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-ost-omop/en

Japanese:
https://nissinbo.github.io/phuse-sde-tokyo-2025-ost-omop/ja

Introduction to OMOP CDM

What is OMOP CDM?

***O**bservational **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
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

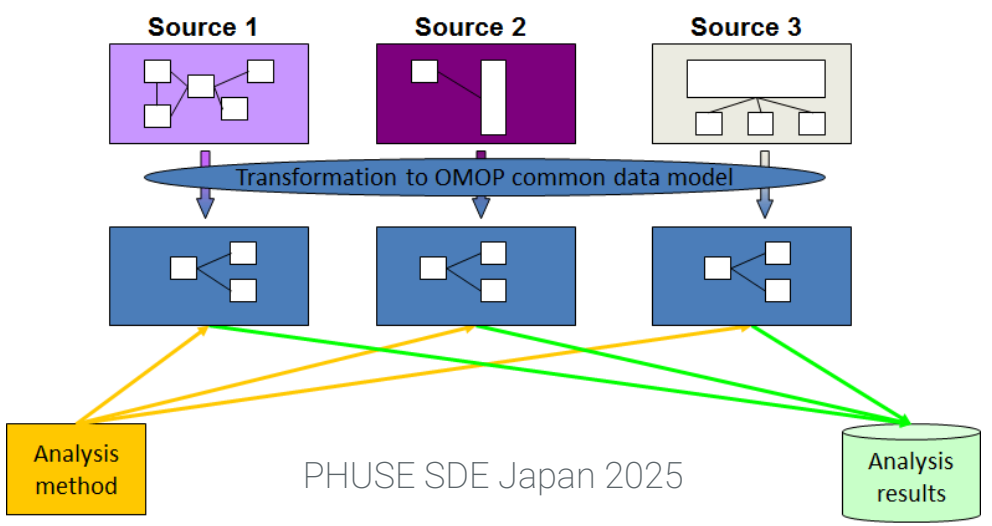
▶ OHDSI

Note

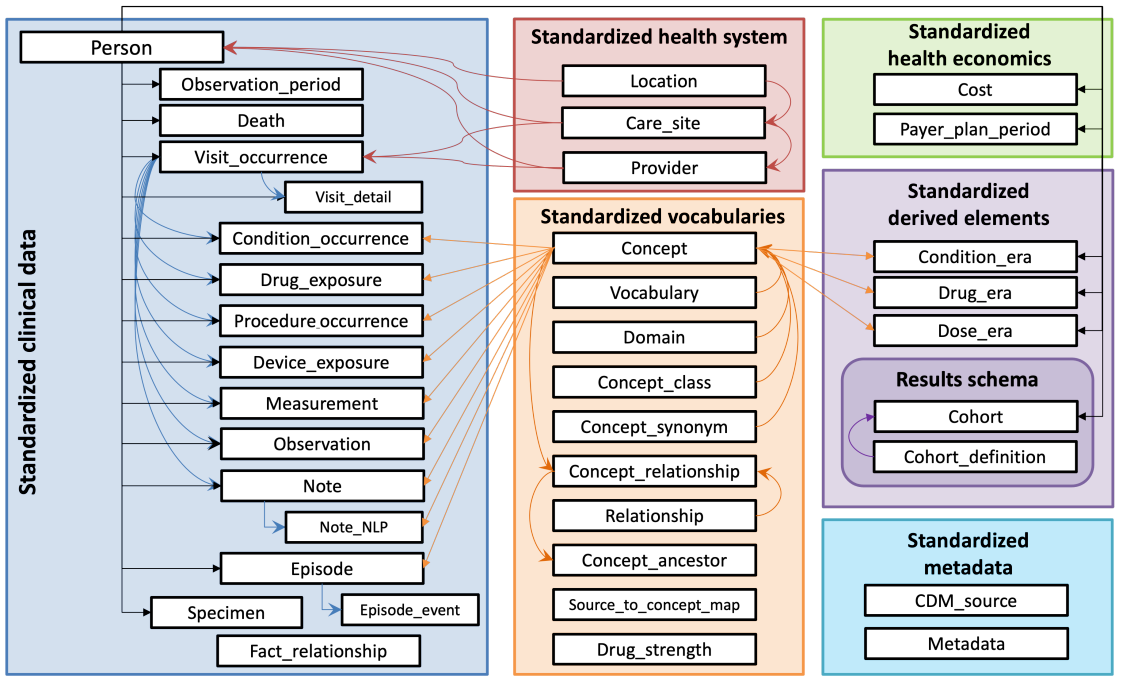
In 2025, **PHUSE Working Group** was established to promote OMOP adoption

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 terminologies



OMOP CDM Structure (v5.4)



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

▶ OMOP CDM

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

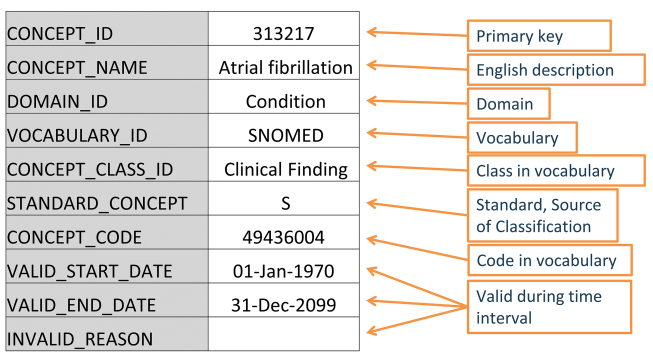
Drug exposure information

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



Example: Hypertension

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

Analyzing OMOP with R

What is HADES?

***H**ealth **A**nalytics **D**ata-to-**E**vidence **S**uite*

- 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...

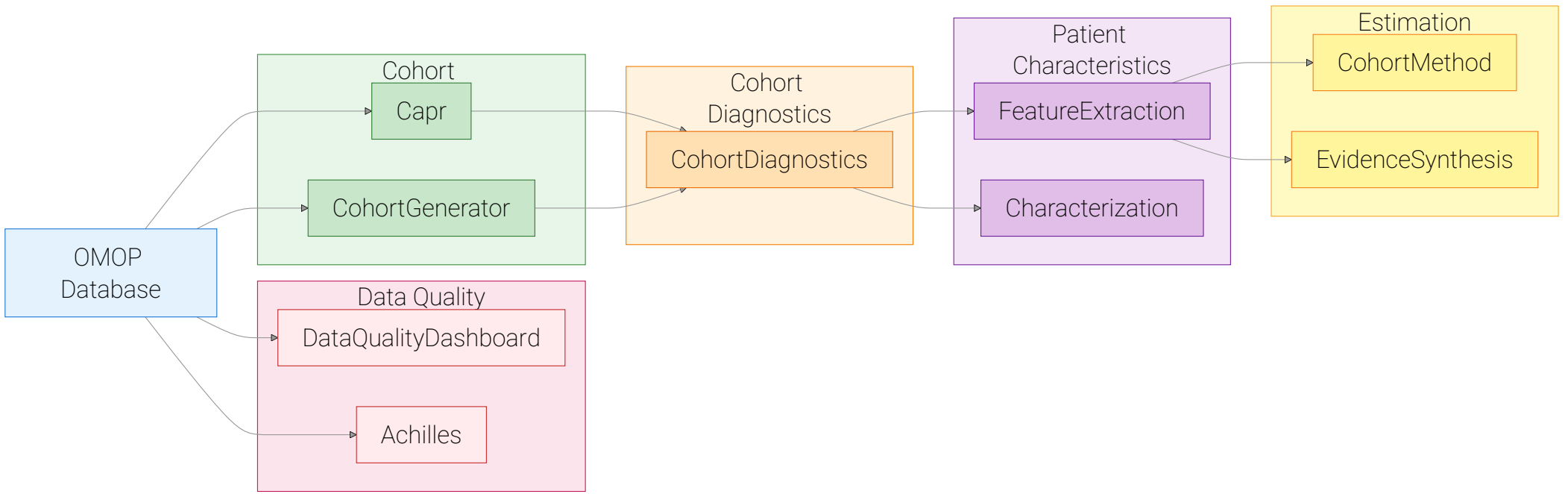
EnsemblePatientLevelPrediction
Building and validating ensemble patient-level predictive models.
Learn more...

▶ 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

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

Download sample data

```
library(CDMConnector)

Sys.setenv("EUNOMIA_DATA_FOLDER" = here::here())
downloadEunomiaData("GiBleed")
```

CDMConnector + Basics

▶ CDMConnector

CDMConnector

Database connection and data access

```r
library(CDMConnector)
library(tidyverse)
library(dbplyr)

# Connect to database
con <- DBI::dbConnect(duckdb::duckdb(), eunomiaDir("GiBleed"))

# List tables
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

```
cdm <- cdmFromCon(con, cdmSchema = "main", writeSchema = "main")
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, 0, …
$ location_id                <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ provider_id                <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ care_site_id               <int> NA, 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", "F", "M", "F…
$ gender_source_concept_id   <int> 0, 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

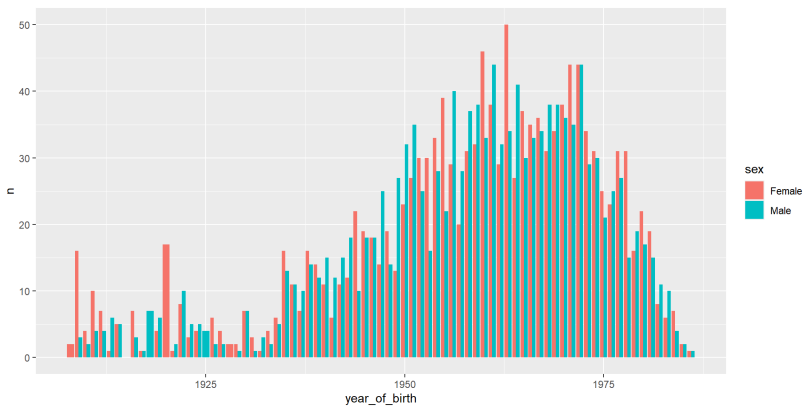
Distribution of conditions among males born after 1975

```
cdm$person |>
  filter(year_of_birth >= 1975, gender_source_value == "M") |>
  left_join(cdm$condition_occurrence, by = "person_id") |>
  summarise(n = n(), .by = condition_concept_id) |>
  left_join(cdm$concept |> select(concept_id, concept_name), by = c("condition_concept_id" = "concept_i
  collect() |>
  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!

```
cdm$person |>
  summarize(n = n(), .by = c(year_of_birth, gender_concept_id)) |>
  mutate(sex = case_when(
    gender_concept_id == 8532 ~ "Female",
    gender_concept_id == 8507 ~ "Male"
  )) |>
  collect() |>
  ggplot(aes(y = n, x = year_of_birth, fill = sex)) +
  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)

```
library(OmopSketch)

cdm |>
  summariseOmopSnapshot() |>
  tableOmopSnapshot(type = "tibble")
```

```
# A tibble: 13 × 3
   Variable           Estimate               [header_name]Database name\n[hea…¹
   <chr>              <chr>                  <chr>
 1 General            Snapshot date          2025-11-20
 2 General            Person count           2,694
 3 General            Vocabulary version     v5.0 18-JAN-19
 4 Observation period N                      5,343
 5 Observation period Start date             1908-09-22
 6 Observation period End date               2019-07-03
 7 Cdm                Source name            Synthea synthetic health database
 8 Cdm                Version                v5.3.1
 9 Cdm                Holder name            OHDSI Community
10 Cdm                Release date           2019-05-25
11 Cdm                Description            SyntheaTM is a Synthetic Patient …
12 Cdm                Documentation reference https://synthetichealth.github.io…
13 Cdm                Source type            duckdb
# ℹ abbreviated name: ¹`[header_name]Database name\n[header_level]Synthea`
```

OmopSketch

Get an overview of `condition_occurrence` (flextable)

```
cdm |>
  summariseClinicalRecords("condition_occurrence") |>
  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)

```
cdm |>
  summariseClinicalRecords("drug_exposure") |>
  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

PatientProfiles

Define a cohort of “patients with bronchitis”

```r
cdm <- cdm |>
  generateConceptCohortSet(
  name = "bronchitis",
  conceptSet = list("any_bronchitis" = c(260139, 258780)),
  limit = "all",
  end = 0
)

cdm$bronchitis |>
  collect()
```

```
# A tibble: 8,232 × 4
  cohort_definition_id subject_id cohort_start_date cohort_end_date
                 <int>      <int> <date>            <date>
1                    1        334 1944-09-16        1944-09-16
2                    1        376 1986-12-31        1986-12-31
3                    1        431 1964-11-03        1964-11-03
4                    1        727 1917-04-10        1917-04-10
# ℹ 8,228 more rows
```

PatientProfiles

Add patient characteristics to cohort

```
library(PatientProfiles)

# Date of birth, age, sex
cdm$bronchitis |>
  addDateOfBirth() |>
  addSex() |>
  addAge()

# Prior/future observation periods from index date
cdm$bronchitis |>
  addPriorObservation() |>
  addFutureObservation()
```

Note

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

IncidencePrevalence

Calculating Prevalence and Incidence

▶ IncidencePrevalence

IncidencePrevalence

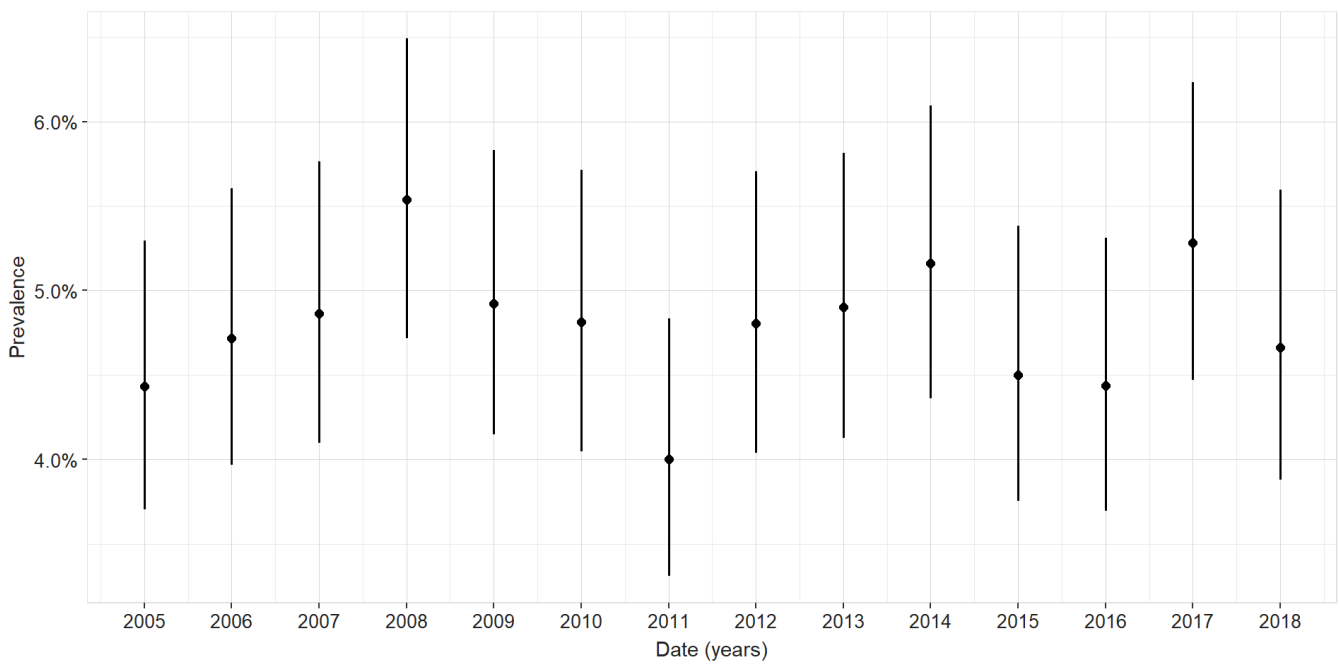
Create “denominator” cohort

```r
library(IncidencePrevalence)

cdm <- cdm |>
  generateDenominatorCohortSet(
    "denom",
    cohortDateRange = c(as.Date("2005-01-01"), as.Date(NA))
  )
```

IncidencePrevalence

Calculate prevalence

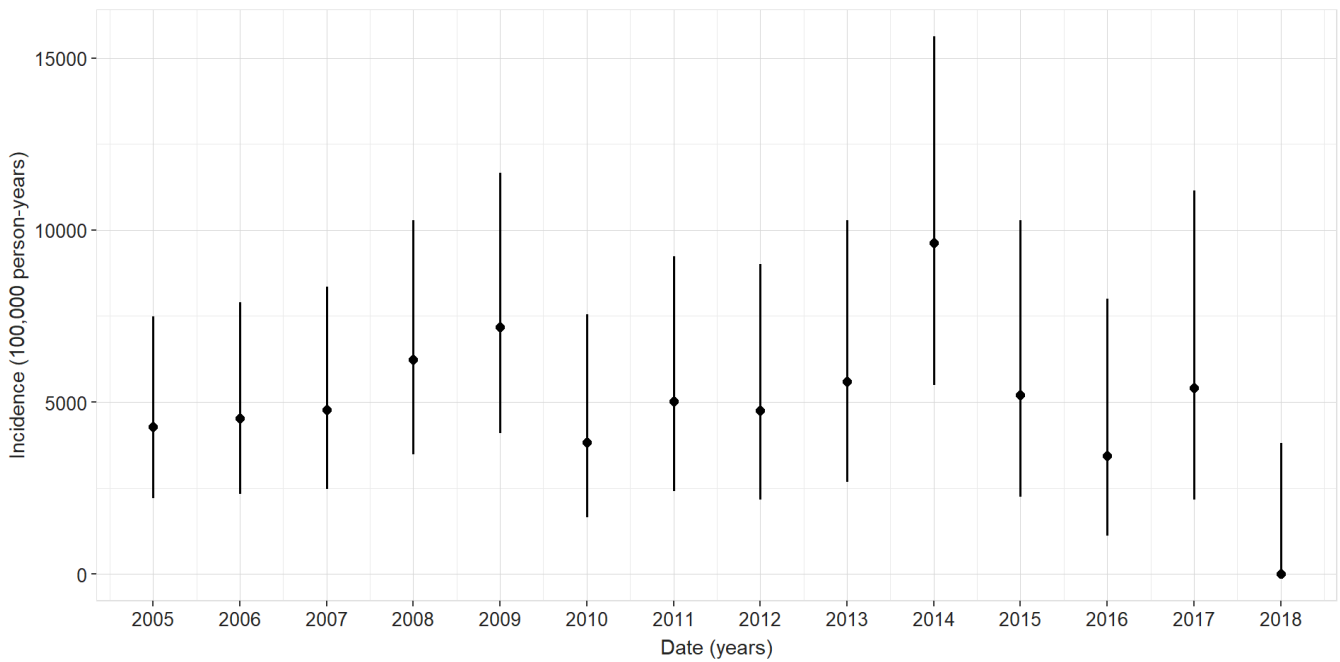
```
cdm |>
  estimatePeriodPrevalence(
    denominatorTable = "denom",
    outcomeTable = "bronchitis"
  ) |>
  plotPrevalence()
```



IncidencePrevalence

Calculate incidence

```
cdm |>
  estimateIncidence(
    denominatorTable = "denom",
    outcomeTable = "bronchitis"
  ) |>
  plotIncidence()
```



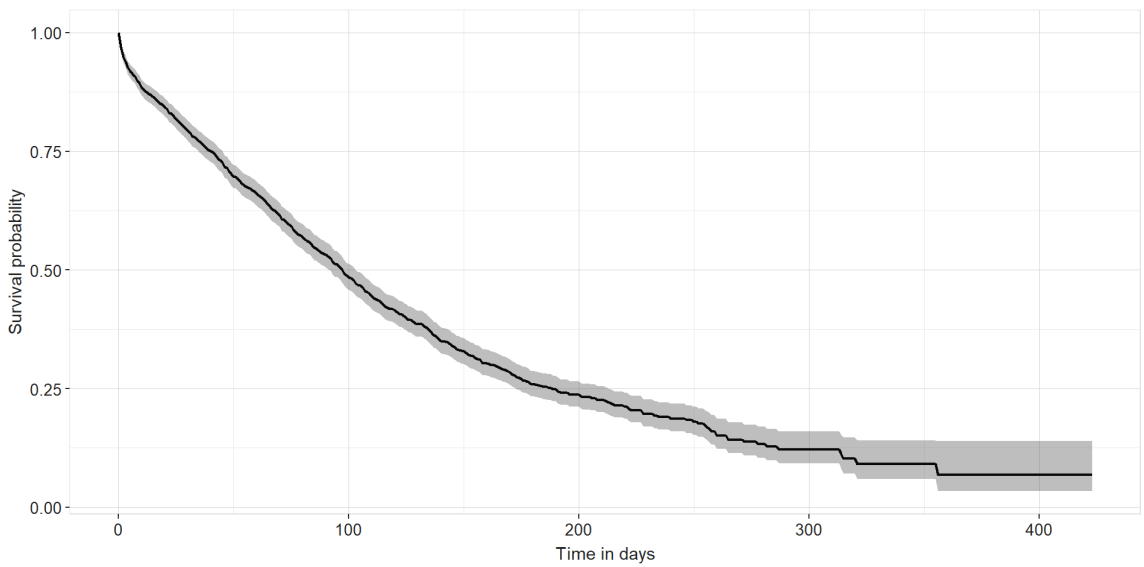
CohortSurvival

Survival Analysis

▶ CohortSurvival

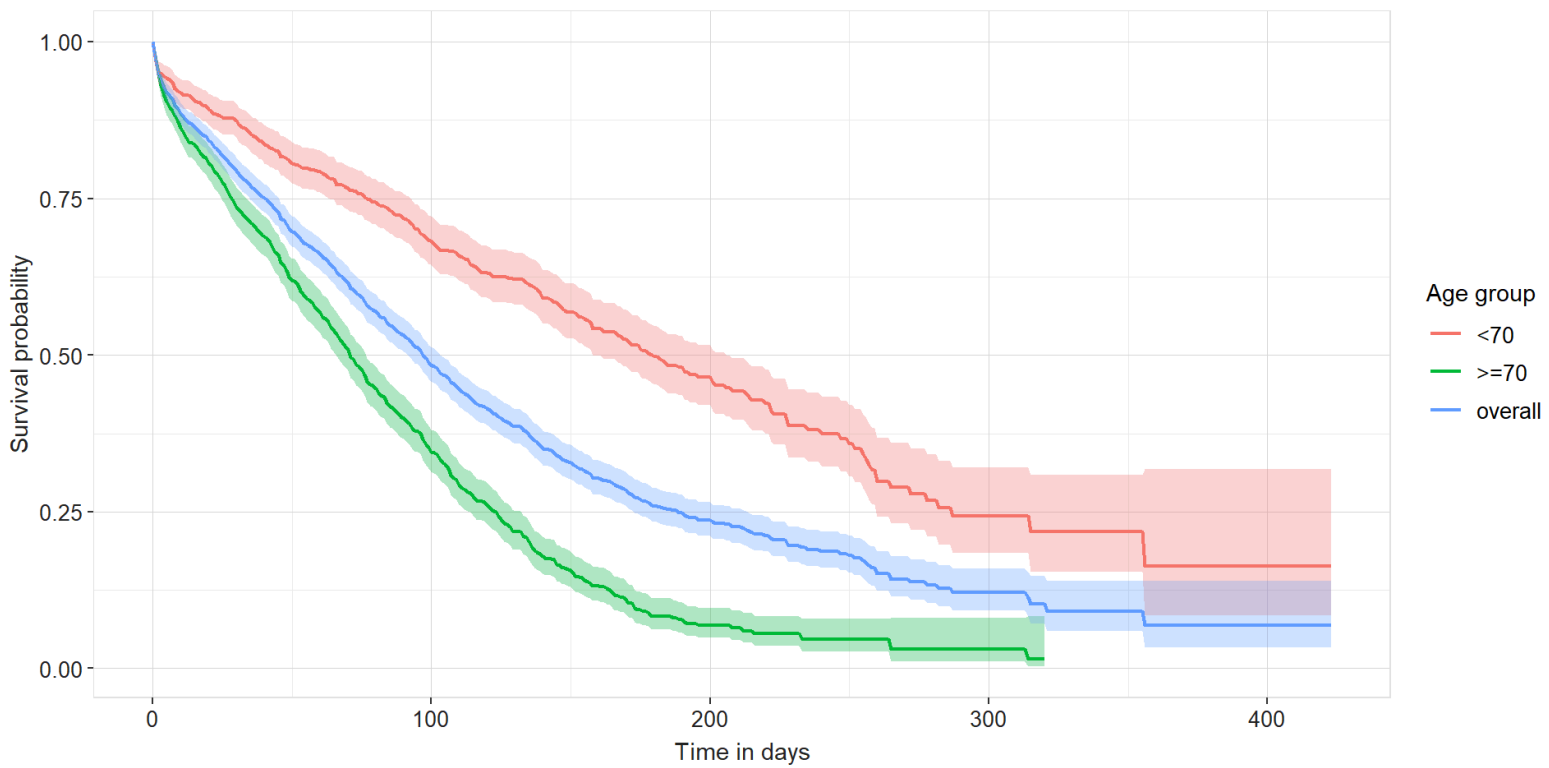
CohortSurvival

```r
library(CohortSurvival)

# Sample data for survival analysis
cdm <- mockMGUS2cdm()

cdm |>
  estimateSingleEventSurvival(
  targetCohortTable = "mgus_diagnosis",
  outcomeCohortTable = "death_cohort"
) |>
  plotSurvival()
```



CohortSurvival

```
cdm |>
  estimateSingleEventSurvival(
    targetCohortTable = "mgus_diagnosis",
    outcomeCohortTable = "death_cohort",
    strata = list(c("age_group"))
) |>
  plotSurvival(colour = "age_group")
```



Summary

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