

RWD Analysis Using Open Source Software

Introduction to OMOP CDM and R Packages for Analysis

PHUSE Japan Open-source Technology Working Group

December 5, 2025

Slides Available Here

English:

<https://nissinbo.github.io/phuse-sde-tokyo-2025-omop/en>



Japanese:

<https://nissinbo.github.io/phuse-sde-tokyo-2025-omop/ja>



Introduction to OMOP CDM

What is OMOP CDM?

Observational **M**edical **O**utcomes **P**artnership **C**ommon **D**ata **M**odel

- Standardized data model for unified RWD analysis
- Enhances reproducibility through common schema and standardized vocabularies
- Developed and maintained by the OHDSI community



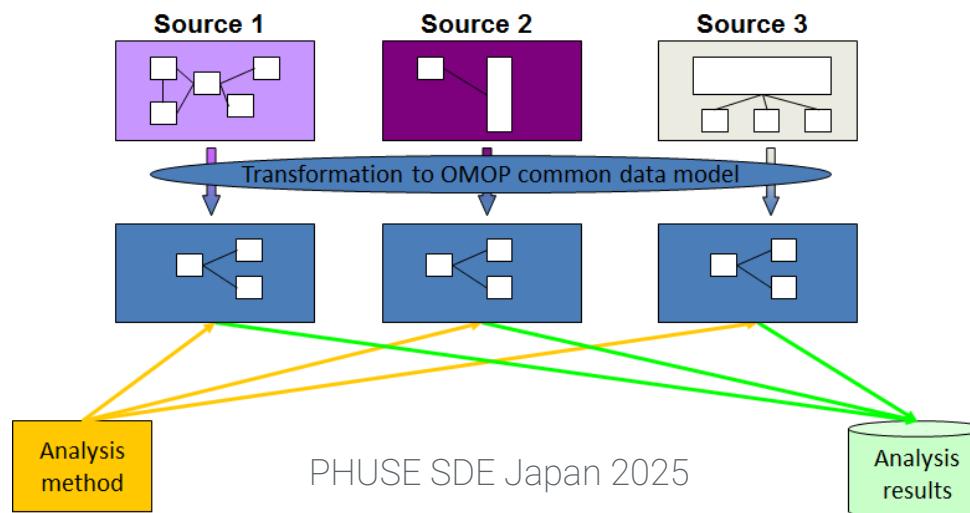
► OHDSI

(i) Note

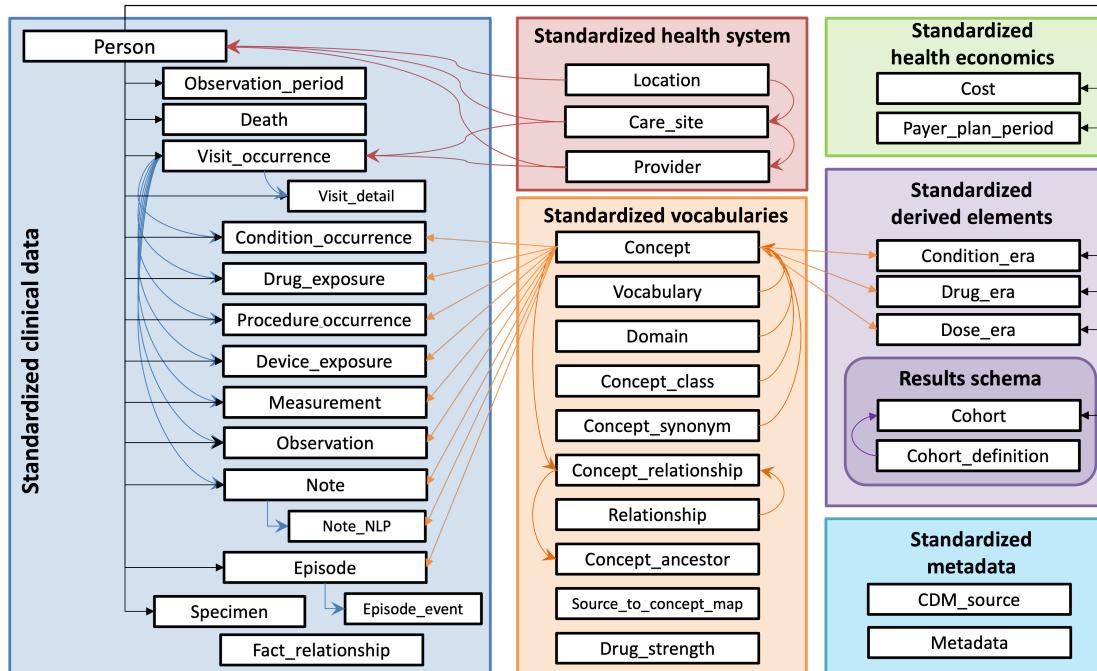
In 2025, **PHUSE Working Group** was established to promote OMOP adoption

Key Features of OMOP CDM

- **Advantages: Facilitates comparative studies across different data sources**
 - Unified terminology through standardized vocabularies
 - Designed with minimal tables necessary for observational research
- **Challenges: Complexity of transformation process**
 - Requires ETL processes specific to each data source
 - Difficulty in mapping to standard vocabularies



OMOP CDM Structure (v5.4)



- **Clinical data:** Person, Observation Period, Visit Occurrence, ...
- **Health system:** Location, Care Site, Provider
- **Vocabularies:** Concept, Vocabulary, Concept Relationship, ...

Core Table: Person

Basic patient information and demographics

Field	Description
person_id	Patient ID
gender_concept_id	Gender
year_of_birth	Birth year
race_concept_id	Race
ethnicity_concept_id	Ethnicity



Tip

All clinical events are linked through **person_id**

Core Table: Visit Occurrence

Healthcare facility visit and admission information

Field	Description
visit_occurrence_id	Visit identifier
person_id	Patient ID
visit_concept_id	Visit type (inpatient/outpatient/emergency)
visit_start_date	Visit start date
visit_end_date	Visit end date

Core Table: Condition Occurrence

Disease and symptom diagnosis information

Field	Description
condition_occurrence_id	Diagnosis identifier
person_id	Patient ID
condition_concept_id	Standard concept ID for condition
condition_start_date	Diagnosis start date
condition_type_concept_id	Record source (EHR/claims)

Core Table: Drug Exposure

Drug exposure information

Field	Description
drug_exposure_id	Drug exposure identifier
person_id	Patient ID
drug_concept_id	Standard concept ID for drug
drug_exposure_start_date	Exposure start date
drug_exposure_end_date	Exposure end date

Code Mapping

Various codes are often mapped to standard concepts (not mandatory)

- **Standard concepts:** Defined by SNOMED CT , RxNorm, etc.
- **Non-standard concepts:** Codes in source data such as ICD10, LOINC
- Concepts can be searched using **ATHENA** web tool

CONCEPT_ID	313217	Primary key
CONCEPT_NAME	Atrial fibrillation	English description
DOMAIN_ID	Condition	Domain
VOCABULARY_ID	SNOMED	Vocabulary
CONCEPT_CLASS_ID	Clinical Finding	Class in vocabulary
STANDARD_CONCEPT	S	Standard, Source of Classification
CONCEPT_CODE	49436004	Code in vocabulary
VALID_START_DATE	01-Jan-1970	Valid during time interval
VALID_END_DATE	31-Dec-2099	
INVALID_REASON		

Example: Hypertension

- **Standard:** SNOMED 38341003
- **Non-standard:** ICD10 I10, MeSH D006973

Analyzing OMOP with R

What is HADES?

Health Analytics Data-to-Evidence Suite

- A collection of R packages specialized for OMOP CDM data analysis
- High interoperability (works seamlessly when using HADES packages together)
- Actively developed by two organizations: OHDSI and DARWIN EU



► HADES

► OHDSI

► DARWIN EU

HADES Package List

Packages

Below are the packages included in HADES. For each package a link is provided with more information, including instructions on how to install and use the package.

Population-level estimation

CohortMethod New-user cohort studies using large-scale regression for propensity and outcome models. Learn more...	SelfControlledCaseSeries Self-Controlled Case Series analysis using few or many predictors, includes splines for age and seasonality. Learn more...	SelfControlledCohort A self-controlled cohort design, where time preceding exposure is used as control. Learn more...
EvidenceSynthesis Routines for combining causal effect estimates and study diagnostics across multiple data sites in a distributed study. Learn more...		

Patient-level prediction

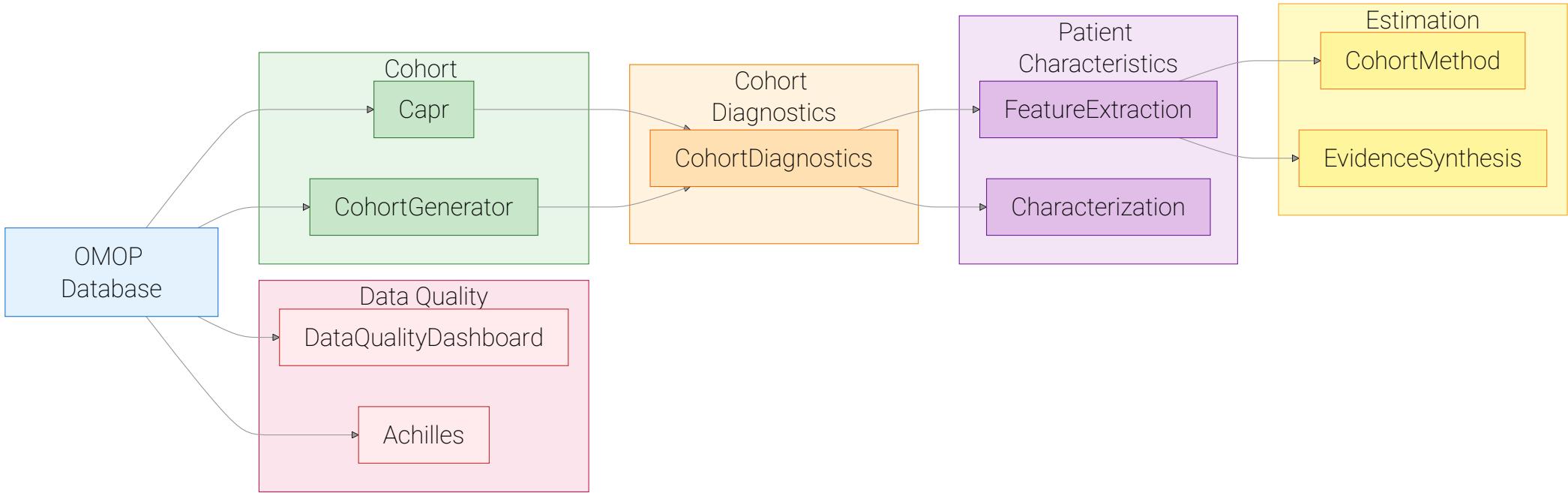
PatientLevelPrediction Build and evaluate predictive models for user-specified outcomes, using a wide array of machine learning algorithms. Learn more...	DeepPatientLevelPrediction Performing patient level prediction using deep learning Learn more...	EnsemblePatientLevelPrediction Building and validating ensemble patient-level predictive models. Learn more...
--	---	---

► [HADES Packages](#)

As of December 2025, 41 packages are registered in HADES!

Example Workflow Using HADES

End-to-end analysis workflow is achievable



Let's Try It Out! 😊

Setup

Install R packages

```
1 install.packages(c("duckdb", "here", "CDMConnector", "OmopSketch",
2                     "PatientProfiles", "IncidencePrevalence", "CohortSurvival"))
```

Download sample data

```
1 library(CDMConnector)
2
3 Sys.setenv("EUNOMIA_DATA_FOLDER" = here::here())
4 downloadEunomiaData("GiBleed")
```

Download completed!

CDMConnector + Basics



► CDMConnector

CDMConnector

Database connection and data access

```
1 library(CDMConnector)
2 library(tidyverse)
3 library(dbplyr)
4
5 # Connect to database
6 con <- DBI::dbConnect(duckdb::duckdb(), eunomiaDir("GiBleed"))
7
8 # List tables
9 DBI::dbListTables(con)
```

```
[1] "care_site"           "cdm_source"          "concept"
[4] "concept_ancestor"    "concept_class"        "concept_relationship"
[7] "concept_synonym"     "condition_era"        "condition_occurrence"
[10] "cost"                "death"                 "device_exposure"
[13] "domain"              "dose_era"              "drug_era"
[16] "drug_exposure"       "drug_strength"        "fact_relationship"
[19] "location"             "measurement"          "metadata"
[22] "note"                 "note_nlp"              "observation"
[25] "observation_period"   "payer_plan_period"    "person"
[28] "procedure_occurrence" "provider"              "relationship"
[31] "source_to_concept_map" "specimen"               "visit_detail"
[34] "visit_occurrence"     "vocabulary"
```

CDMConnector

Create CDM object

Use `cdmFromCon()` to create an OMOP-specific object format

```
1 cdm <- cdmFromCon(con, cdmSchema = "main", writeSchema = "main")
2 cdm
```

cdm Object

Access each table using \$

```
1 cdm$person |>  
2   collect() |>  
3   glimpse()
```

Rows: 2,694

Columns: 18

\$ person_id	<int> 6, 123, 129, 16, 65, 74, 42, 187, 18, 111, ...
\$ gender_concept_id	<int> 8532, 8507, 8507, 8532, 8532, 8532, 8532, ...
\$ year_of_birth	<int> 1963, 1950, 1974, 1971, 1967, 1972, 1909, ...
\$ month_of_birth	<int> 12, 4, 10, 10, 3, 1, 11, 7, 11, 5, 8, 3, 3...
\$ day_of_birth	<int> 31, 12, 7, 13, 31, 5, 2, 23, 17, 2, 19, 13...
\$ birth_datetime	<dttm> 1963-12-31, 1950-04-12, 1974-10-07, 1971-...
\$ race_concept_id	<int> 8516, 8527, 8527, 8527, 8516, 8527, 8527, ...
\$ ethnicity_concept_id	<int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
\$ location_id	<int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
\$ provider_id	<int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
\$ care_site_id	<int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
\$ person_source_value	<chr> "001f4a87-70d0-435c-a4b9-1425f6928d33", "0...
\$ gender_source_value	<chr> "F", "M", "M", "F", "F", "F", "M", "F...
\$ gender_source_concept_id	<int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
\$ race_source_value	<chr> "black", "white", "white", "white", "black...
\$ race source concept id	<int> 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. ...

Basic Data Operations

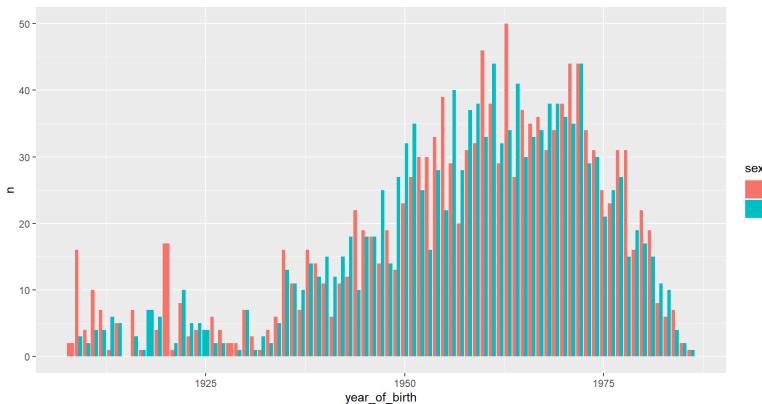
Distribution of conditions among males born after 1975

```
1 cdm$person |>
2   filter(year_of_birth >= 1975, gender_source_value == "M") |>
3   left_join(cdm$condition_occurrence, by = "person_id") |>
4   summarise(n = n(), .by = condition_concept_id) |>
5   left_join(cdm$concept |> select(concept_id, concept_name), by = c("condition_concept_id" = "concept_i
6   collect() |>
7   arrange(desc(n))

# A tibble: 62 × 3
  condition_concept_id     n concept_name
          <int> <dbl> <chr>
1        40481087    744 Viral sinusitis
2        4112343     419 Acute viral pharyngitis
3        260139      354 Acute bronchitis
4        372328      210 Otitis media
# i 58 more rows
```

Visualization Too!

```
1 cdm$person |>
2   summarize(n = n(), .by = c(year_of_birth, gender_concept_id)) |>
3   mutate(sex = case_when(
4     gender_concept_id == 8532 ~ "Female",
5     gender_concept_id == 8507 ~ "Male"
6   )) |>
7   collect() |>
8   ggplot(aes(y = n, x = year_of_birth, fill = sex)) +
9   geom_col(position = "dodge")
```



Note

`tidyverse` style data handling is possible!

OmopSketch

Understanding Database Overview



► OmopSketch

OmopSketch

Get an overview of the entire database (tibble)

```
1 library(OmopSketch)
2
3 cdm |>
4 summariseOmopSnapshot() |>
5 tableOmopSnapshot(type = "tibble")
```

A tibble: 13 × 3

	Variable	Estimate
1	General	Snapshot date
2	General	Person count
3	General	Vocabulary version
4	Observation period	N
5	Observation period	Start date
6	Observation period	End date
7	Cdm	Source name
8	Cdm	Version
9	Cdm	Holder name
10	Cdm	Release date
11	Cdm	Description
12	Cdm	Documentation reference
13	Cdm	Source type

[header_name]Database name\n[hea...¹

<chr>

2025-12-06

2,694

v5.0 18-JAN-19

5,343

1908-09-22

2019-07-03

Synthea synthetic health database

v5.3.1

OHDSI Community

2019-05-25

SyntheaTM is a Synthetic Patient ...

<https://synthetichealth.github.io...>

duckdb

i abbreviated name: ^{1`[header_name]Database name\n[header_level]Synthea`}

OmopSketch

Get an overview of `condition_occurrence` (flextable)

```
1 cdm |>
2 summariseClinicalRecords("condition_occurrence") |>
3 tableClinicalRecords(type = "flextable")
```

Variable name	Variable level	Estimate name	Database name
			Synthea
condition_occurrence			
Number records	–	N	65,332
Number subjects	–	N (%)	2,694 (100.00%)
Records per person	–	Mean (SD)	24.25 (7.41)
		Median [Q25 - Q75]	23 [19 - 29]
		Range [min to max]	[5 to 65]
In observation	No	N (%)	450 (0.69%)
	Yes	N (%)	64,882 (99.31%)
Domain	Condition	N (%)	65,332 (100.00%)
Source vocabulary	Icd10cm	N (%)	479 (0.73%)
	No matching concept	N (%)	27 (0.04%)
	Snomed	N (%)	64,826 (99.23%)
Standard concept	S	N (%)	65,332 (100.00%)
Type concept id	Ehr encounter diagnosis	N (%)	65,332 (100.00%)

OmopSketch

Get an overview of `drug_exposure` (flextable)

```
1 cdm |>
2 summariseClinicalRecords("drug_exposure") |>
3 tableClinicalRecords(type = "flextable")
```

Variable name	Variable level	Estimate name	Database name
			Synthea
drug_exposure			
Number records	-	N	67,713
Number subjects	-	N (%)	2,694 (100.00%)
Records per person	-	Mean (SD)	25.13 (5.25)
		Median [Q25 - Q75]	25 [22 - 28]
		Range [min to max]	[7 to 54]
In observation	No	N (%)	251 (0.37%)
	Yes	N (%)	67,462 (99.63%)
Domain	Drug	N (%)	67,713 (100.00%)
Source vocabulary	Cvx	N (%)	25,713 (37.97%)
	Ndc	N (%)	2,694 (3.98%)
	No matching concept	N (%)	35 (0.05%)
	Rxnorm	N (%)	39,271 (58.00%)
Standard concept	S	N (%)	67,713 (100.00%)
Type concept id	Dispensed in outpatient office	N (%)	25,713 (37.97%)
	Prescription written	N (%)	42,000 (62.03%)

PatientProfiles

Adding Patient Characteristics



► PatientProfiles

Patient Profiles

Define a cohort of “patients with bronchitis”

```
1 cdm <- cdm |>
2   generateConceptCohortSet(
3     name = "bronchitis",
4     conceptSet = list("any_bronchitis" = c(260139, 258780)),
5     limit = "all",
6     end = 0
7   )
8
9 cdm$bronchitis |>
10 collect()

# A tibble: 8,232 × 4
  cohort_definition_id subject_id cohort_start_date cohort_end_date
              <int>      <int>    <date>            <date>
1                      1         65 2001-09-15    2001-09-15
2                      1        184 1964-01-25    1964-01-25
3                      1        332 1978-04-12    1978-04-12
4                      1        366 1970-11-13    1970-11-13
# i 8,228 more rows
```

PatientProfiles

Add patient characteristics to cohort

```
1 library(PatientProfiles)
2
3 # Date of birth, age, sex
4 cdm$bronchitis |>
5   addDateOfBirth() |>
6   addSex() |>
7   addAge()
8
9 # Prior/future observation periods from index date
10 cdm$bronchitis |>
11   addPriorObservation() |>
12   addFutureObservation()
```

Note

Nothing much to say here. It's incredibly simple!

IncidencePrevalence

Calculating Prevalence and Incidence



► IncidencePrevalence

IncidencePrevalence

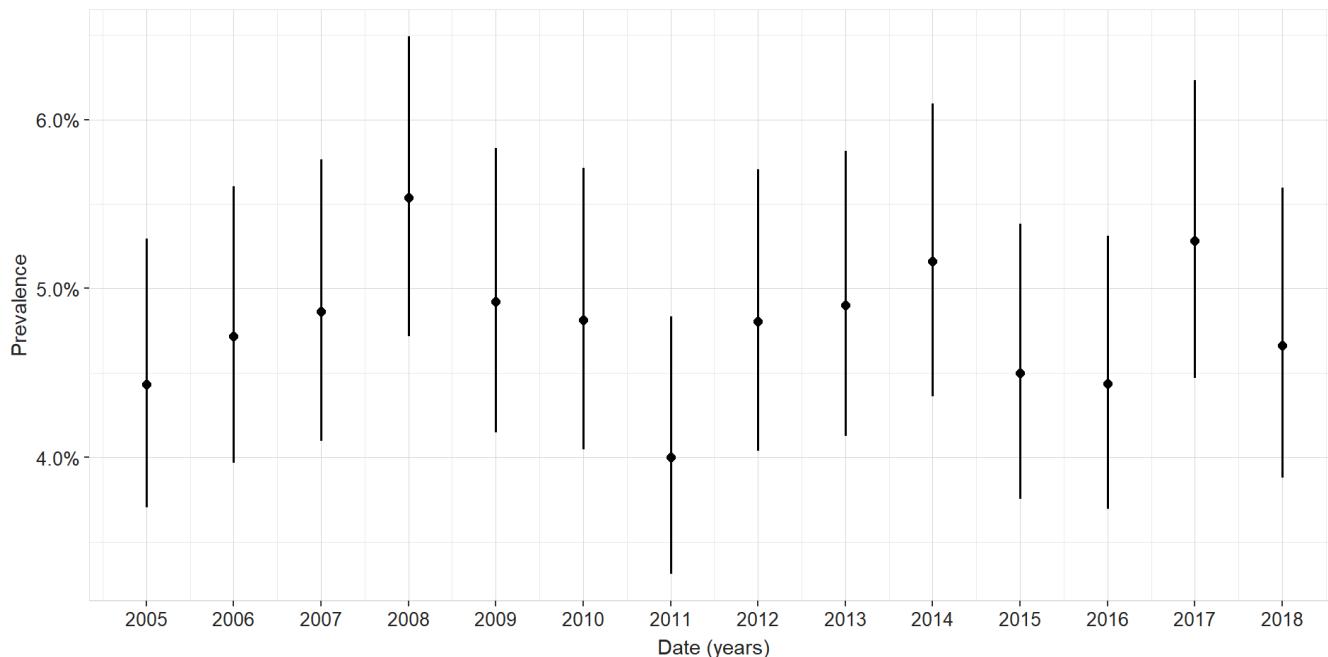
Create “denominator” cohort

```
1 library(IncidencePrevalence)
2
3 cdm <- cdm |>
4   generateDenominatorCohortSet(
5     "denom",
6     cohortDateRange = c(as.Date("2005-01-01"), as.Date(NA)))
7   )
```

IncidencePrevalence

Calculate prevalence

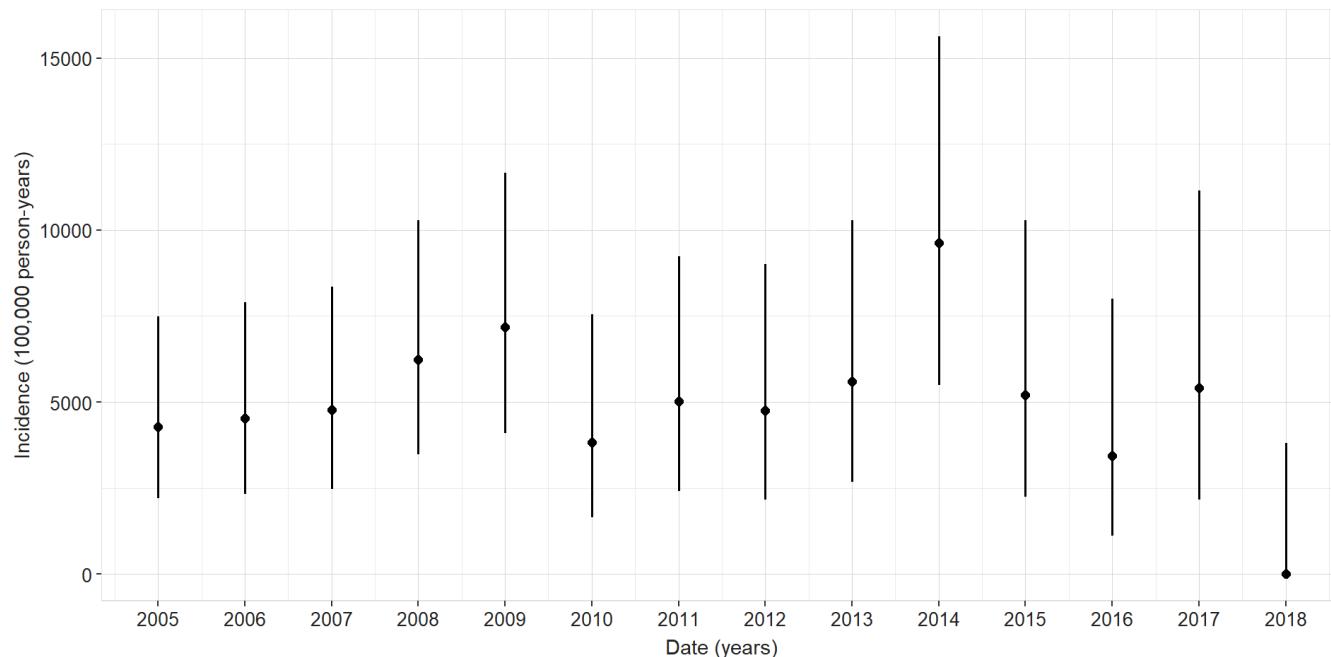
```
1 cdm |>
2   estimatePeriodPrevalence(
3     denominatorTable = "denom",
4     outcomeTable = "bronchitis"
5   ) |>
6   plotPrevalence()
```



IncidencePrevalence

Calculate incidence

```
1 cdm |>
2   estimateIncidence(
3     denominatorTable = "denom",
4     outcomeTable = "bronchitis"
5   ) |>
6   plotIncidence()
```



CohortSurvival

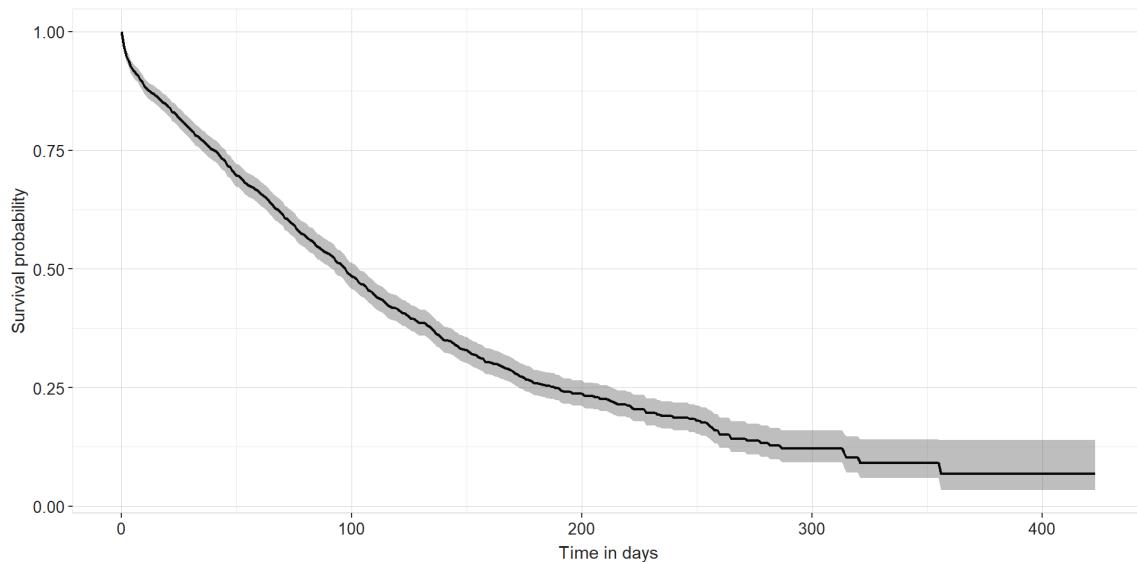
Survival Analysis



► CohortSurvival

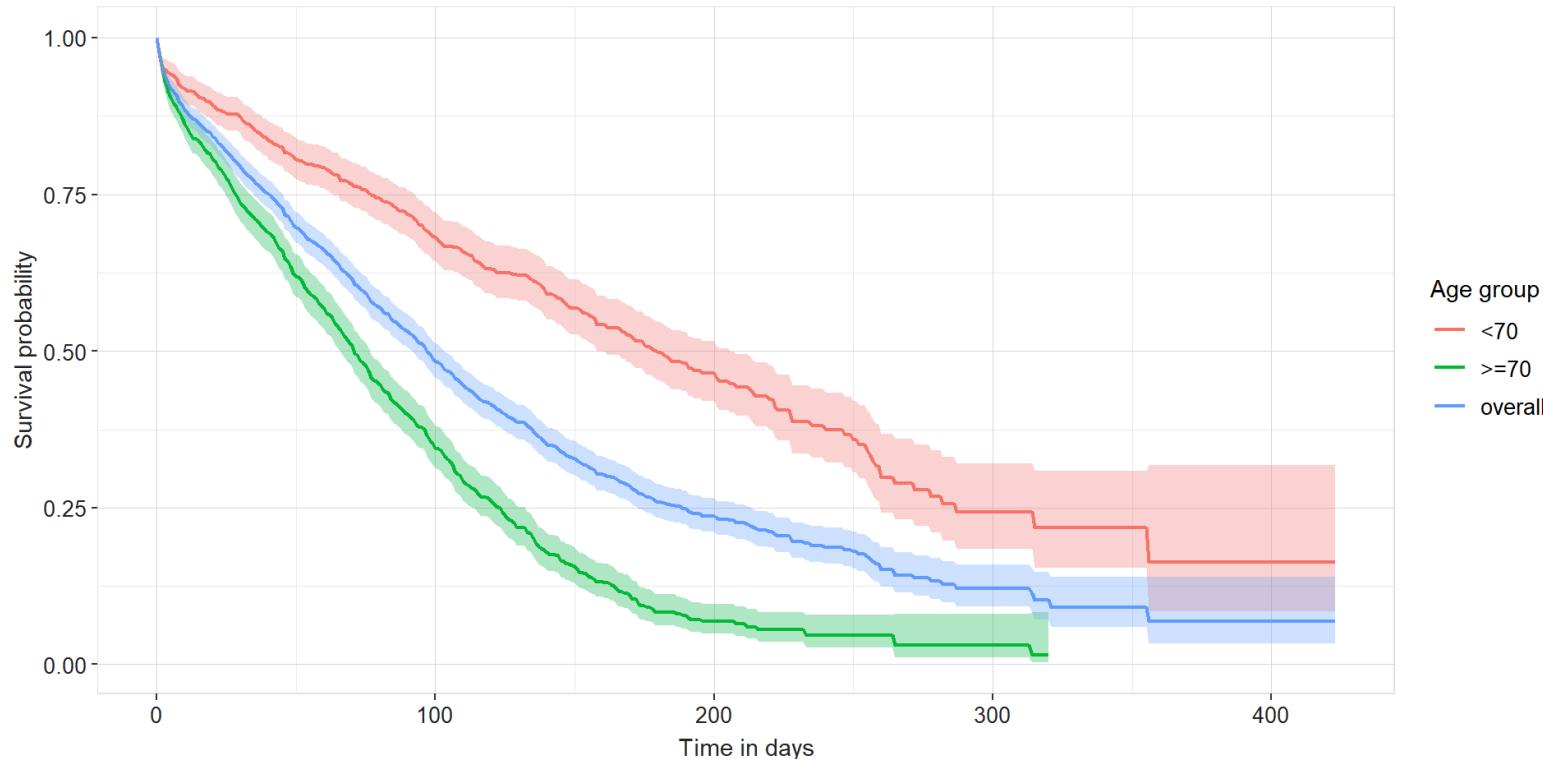
CohortSurvival

```
1 library(CohortSurvival)
2
3 # Sample data for survival analysis
4 cdm <- mockMGUS2cdm()
5
6 cdm |>
7   estimateSingleEventSurvival(
8     targetCohortTable = "mgus_diagnosis",
9     outcomeCohortTable = "death_cohort"
10 ) |>
11   plotSurvival()
```



CohortSurvival

```
1 cdm |>
2   estimateSingleEventSurvival(
3     targetCohortTable = "mgus_diagnosis",
4     outcomeCohortTable = "death_cohort",
5     strata = list(c("age_group")))
6 ) |>
7   plotSurvival(colour = "age_group")
```



Summary

- **OMOP CDM**

- **OMOP CDM**
 - Common data model for observational and RWD research
 - Standardizes different databases to enable reproducible analysis

- **R Packages for Analysis**

- **R Packages for Analysis**
 - Ecosystem centered around HADES
 - Many convenient packages specialized for OMOP analysis

Learning Resources

- The Book of OHDSI
- The Book of OHDSI (Japanese translation)
- OMOP CDM Documentation
- HADES
- DARWIN EU
- ATHENA
- Prieto-Alhambra Group - University of Oxford
 - Quarto Pub
 - Tidy R programming with the OMOP common data model

Community

- OHDSI Japan
- OHDSI Forums
- PHUSE OSS Technology WG