

OSS を用いた RWD 解析

OMOP CDM と分析用 R パッケージの紹介

PHUSE Japan Open-source Technology Working Group

2025年12月5日

スライドはこちら

English:

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



Japanese:

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



OMOP CDM 入門

OMOP CDMとは？

Observational Medical Outcomes Partnership Common Data Model

- 異なるRWDを統一的に解析するためのデータモデル
- 共通スキーマと標準ボキャブラリによる標準化で、再現性を高める
- OHDSIというコミュニティによって開発・発展



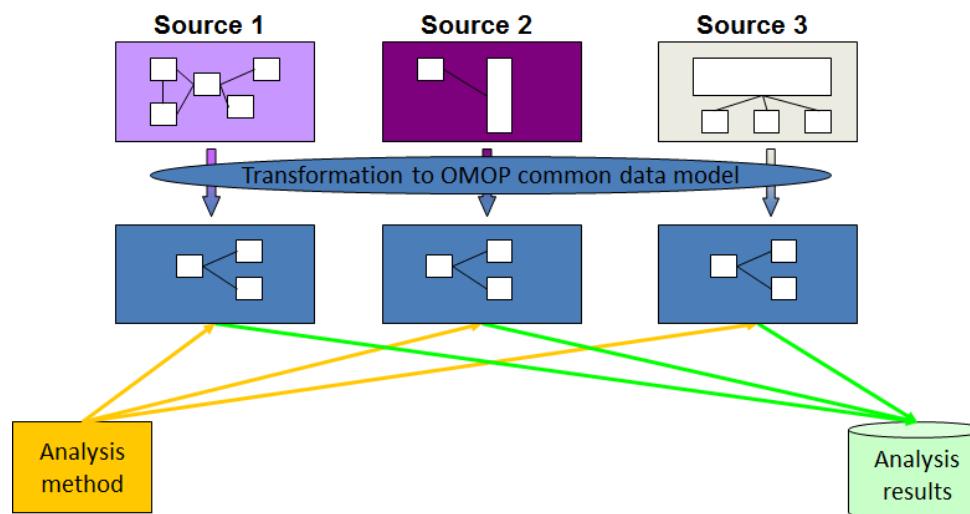
► OHDSI

① ノート

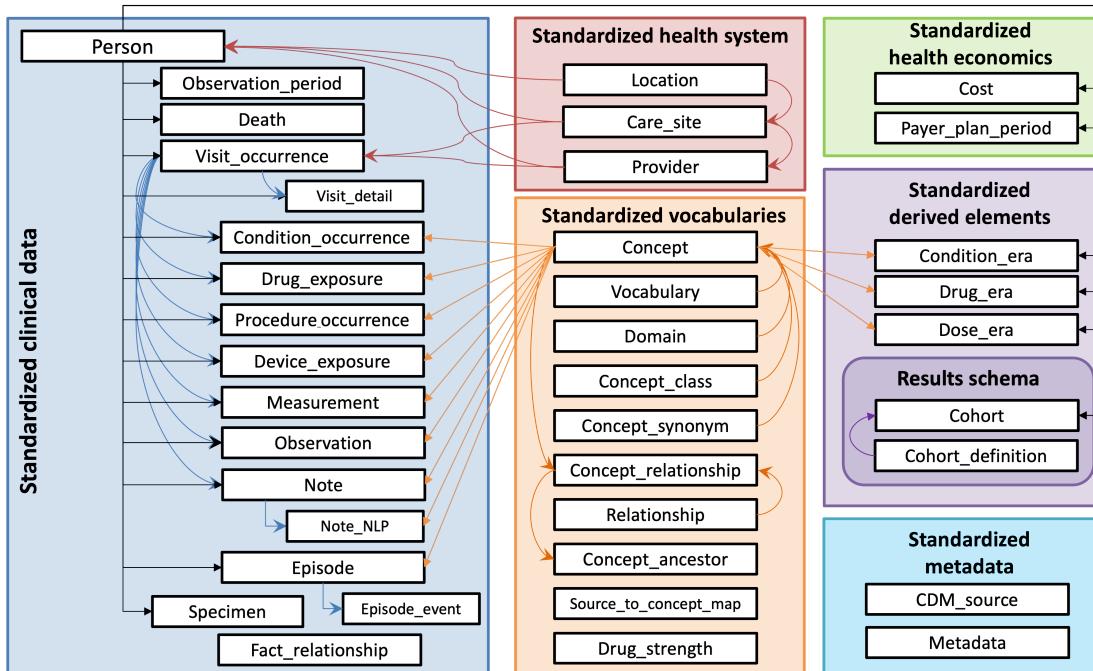
2025年には製薬業界でのOMOP活用を推進する**PHUSE Working Group**が発足

OMOP CDMの特徴

- **利点: 異なるデータソースの比較研究が容易**
 - 標準ボキャブラリによる用語統一
 - 観察研究に必要な最小限のテーブルで設計
- **課題: 変換工程の複雑さ**
 - データソースごとにETLプロセスが必要
 - 標準ボキャブラリへのマッピングの難易度



OMOP CDMの構造 (v5.4)



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

主要テーブル: Person

患者の基本情報と人口統計学的データ

フィールド	説明
person_id	患者ID
gender_concept_id	性別
year_of_birth	生年
race_concept_id	人種
ethnicity_concept_id	民族



全ての臨床イベントは **person_id** で紐付けられる

主要テーブル: Visit Occurrence

医療機関への来院・入院情報

フィールド	説明
visit_occurrence_id	来院の識別子
person_id	患者ID
visit_concept_id	来院タイプ（入院/外来/救急）
visit_start_date	来院開始日
visit_end_date	来院終了日

主要テーブル: Condition Occurrence

疾患・症状の診断情報

フィールド	説明
condition_occurrence_id	診断の識別子
person_id	患者ID
condition_concept_id	疾患の標準コンセプトID
condition_start_date	診断開始日
condition_type_concept_id	レコードの出所 (EHR/レセプト)

主要テーブル: Drug Exposure

薬剤曝露の情報

フィールド	説明
drug_exposure_id	薬剤曝露の識別子
person_id	患者ID
drug_concept_id	薬剤の標準コンセプトID
drug_exposure_start_date	曝露開始日
drug_exposure_end_date	曝露終了日

コードのマッピング

各種コードを標準コンセプトにマッピングすることが多い(必須ではない)

- **標準コンセプト**: SNOMED CT(疾患)やRxNorm(薬剤)などによって定義
- **非標準コンセプト**: ICD10やLOINCなど、元データで使用されるコード
- コンセプトは**ATHENA**というWebツールで検索可能

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		

例: 高血圧

- **標準**: SNOMED 38341003
- **非標準**: ICD10 I10, MeSH D006973

RでOMOPを解析する

HADESとは

Health Analytics Data-to-Evidence Suite

- OMOP CDMデータの分析に特化したRパッケージ群
- 相互運用性が高い (HADESでそろえればうまく動く)
- OHDSIとDARWIN EUという2つの組織を中心に開発が進んでいる



HADES
HEALTH ANALYTICS DATA-TO-EVIDENCE SUITE

► HADES

► OHDSI

► DARWIN EU

HADESのパッケージリスト

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

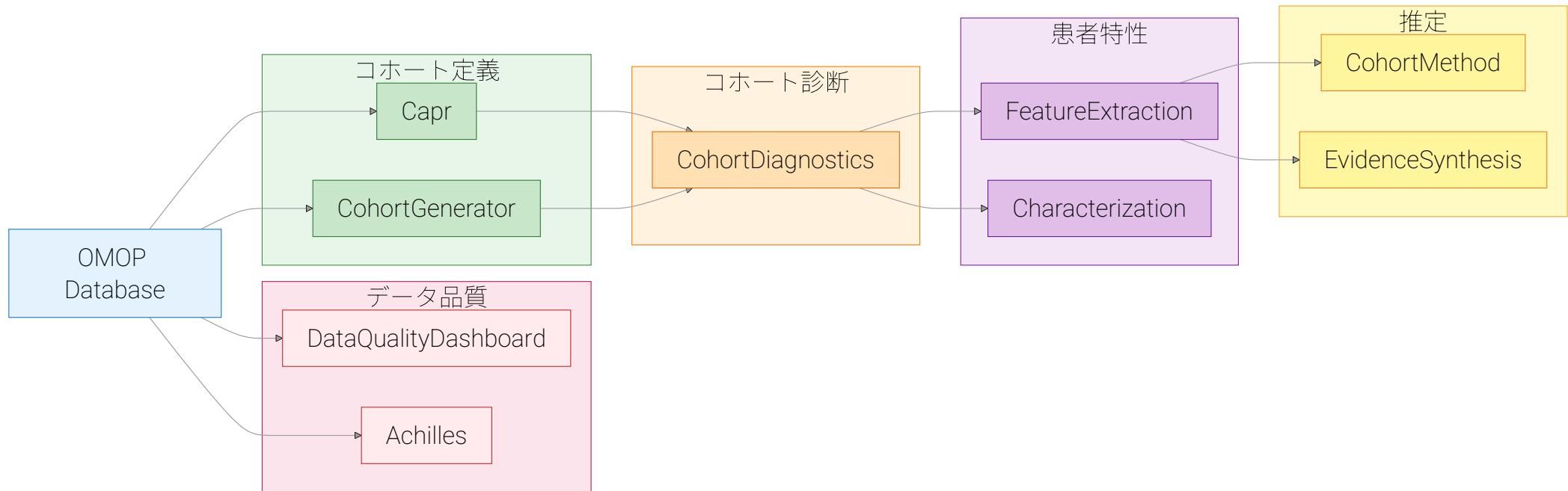
PatientLevelPrediction Build and evaluate predictive models for user-specified outcomes, using a wide array of machine learning algorithms. Learn more...	DeepPatientLevelPrediction Performing patient level prediction using deep learning Learn more...	EnsemblePatientLevelPrediction Building and validating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

▶ HADES Packages

2025年12月現在、41個ものパッケージがHADESに登録されている！

HADESを用いたワークフローの一例

一気通貫の分析フローが実現可能



実際にやってみよう



準備

R パッケージのインストール

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

サンプルデータのダウンロード

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

CDMConnector + 基本的な操作



▶ CDMConnector

CDMConnector

データベース接続とデータアクセス

```
1 library(CDMConnector)
2 library(tidyverse)
3 library(dbplyr)
4
5 # データベースへの接続
6 con <- DBI::dbConnect(duckdb::duckdb(), eunomiaDir("GiBleed"))
7
8 # テーブル一覧を表示
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

CDMオブジェクトの作成

`cdmFromCon()`を使って、OMOP専用のオブジェクト形式にする

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

cdm オブジェクト

\$を使って各テーブルにアクセス可能

```
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, ...
$ provider_id                 <int> NA, ...
$ care_site_id                 <int> 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, ...
$ 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, ...
```

基本的なデータ操作

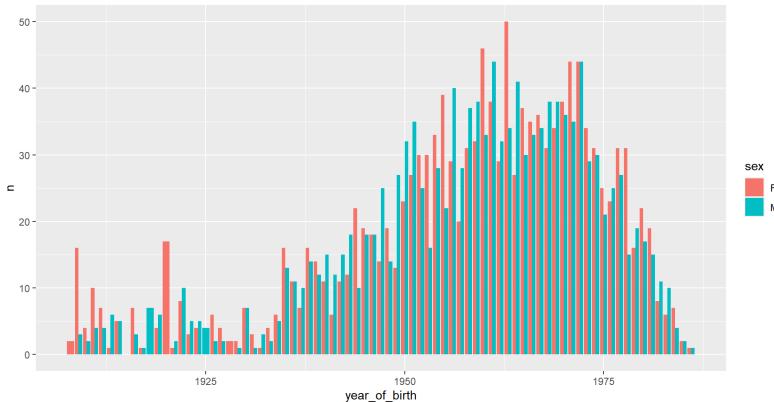
1975年以降生まれの男性の疾患分布

```
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
```

可視化も！

```
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")
```



① ノート

tidyverse スタイルのデータハンドリングが可能！

OmopSketch

データベースの概要をつかむ



▶ OmopSketch

OmopSketch

DB全体の概要をつかむ (tibble)

```
1 library(OmopSketch)
2
3 cdm |>
4 summariseOmopSnapshot() |>
5 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

condition_occurrence の概要をつかむ (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

drug_exposure の概要をつかむ (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

患者特性の追加



► PatientProfiles

Patient Profiles

「気管支炎を有する患者」のコホートを設定

```
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         334 1963-11-22  1963-11-22
2                   1         384 2008-02-10  2008-02-10
3                   1         431 1992-05-11  1992-05-11
4                   1         727 1927-04-03  1927-04-03
# i 8,228 more rows
```

PatientProfiles

コホートに患者特性を追加

```
1 library(PatientProfiles)
2
3 # 生年月日、年齢、性別
4 cdm$bronchitis |>
5   addDateOfBirth() |>
6   addSex() |>
7   addAge()
8
9 # index dateを起点とした過去/未来の観察期間
10 cdm$bronchitis |>
11   addPriorObservation() |>
12   addFutureObservation()
```

i ノート

特に何か言うことはありません。簡単すぎる！

IncidencePrevalence

有病割合・罹患率の計算



▶ IncidencePrevalence

IncidencePrevalence

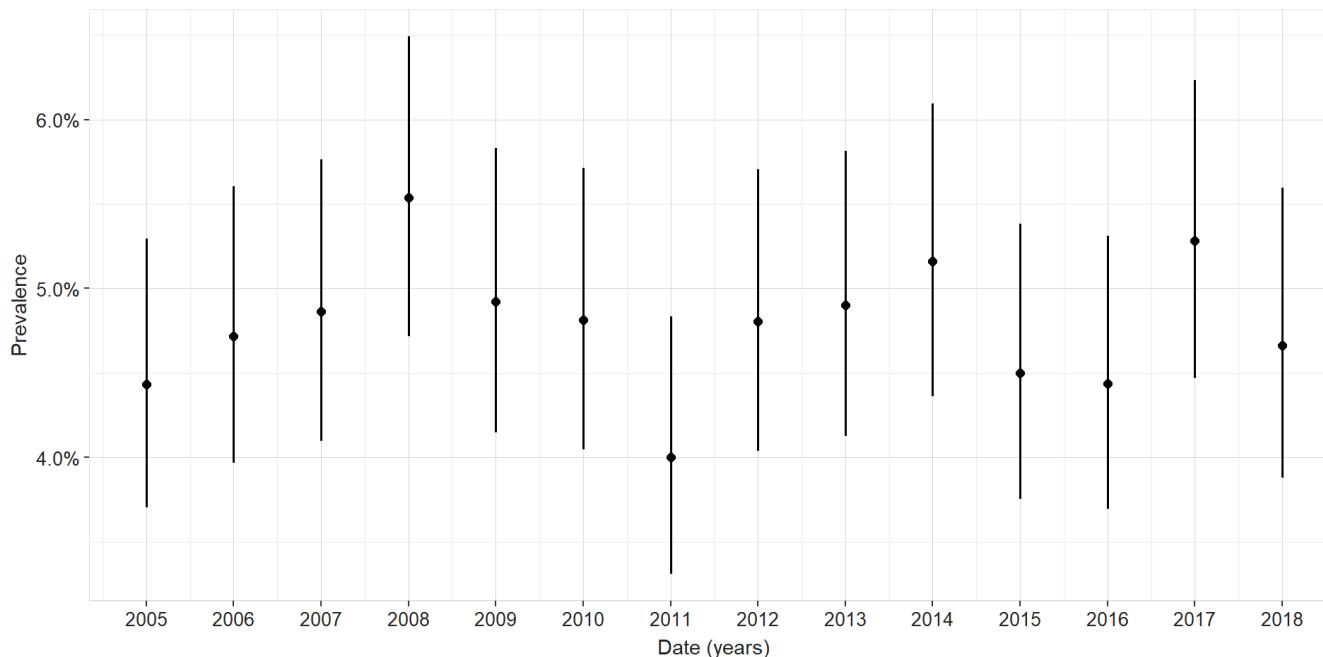
「分母」となるコホートの作成

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

IncidencePrevalence

有病割合の算出

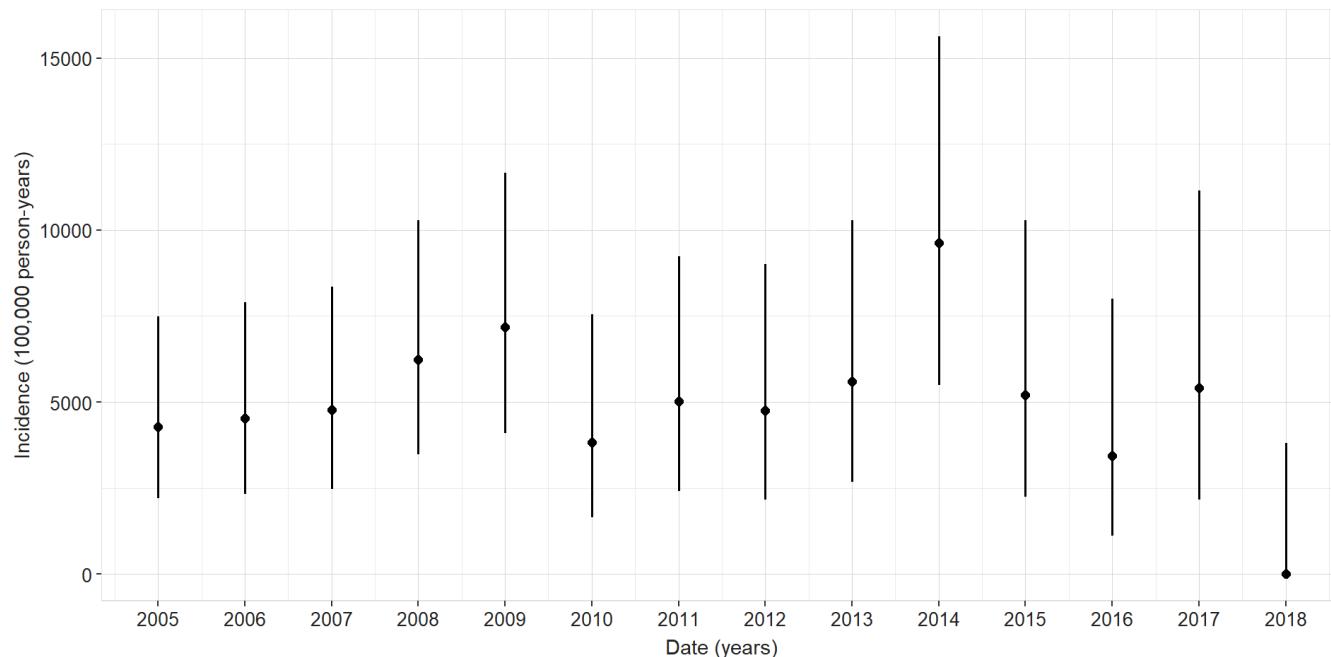
```
1 cdm |>
2   estimatePeriodPrevalence(
3     denominatorTable = "denom",
4     outcomeTable = "bronchitis"
5   ) |>
6   plotPrevalence()
```



IncidencePrevalence

罹患率の算出

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



CohortSurvival

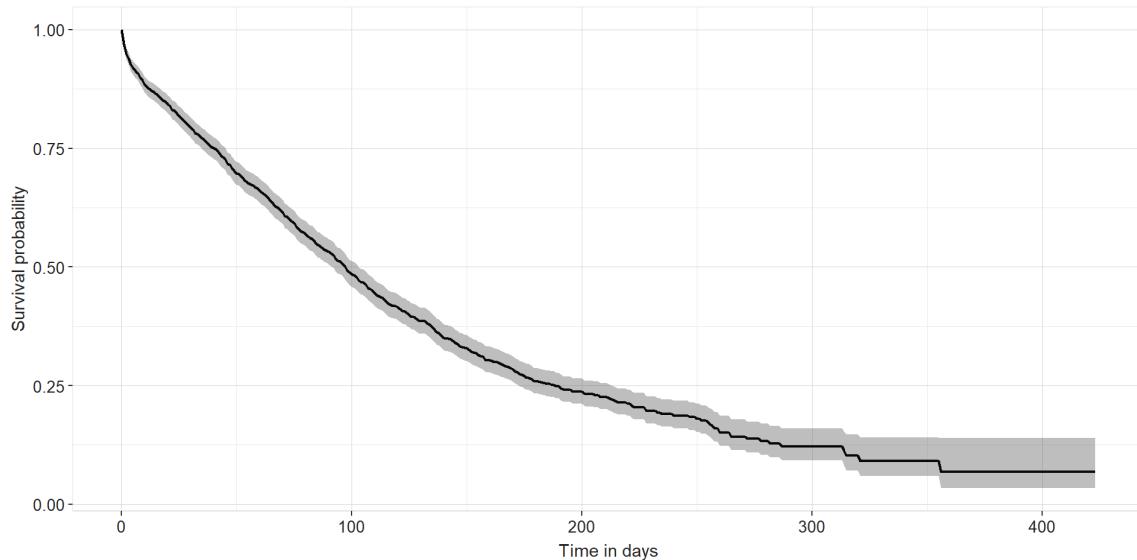
生存時間分析



► CohortSurvival

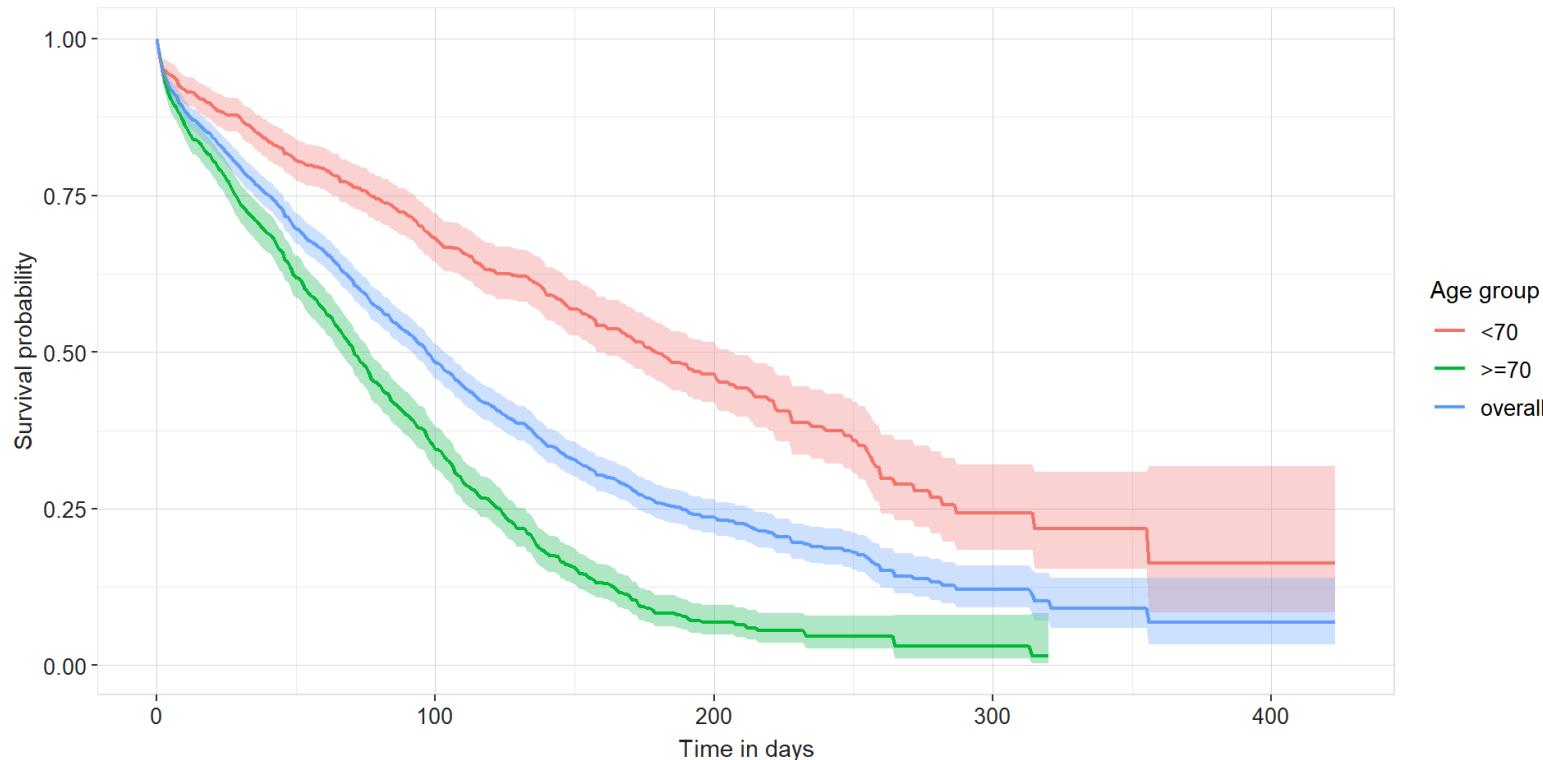
CohortSurvival

```
1 library(CohortSurvival)
2
3 # 生存時間分析のためのサンプルデータ
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")
```



まとめ

- **OMOP CDM**
 - 観察研究・RWD研究のための共通データモデル
 - 異なるデータベースを標準化し、再現性の高い分析を可能に
- **解析用Rパッケージ**
 - HADESを中心としたエコシステム
 - OMOP解析に特化した便利なパッケージが多く存在

学習リソース

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

コミュニティ

- OHDSI Japan
- OHDSI Forums
- PHUSE OSS Technology WG