortDefinition(drugConceptSet = DPP4)
```

Cohort Definition

Circe

Atlas

Capr

json


```
{
 "ConceptSets": [
  {
   "id": 0,
   "name": "Alzheimer Disease ",
   "expression": {
    "items": [
     {
      "concept": {
       "CONCEPT_CLASS_ID": "Clinical Finding",
       "CONCEPT_CODE": "26929004",
       "CONCEPT_ID": 378419,
       "CONCEPT_NAME": "Alzheimer's disease",
       "DOMAIN_ID": "Condition",
       "INVALID_REASON": "V",
       "INVALID_REASON_CAPTION": "Valid",
       "STANDARD_CONCEPT": "S",
       "STANDARD_CONCEPT_CAPTION": "Standard",
       "VOCABULARY_ID": "SNOMED"
      },
      "includeDescendants": true
     }
    ]
   }
  },
  {
```


SQL

```
CREATE TABLE #Codesets (
  codeset_id int NOT NULL,
  concept_id bigint NOT NULL
)
;

INSERT INTO #Codesets (codeset_id, concept_id)
SELECT 0 as codeset_id, c.concept_id FROM (select distinct I.concept_id
(
  select concept_id from @vocabulary_database_schema.CONCEPT where cond
UNION  select c.concept_id
  from @vocabulary_database_schema.CONCEPT c
  join @vocabulary_database_schema.CONCEPT_ANCESTOR ca on c.concept_id
  and ca.ancestor_concept_id in (378419)
  and c.invalid_reason is null

) I
) C UNION ALL
SELECT 2 as codeset_id, c.concept_id FROM (select distinct I.concept_id
(
  select concept_id from @vocabulary_database_schema.CONCEPT where cond
UNION  select c.concept_id
  from @vocabulary_database_schema.CONCEPT c
  join @vocabulary_database_schema.CONCEPT_ANCESTOR ca on c.concept_id
  and ca.ancestor_concept_id in (37312036,37312035,4041685,37312031,373
  and c.invalid_reason is null

) I
LEFT JOIN
(
  select concept_id from @vocabulary_database_schema.CONCEPT where cond

) E ON I.concept_id = E.concept_id
WHERE E.concept_id is null
) C UNION ALL
SELECT 3 as codeset_id, c.concept_id FROM (select distinct I.concept_id
```

Cohort Generation

Circe

CohortGenerator

json

```
{
 "ConceptSets": [
  {
   "id": 0,
   "name": "Alzheimer Disease ",
   "expression": {
    "items": [
     {
      "concept": {
       "CONCEPT_CLASS_ID": "Clinical Finding",
       "CONCEPT_CODE": "26929004",
       "CONCEPT_ID": 378419,
       "CONCEPT_NAME": "Alzheimer's disease",
       "DOMAIN_ID": "Condition",
       "INVALID_REASON": "V",
       "INVALID_REASON_CAPTION": "Valid",
       "STANDARD_CONCEPT": "S",
       "STANDARD_CONCEPT_CAPTION": "Standard",
       "VOCABULARY_ID": "SNOMED"
      },
      "includeDescendants": true
     }
    ]
   }
  },
  {
```

SQL

```
CREATE TABLE #Codesets (
  codeset_id int NOT NULL,
  concept_id bigint NOT NULL
)
;

INSERT INTO #Codesets (codeset_id, concept_id)
SELECT 0 as codeset_id, c.concept_id FROM (select distinct I.concept_id
(
  select concept_id from @vocabulary_database_schema.CONCEPT where cond
UNION  select c.concept_id
  from @vocabulary_database_schema.CONCEPT c
  join @vocabulary_database_schema.CONCEPT_ANCESTOR ca on c.concept_id
  and ca.ancestor_concept_id in (378419)
  and c.invalid_reason is null

) I
) C UNION ALL
SELECT 2 as codeset_id, c.concept_id FROM (select distinct I.concept_id
(
  select concept_id from @vocabulary_database_schema.CONCEPT where cond
UNION  select c.concept_id
  from @vocabulary_database_schema.CONCEPT c
  join @vocabulary_database_schema.CONCEPT_ANCESTOR ca on c.concept_id
  and ca.ancestor_concept_id in (37312036,37312035,4041685,37312031,373
  and c.invalid_reason is null

) I
LEFT JOIN
(
  select concept_id from @vocabulary_database_schema.CONCEPT where cond

) E ON I.concept_id = E.concept_id
WHERE E.concept_id is null
) C UNION ALL
SELECT 3 as codeset_id, c.concept_id FROM (select distinct I.concept_id
```

CDM Database

Generated Cohort

CohortGenerator & RObdsiWebApi

- Options for defining cohorts
  - Define cohorts using ATLAS and download using RObdsiWebApi
  - Use Capr to define cohorts
- Generate cohorts against your CDM using CohortGenerator
- Advanced topics:
  - Define subsets using demographics, nested by indication, etc using CohortGenerator

Cohorts of our examples

- Indication cohorts:
  - Hypertension — People with hypertension, while having hypertension
- Exposures cohorts :
  - ACE Inhibitors — People on ACE Inhibitors, while on the drug
  - Diuretic — People on a diuretic, while on the drug
- Outcomes cohorts :
  - Acute myocardial infarction (**AMI**) — People with AMI, at the time of AMI
  - Angioedema — People with Angioedema, while having angioedema

These same cohorts can be re-used to answer different questions

The OHDSI Phenotype Library

PhenotypeLibrary 3.36.0 | Reference | Articles | Changelog | HADES

The OHDSI phenotype library

R-CMD-check passing | codecov 100%

PhenotypeLibrary is part of HADES.

Gowtham A Rao

Annoucements on OHDSI forums
Release notes
Definition catalog

The Observational Health Data Sciences and Informatics (OHDSI) community has developed a publicly accessible, version-controlled Phenotype Library to guide real-world evidence towards the FAIR principles: Findability, Accessibility, Reproducibility, and Interoperability.[1] This library aims to foster the submission and retrieval of high-quality cohort definitions, cataloging of metadata, attribution and promotion of discovery and reuse in scientific research.

Within the OHDSI Phenotype Library (OHDSI PL), each entry represents a unique cohort definition identifiable by a stable, externally referenceable ID. Comprehensive metadata about each cohort definition is cataloged and made searchable for researchers.[2] Content in the library is subject to version control, with each version is assigned a specific DOI.

The OHDSI PL employs a community engagement and contribution process, crediting contributors via ORCID when provided. Submitted cohort definitions are subject to a voluntary, open peer review process managed by the OHDSI Phenotype Development and Evaluation Workgroup. All cohort definitions are computable and portable and conform to the specifications of the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) promoting efficient implementation, standard terminology use, seamless conversions between computable and human-readable definitions, and consistent understanding of the logic.[3-6]

Links

Browse source code
Report a bug
Ask a question

License

Apache License

Citation

Citing PhenotypeLibrary

Developers

Gowtham Rao
Maintainer

The OHDSI Phenotype Library

- Globally unique cohort IDs
- Peer reviewed
- Clinical descriptions
- Good starting point for cohort building

Cohort Definitions in OHDSI Phenotype Library

Gowtham A. Rao

2025-04-02

Source: `vignettes/CohortDefinitionsInOhdsiPhenotypeLibrary.Rmd`

Download Cohorts.csv

Search

Cohort Id	ame Long	Librarian	St…	Added Version	Logic Description	Hash …	Is Circe Json
3	Sputum	rao@ohd…	Pen…		All events of cough or …	#Sympt…	1
4		rao@ohd…	Pen…		All events of diarrhea i…	#Sympt…	1
5		rao@ohd…	Pen…		All events of dyspnea i…	#Sympt…	1
6		rao@ohd…	Pen…		All events of fever or e…	#Sympt…	1
7	e or Headac…	rao@ohd…	Pen…		All events of Headach…	#Sympt…	1
8	nell or tast…	rao@ohd…	Pen…		Altered smell or taste i…	#Sympt…	1
9	at	rao@ohd…	Pen…		All events of Sore throat	#Sympt…	1
10	r Vomiting	rao@ohd…	Pen…		All events of Nausea o…	#Sympt…	1
11	nd or fatigue	rao@ohd…	Pen…		All events of Malaise a…	#Sympt…	1
12	r common …	rao@ohd…	Pen…		All events of Rhinitis o…	#Sympt…	1
13	not explain…	rao@ohd…	Pen…		All events of Myalgia. …	#Sympt…	1
14		rao@ohd…	Pen…		All events of Myalgia	#Sympt…	1
20	s or Bronch…	rao@ohd…	Pen…		All events of Bronchitis	#Sympt…	1
21	piratory dis…	rao@ohd…	Pen…		All events of Acute res…		1

CohortDiagnostics

Provides several diagnostics on cohort definitions:

- Concept set diagnostics: check sensitivity/specificity of concept sets
- Cohort characterization: clinical plausibility
- Incidence rates: temporal trends, identify database artifacts
- Attrition: evaluate impact of each inclusion criteria
- Time distributions: time in cohort, observation time prior to and after.
- Cohort overlap: check if cohorts are subsets of one another

Topics to keep in mind when evaluating cohorts:

- Sensitivity / Specificity tradeoff
- Capture the correct index date
- Incidence vs prevalent cases

CohortExplorer

ACE inhibitor

Database id

synthea1m

Cohort id

20126

Subject ID

2290686

Age at First Index Date

67

Gender

FEMALE

< >

Domains

- [ ] Feature Cohort Data
- [ ] Condition Era
- [x] Condition Occurrence
- [x] Drug Era
- [ ] Drug Exposure
- [x] Procedure Occurrence
- [x] Measurement
- [x] Observation
- [x] Visit Occurrence

Concept Name Filter (keep)

regex

Concept Name Filter (remove)

regex

Absolute days from Start

9999

Absolute days to Start

9999

Search

Days To First	Concept Name	Type Concept Name	Start Date	End Date	Domain	Cdm Table	Records
0	Cohort entry	Cohort entry	2007-05-02	2007-05-02	Cohort	Cohort	
0	Acute non-ST segment elevation myocardial infarction	Prescription dispensed in pharmacy	2007-05-02	2007-05-02	Condition	condition_occurrence	1
0	aspirin	No matching concept	2007-05-02	2007-05-02	Drug	drug_era	1
0	atenolol	No matching concept	2007-05-02	2007-05-02	Drug	drug_era	1
0	simvastatin	No matching concept	2007-05-02	2007-05-02	Drug	drug_era	1
0	prasugrel	No matching concept	2007-05-02	2007-05-02	Drug	drug_era	1
0	Inpatient Visit	Visit derived from encounter on claim	2007-05-02	2007-05-03	Visit	visit_occurrence	1
0	Emergency Room Visit	Visit derived from encounter on claim	2007-05-02	2007-05-02	Visit	visit_occurrence	1
-1477	Chronic kidney disease stage 2	Prescription dispensed in pharmacy	2011-05-18	2011-05-18	Condition	condition_occurrence	1
-1477	Microalbuminuria due to type 2 diabetes	Prescription dispensed in pharmacy	2011-05-18	2011-05-18	Condition	condition_occurrence	1

Cohort building examples

- ACE inhibitors
- ACE inhibitors nested in Hypertension
- Acute myocardial infarction

Other phenotype validation tools

- PheValuator
  - Build a predictive model on definitely true and definitely false cases
  - Use model to generate annotated silver standard
  - Compute sensitivity and specificity
- KEEPER
  - Define patient attributes to extract (symptoms, treatments, diagnostic procedures, alternative diagnoses, etc.)
  - Extract attributes per patient
  - Have human or LLM adjudicate patient profiles

Journal of Biomedical Informatics 135 (2022) 104177

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

PheValuator 2.0: Methodological improvements for the PheValuator approach to semi-automated phenotype algorithm evaluation

Joel N. Swerdel [a,c,*], Martijn Schuemie [a,c], Gayle Murray [a], Patrick B. Ryan [a,b,c]

[a] Janssen Research and Development, Titusville, NJ, USA
[b] Columbia University, New York, NY, USA

Journal of the American Medical Informatics Association, 2024, 31(1), 119–129
https://doi.org/10.1093/jamia/ocad202
Advance access publication 17 October 2023
Research and Applications

Research and Applications

Scalable and interpretable alternative to chart review for phenotype evaluation using standardized structured data from electronic health records

Anna Ostropolets, MD, PhD*,[1], George Hripcsak, MD, MS[1,2], Syed A. Husain, MD, MPH[3], Lauren R. Richter, MD, MS[1], Matthew Spotnitz, MD, MPH[1], Ahmed Elhussein, MD, MS[1], Patrick B. Ryan, PhD[1,4]

[1]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, United States, [2]Medical Informatics

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01433-4

Standardized patient profile review using large language models for case adjudication in observational research

Martijn J. Schuemie [1,2,3], Anna Ostropolets[1,4], Aleh Zhuk[1,5], Uladzislau Korsik[1,5], Seung In Seo[1,6], Marc A. Suchard[1,3], George Hripcsak[1,4] & Patrick B. Ryan[1,2,4]

Characterization: CohortIncidence package

Amongst patients who are new users of **GLP-1s,** how many patients experienced **Acute Myocardial Infarction** within **drug exposure**?

- Target: **GLP-1**
- Outcome: **AMI**

Computes the incidence rate of the Outcome cohort in some Target cohort

- Standardized computation of incidence rates
- Default: overall and stratified by age, sex, and calendar time

PatientLevelPrediction package

For a given patient who is a new user of **GLP-1s**, what is the probability that they will experience **an AMI** while **exposed to the drug**?

- Target: **GLP-1**, restricted to those with **T2DM** (and first use only)
- Outcome: **AMI**

Builds a model to predict who in the Target will have the Outcome

- Uses all observed data up to Target start
- Implements many machine learning / deep learning algorithms

Unique feature: external validation

- When using the PatientLevelPrediction package, models fit in one database can easily be validated in other databases
  - Portability of code
  - Standardized construction of features

Example of external validation

John *et al. BMC Medicine* (2024) 22:308
https://doi.org/10.1186/s12916-024-03530-9
BMC Medicine

RESEARCH ARTICLE **Open Access**

Development and validation of a patient-level model to predict dementia across a network of observational databases

Luis H. John[1*], Egill A. Fridgeirsson[1], Jan A. Kors[1], Jenna M. Reps[2], Ross D. Williams[1], Patrick B. Ryan[2] and Peter R. Rijnbeek[1]

Abstract

Background A prediction model can be a useful tool to quantify the risk of a patient developing dementia

CohortMethod package

Does exposure to **GLP-1s** have a different risk of experiencing **AMI** while **exposed to drug**, relative to **DPP-4s**?

- Target: **GLP-1**, restricted to those with **T2DM** (and first use only)
- Comparator: **DPP-4**, restricted to those with **T2DM** (and first use only)
- Outcome: **AMI**

Computes the hazard of the Outcome cohort in the Target cohort compared to the Comparator

Unique feature: Large-scale propensity scores

- Treatment assignment is often non-random, which can cause confounding
  - E.g. GLP-1 may be prescribed more often to obese, who already have a higher risk of AMI
- Propensity scores are an establish way to address this
  - Fit a model to predict treatment assignment, and use to compute probability (propensity score)
  - Match subjects in Target to Comparator with similar propensity scores
- Traditionally, expert pick a few variables to use in the prediction model
- Large-scale propensity scores include all baseline covariates, and uses machine learning

Unique feature: objective diagnostics

- Whether study results are reliable depends on whether certain assumptions have been met
  - E.g. we assume our PS adjustment makes our treatment groups comparable
- Most of these assumptions are testable through diagnostics
  - E.g. we can test whether our PS adjustment achieved balance by computing the standardized difference of means (SDM)
- By 'objective' diagnostics we mean diagnostics that are evaluated while blinded to the results of the study
  - E.g. Pre-specify that we will not look at results where $\max(|SDM|) > 0.1$
  - Unique: negative controls

Example of a negative control

RESEARCH PAPER
Multiple Sclerosis 2008; **14**: 307–313

Selective association of multiple sclerosis with infectious mononucleosis

BM Zaadstra[1,2], AMJ Chorus[1], S van Buuren[1,3], H Kalsbeek[1] and JM van Noort[4]

Example of a negative control

Odds ratio:

Infectious mononucleosis	2.22 *	Multiple sclerosis
Rubella	1.31 *	
Measles	1.42 *	

* P < .05

RESEARCH PAPER Multiple Sclerosis 2008; 14: 307–313

Selective association of multiple sclerosis with infectious mononucleosis

BM Zaadstra[1,2], AMJ Chorus[1], S van Buuren[1,3], H Kalsbeek[1] and JM van Noort[4]

Example of a negative control

How to interpret negative control findings?

- Unique: use a sample ($n > 50$) of negative controls to understand distribution of bias
- Systematic error distribution can be used as
  - Diagnostic: if too much systematic error, we stop
  - Calibration: can adjust p-values and confidence intervals to take into account possible systematic error

Quantifying systematic error

Need to execute estimation studies for 66 Target-Outcome combinations.

OHDSI tools readily allow for this (simply swapping out the outcome cohort for the negative controls)

Received: 8 July 2022 | Revised: 30 September 2022 | Accepted: 8 December 2022
DOI: 10.1002/sim.9631

RESEARCH ARTICLE

Statistics in Medicine WILEY

Adjusting for both sequential testing and systematic error in safety surveillance using observational data: Empirical calibration and MaxSPRT

Martijn J. Schuemie[1,2] | Fan Bu[2,3] | Akihiko Nishimura[4] | Marc A. Suchard[2,3,5]

[1]Observational Health Data Analytics, Janssen Research & Development, Titusville, New Jersey,
[2]Department of Biostatistics, University of California, Los Angeles, California,
[3]Department of Human Genetics, University of California, Los Angeles,

Post-approval safety surveillance of medical products using observational healthcare data can help identify safety issues beyond those found in pre-approval trials. When testing sequentially as data accrue, maximum sequential probability ratio testing (MaxSPRT) is a common approach to maintaining nominal type 1 error. However, the true type 1 error may still deviate from the

Quantifying systematic error

Expected Absolute Systematic Error (**EASE**) summarizes this distribution

We use a **prespecified** EASE threshold (EASE < 0.25) for go – no go decisions for our studies

Combining evidence across a federated network: What to share when estimating causal effects?

- Cannot share patient-level data
- Usually **propensity-score stratified time-to-event** or **conditional Poisson regression**: no 2-by-2 tables
- Point-estimates + standard errors?

Normal assumption violated when counts are low

Hazard Ratio = 1.02 (0.27 – 3.78) Assuming normal distribution

Hazard Ratio = 1.02 (0.22 – 3.31) No shape assumption

	Subjects	Outcomes
Target	22,002	3
Comparator	130,200	25

* Real data, no simulation

Even more when counts are 0

	Subjects	Outcomes
Target	2,834	0
Comparator	15,168	10

* Real data, no simulation

Solution

Data sites share actual shape of likelihood, instead of just the hazard ratio + confidence interval

Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2022, Vol. 31(3) 438–450
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802211060518
journals.sagepub.com/home/smm
SAGE

Combining cox regressions across a heterogeneous distributed research network facing small and zero counts

Martijn J. Schuemie[1,2,3], Yong Chen[4], David Madigan[1,5], and Marc A. Suchard[1,3,6]

Abstract

Studies of the effects of medical interventions increasingly take place in distributed research settings using data from multiple clinical data sources including electronic health records and administrative claims. In such settings, privacy concerns typically prohibit sharing of individual patient data, and instead, cross-network analyses can only utilize summary statistics from the individual databases such as hazard ratios and standard errors. In the specific but very common context of the Cox proportional hazards model, we show that combining such per site summary statistics into a single network-wide estimate using standard meta-analysis methods leads to substantial bias when outcome counts are small. This bias derives

Solution

Data sites share actual shape of likelihood, instead of just the hazard ratio + confidence interval

Current best-practice:

- Adaptive grid

Distributed Research Network

- Data sites share shape of likelihood

Visualization

Normal approximation

Source	Hazard Ratio (95% CI)
Site 1	2.38 (0.51 - 11.08)
Site 2	-
Site 3	1.34 (0.37 - 4.87)
Site 4	2.23 (0.22 - 22.08)
Site 5	-
Site 6	1.93 (0.19 - 19.45)
Site 7	1.11 (0.22 - 5.54)
Site 8	-
Site 9	-
Site 10	1.55 (0.32 - 7.55)
Summary (tau = 0.25)	1.55 (0.77 - 3.25)

Adaptive grid approximation

Source	Hazard Ratio (95% CI)
Site 1	2.38 (0.69 - 10.00)
Site 2	-
Site 3	1.34 (0.45 - 5.81)
Site 4	2.23 (0.42 - 10.00)
Site 5	-
Site 6	1.93 (0.36 - 10.00)
Site 7	1.11 (0.29 - 6.81)
Site 8	-
Site 9	-
Site 10	1.55 (0.43 - 9.38)
Summary (tau = 0.28)	2.67 (1.33 - 5.26)

Summary

Unique features of HADES analytics

- Re-use of cohort definitions
- Standardization of analytics in open-source software
  - Many opportunities for testing, review, fixing bugs, etc.
  - Making it hard to do the wrong thing (opinionated)
- Advanced methods to reduce bias
  - Splines for time in self-controlled case series
  - Large-scale propensity scores in cohort method
- Objective study diagnostics to improve reliability of evidence
  - Including negative controls
- Designed to run across a network of databases
  - Without sharing patient-level data