The Book of OHDSI Korea

OHDSI-Korea

2019-08-11

Contents

Chapter 1

Introduction

. , , .
, , .

, (Common Data Model, CDM) (Distributed Research Network, DRN) . CDM / DRN , (OMOP-CDM) (OHDSI network) .

OHDSI (open-source research network) ,
.

, OHDSI .

Chapter 2

The OHDSI Community

Seng Chan You

2.1 OHDSI

OHDSI . OHDSI .

2.2

2.2.1 (Distributed Research Network)

, , , .
, , .
. .
, / (, , , ,) .
, . .

2.2.2 (Common Data Model)

, . .
. .
.
. ETL(Extract, Transform, Load) .
.
(Observational Health Data and Informatics, OHDSI) FDA (Sentinel CDM), (The National Patient-Centered Clinical Outcomes Research Network, PCORnet) .
OHDSI . OHDSI 2008 Observational Medical Outcomes Partnership(OMOP) .
, , , .
OMOP 2013 , OMOP CDM OHDSI . OMOP , OHDSI , , , , .

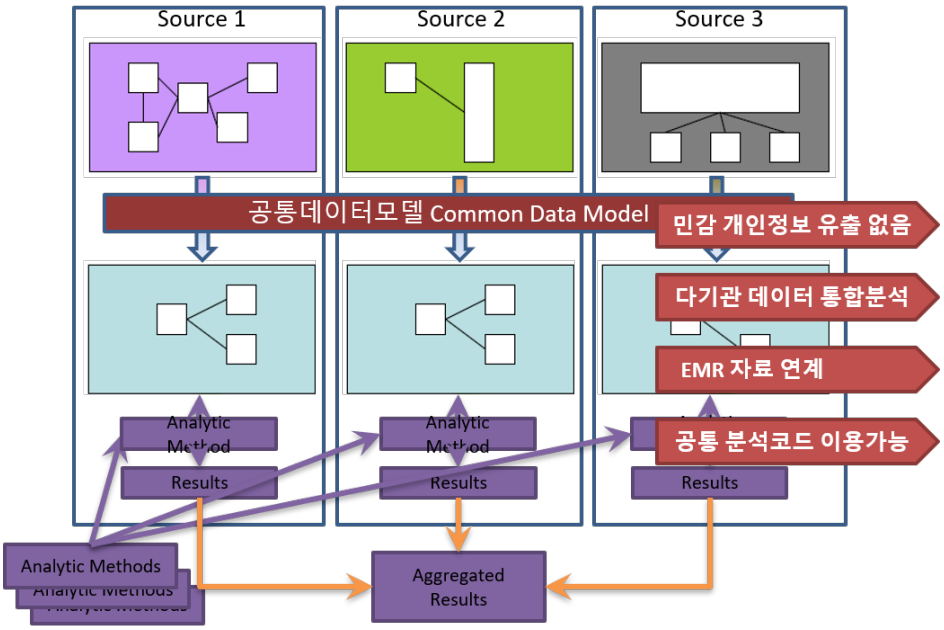


Figure 2.1: Distributed Research Network

CDM Sentinel Initiative (Food and Drug Administration, FDA) . Sentinel . FDA FDA .

Sentinel . Sentinel CDM . Sentinel CDM . . Sentinel Distributed Database(SDD) .

CDM PCORnet The Patient-Centered Outcomes Research Institute(PCORI) 2013 (Electronic health records, EHR) (Comparative effectiveness research, CER) . 50 11 (Clinical data research networks, CDRNs) 18 (Patient-powered research networks, PPRNs) . PCORnet (patient-centered approach) .

2.3 (OHDSI network)

(The Observational Health Data Sciences and Informatics, OHDSI network) OMOP (Observational Medical Outcomes Partnership) . (CDM) (Distributed Research Network) , OHDSI . OHDSI OMOP-CDM , , , .

2.4 OHDSI

2008 (FDA) , , OMOP (Observational Medical Outcomes Partnership) ref. 2009 OMOP-CDM version 1 ref. OMOP common data model (CDM) (claim data) (eletronic heatlh record) , 2013 Reagan-Udall . FDA OMOP , Columbia (coordinating center), George Hripsack (OHDSI) . ODYSSEY , OHDSI . 2014 Columbia Face-to-Face (F2F meeting) 2015 (Washington DC) . (Bethesda) . OHDSI .

Figure 2.2: OHDSI International Symposium 2017 in Korea

Figure 2.3: OHDSI International Symposium 2017 in Korea

2.4.1

2014 OMOP-CDM , 2015 OHDSI .
OMOP-CDM, OHDSI , 2016 OHDSI committee Korean chapter , OHDSI
. 2017 3 , 3 .

OHDSI 2017 3 7 2 / OHDSI OHDSI .

2.5 , ,

OHDSI mission, vision, value page .

2.5.1 OHDSI

.

> To improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.

2.5.2 OHDSI

.

Figure 2.4: Tutorial in the OHDSI International Symposium 2017

A world in which observational research produces a comprehensive understanding of health and disease.

2.5.3 OHDSI

- **Innovation**: .

  Observational research is a field which will benefit greatly from disruptive thinking. We actively seek and encourage fresh methodological approaches in our work.

- **Reproducibility**: , , .

  Accurate, reproducible, and well-calibrated evidence is necessary for health improvement.

- **Community**: .

  Everyone is welcome to actively participate in OHDSI, whether you are a patient, a health professional, a researcher, or someone who simply believes in our cause.

- **Openness**: , , , .

  We strive to make all our community's proceeds open and publicly accessible, including the methods, tools and the evidence that we generate.

- **Collaboration**: , .

  We work collectively to prioritize and address the real world needs of our community's participants.

- **Beneficence**: .

  We seek to protect the rights of individuals and organizations within our community at all times.

OHDSI , ‘ , ’ .

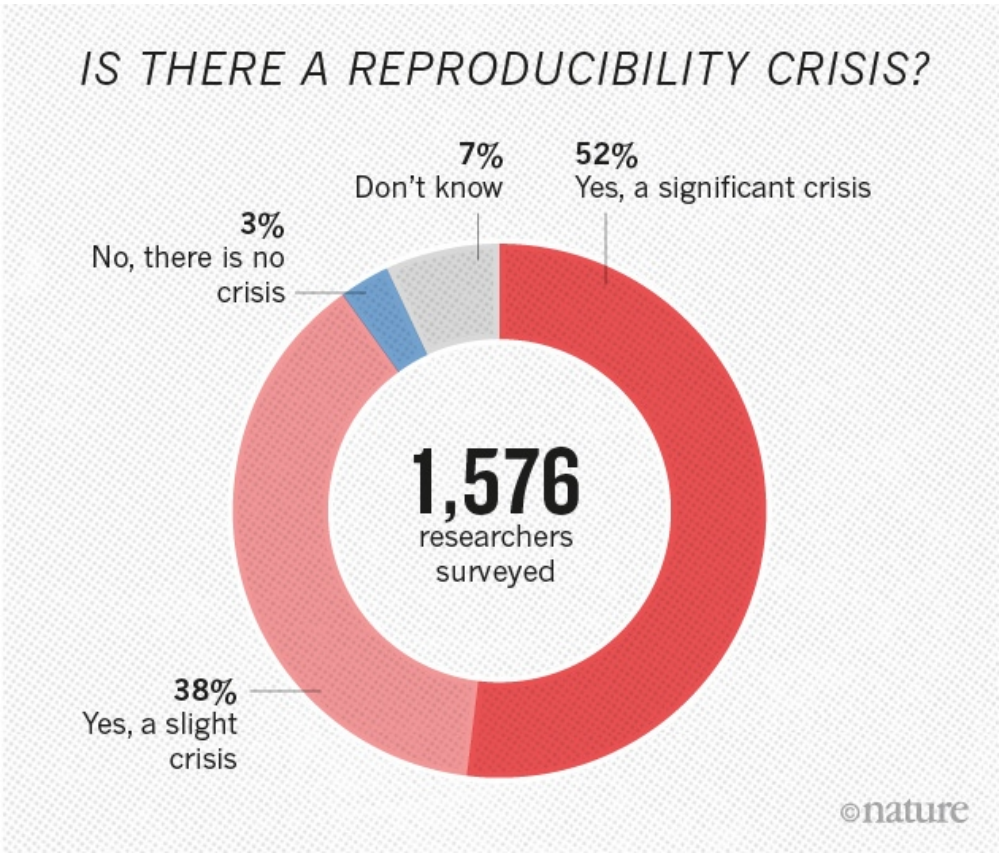


Figure 2.5: ?

2.6 (Open Science and Reproducible Research)

OHDSI CDM DRN , .

2.6.1 (Reproducibility Crisis)

‘ 20 ? *How on Earth could we have spent 20 yeares and hundreds of millions of dollars studying pure noise?*’ref 20 18 (major depression) , Matthew Keller . (false positive) . [ref, Richard Border et al., American Journal Of Psychiatry, 2019]

2016 Nature 1576 , 70% , 50% . 52% . ref

(Reproducible research) ’ (raw data) ’ .

, . PLOS Medicine 1/10 , (Discovery-oriented exploratory research with massive testing) 1000 1 [ref, *Ioannidis, 2005 PLOS Medicine*].

? 4 [ref, *Bishop, 2019, Nature*].

- Publication Bias
- Low Statistical Power

1 – β	*R*	*u*	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.
RCT, randomized controlled trial.
DOI: 10.1371/journal.pmed.0020124.t004

Figure 2.6: PPV of Research Findings for various combinations of power, ratio of Tru to Not-True Relationship, and Bias

- *P* *P*-value Hacking
- HARKing (Hypothesizing After Results are Known)

2.6.1.1 Publication Bias

. . ' *(file drawer effect)* .
, . [ref, , . 165]

p-value hacking HARKing . impact factor (IF) ,
IF .

2.6.1.2 Low Statistical Power

(small sample size) (small effect), . (underpowered)
, [ref, R. G. Newcombe Br. Med. J. (Clin. Res. Ed.) 295, 656–659; 1987] . Low
Statistical Power , *p*-value hacking HARKing . IF
.

2.6.1.3 *P* *P*-value Hacking and HARKing (Hypothesizing After Results are Known)

P-value hacking HARKing , 'Discovery-oriented exploratory research with massive testing' . IF , *P*-value ($P < 0.05$) ,
. 'False-Positive Psychology' Simmons 4 (, , ,) '
, [ref, Simmons et al., 2011 Psychological Science]. Hack Your Ways To Scientific Glory *P*-hacking , testing 0.05 *P* .

2.6.2

Simmons , 6 (disclosure based solution) . , ,
, .

```
1. Authors must decide the rules for the terminating data collection before data collectino begins and r
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data coll
3. Authors must list all variables collected in a study
4. Authors must report all experimental conditions, including failed manipulations
5. If observation are elimiated, authors must also report what the statistical results are if those obse
6. If an analysis includes a covariate, authors must report the statistical results of the analysis with
```

OHDSI ' ' . . OHDSI .
OHDSI Studies GitHub OHDSI Study Protocol Github .

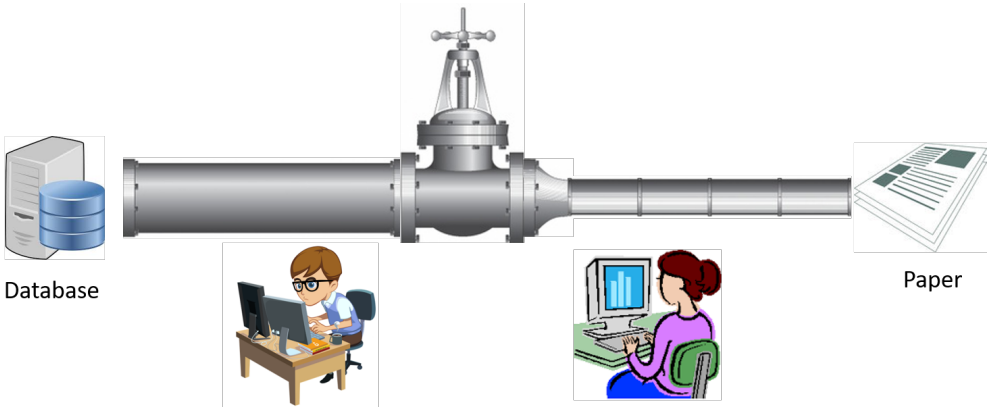


Figure 2.7: An OHDSI study shoul be look like a pipeline

Chapter 3

The OMOP-CDM

- OMOP-CDM : CDM wiki page OMOP-CDM .
- OMOP-CDM Github : OMOP-CDM github
- OMOP-CDM : OMOP-CDM OMOP-CDM .
- OHDSI tutorial : OHDSI past event OMOP-CDM tutorial .

3.1

OMOP-CDM . OMOP-CDM .

- **Suitability for purpose:** The CDM aims to provide data organized in a way optimal for analysis, rather than for the purpose of addressing the operational needs of health care providers or payers.
- **Data protection:** All data that might jeopardize the identity and protection of patients, such as names, precise birthdays etc. are limited. Exceptions are possible where the research expressly requires more detailed information, such as precise birth dates for the study of infants.
- **Design of domains:** The domains are modeled in a person-centric relational data model, where for each record the identity of the person and a date is captured as a minimum.
- **Rationale for domains:** Domains are identified and separately defined in an entity-relationship model if they have an analysis use case and the domain has specific attributes that are not otherwise applicable. All other data can be preserved as an observation in an entity-attribute-value structure.
- **Standardized Vocabularies:** To standardize the content of those records, the CDM relies on the Standardized Vocabularies containing all necessary and appropriate corresponding standard healthcare concepts.
- **Reuse of existing vocabularies:** If possible, these concepts are leveraged from national or industry standardization or vocabulary definition organizations or initiatives, such as the National Library of Medicine, the Department of Veterans' Affairs, the Center of Disease Control and Prevention, etc.
- **Maintaining source codes:** Even though all codes are mapped to the Standardized Vocabularies, the model also stores the original source code to ensure no information is lost.
- **Technology neutrality:** The CDM does not require a specific technology. It can be realized in any relational database, such as Oracle, SQL Server etc., or as SAS analytical datasets.
- **Scalability:** The CDM is optimized for data processing and computational analysis to accommodate data sources that vary in size, including databases with up to hundreds of millions of persons and billions of clinical observations.
- **Backwards compatibility:** All changes from previous CDMs are clearly delineated in the github repository (https://github.com/OHDSI/CommonDataModel). Older versions of the CDM can be easily created from the CDMv5, and no information is lost that was present previously.

3.2 Data Model Conventions

There are a number of implicit and explicit conventions that have been adopted in the CDM. Developers of methods that run against the CDM need to understand these conventions.

3.2.1 General conventions of the model

The OMOP CDM is considered a "person-centric" model, meaning that the people (or patients) drive the event and observation tables. At a minimum, the tables have a foreign key into the PERSON table and a date. This allows for a longitudinal view on all healthcare-relevant events by person. The exceptions from this rule are the standardized health system data tables, which are linked directly to events of the various domains.

3.2.2 General conventions of schemas

New to CDM v6.0 is the concept of schemas. This allows for more separation between read-only and writeable tables. The clinical data, event, and vocabulary tables are in the 'CDM' schema and are considered read-only to the end user. This means that the tables can be queried but no information can be accidentally removed or written over except by the database administrator. Tables that need to be manipulated by web-based tools or end users have moved to the 'Results' schema. Currently the only two tables in the 'Results' schema are COHORT and COHORT_DEFINITON, **add a sentence explaining that these tables describe groups of interest that the user might define, put in links to the later sections** though likely more will be added over the course of v6.0 point releases. These tables can be written to, meaning that a cohort created in ATLAS or by a user can be stored in the COHORT table and accessed at a later date. This does mean that cohorts in the COHORT table can be manipulated by anyone so it is always recommended that the SQL code used to create the cohort be saved along with the project or analysis in the event it needs to be regenerated.

3.2.3 General conventions of data tables

The CDM is platform-independent. Data types are defined generically using ANSI SQL data types (VARCHAR, INTEGER, FLOAT, DATE, DATETIME, CLOB). Precision is provided only for VARCHAR. It reflects the minimal required string length and can be expanded within a CDM instantiation. The CDM does not prescribe the date and datetime format. Standard queries against CDM may vary for local instantiations and date/datetime configurations.

In most cases, the first field in each table ends in '_ID', containing a record identifier that can be used as a foreign key in another table. For example, the CONDITION_OCCURRENCE table contains the field VISIT_OCCURRENCE_ID which is a foreign key to the VISIT_OCCURRENCE table where VISIT_OCCURRENCE_ID is the primary key.

3.2.4 General conventions of fields

Variable names across all tables follow one convention:

Notation	Description
_SOURCE_VALUE	Verbatim information from the source data, typically used in ETL to map to CONCEPT_ID, and not to be used by any standard analytics. For example, CONDITION_SOURCE_VALUE = '787.02' was the ICD-9 code captured as a diagnosis from the administrative claim.

Notation	Description
_ID	Unique identifiers for key entities, which can serve as foreign keys to establish relationships across entities. For example, PERSON_ID uniquely identifies each individual. VISIT_OCCURRENCE_ID uniquely identifies a PERSON encounter at a point of care.
_CONCEPT_ID	Foreign key into the Standardized Vocabularies (i.e. the standard_concept attribute for the corresponding term is true), which serves as the primary basis for all standardized analytics. For example, CONDITION_CONCEPT_ID = 31967 (http://athena.ohdsi.org/search-terms/terms/31967) contains the reference value for the SNOMED concept of 'Nausea'
_SOURCE_CONCEPT_ID	Foreign key into the Standardized Vocabularies representing the concept and terminology used in the source data, when applicable. For example, CONDITION_SOURCE_CONCEPT_ID = 45431665 (http://athena.ohdsi.org/search-terms/terms/45431665) denotes the concept of 'Nausea' in the Read terminology; the analogous CONDITION_CONCEPT_ID might be 31967, since SNOMED-CT is the Standardized Vocabulary for most clinical diagnoses and findings.
_TYPE_CONCEPT_ID	Delineates the origin of the source information, standardized within the Standardized Vocabularies. For example, DRUG_TYPE_CONCEPT_ID can allow analysts to discriminate between 'Pharmacy dispensing' and 'Prescription written'

3.2.5 Representation of content through Concepts

In CDM data tables the content of each record is represented using Concepts. Concepts are stored in event tables with their CONCEPT_IDs as foreign keys to the CONCEPT table, which contains Concepts necessary to describe the healthcare experience of a patient. If a Standard Concept does not exist or cannot be identified, the the CONCEPT_ID 0 is used, representing a non-existing concept or un-mappable source value.

Records in the CONCEPT table contain detailed information about each concept (name, domain, class etc.). Concepts, Concept Relationships, Concept Ancestors and other information relating to Concepts is contained in the tables of the Standardized Vocabularies.

3.2.6 Difference between Concept IDs and Source Values

Many tables contain equivalent information in multiple places: As a Source Value, a Source Concept and as a Standard Concept.

- Source Values contain the codes from public code systems such as ICD-9-CM, NDC, CPT-4, READ etc. or locally controlled vocabularies (such as F for female and M for male) copied from the source data. Source Values are stored in the _SOURCE_VALUE fields in the data tables.
- Concepts are CDM-specific entities that represent the meaning of a clinical fact. Most concepts are based on code systems used in healthcare (called Source Concepts), while others were created de-novo (CONCEPT_CODE = 'OMOP generated'). Concepts have unique IDs across all domains.
- Source Concepts are the concepts that represent the code used in the source. Source Concepts are only used for common healthcare code systems, not for OMOP-generated Concepts. Source Concepts are stored in the _SOURCE_CONCEPT_ID field in the data tables.
- Standard Concepts are those concepts that are used to define the unique meaning of a clinical entity. For each entity there is one Standard Concept. Standard Concepts are typically drawn from existing public vocabulary sources. Concepts that have the equivalent meaning to a Standard Concept are

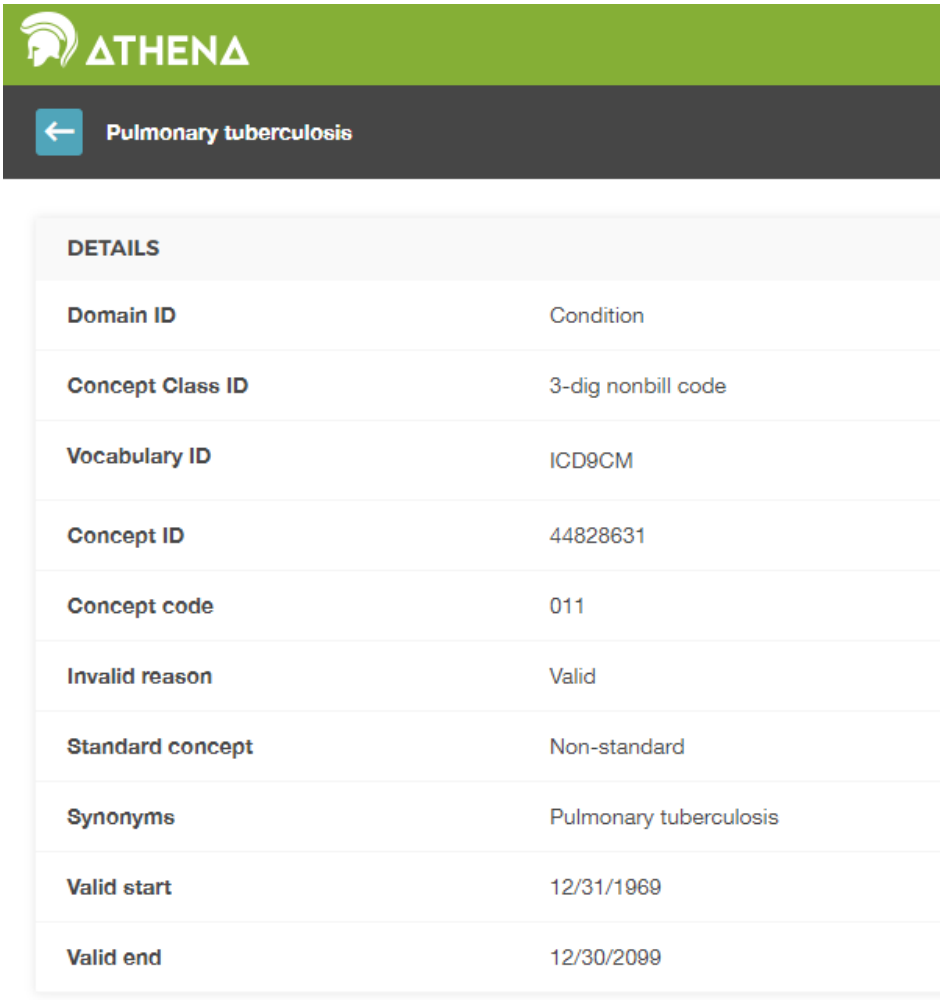
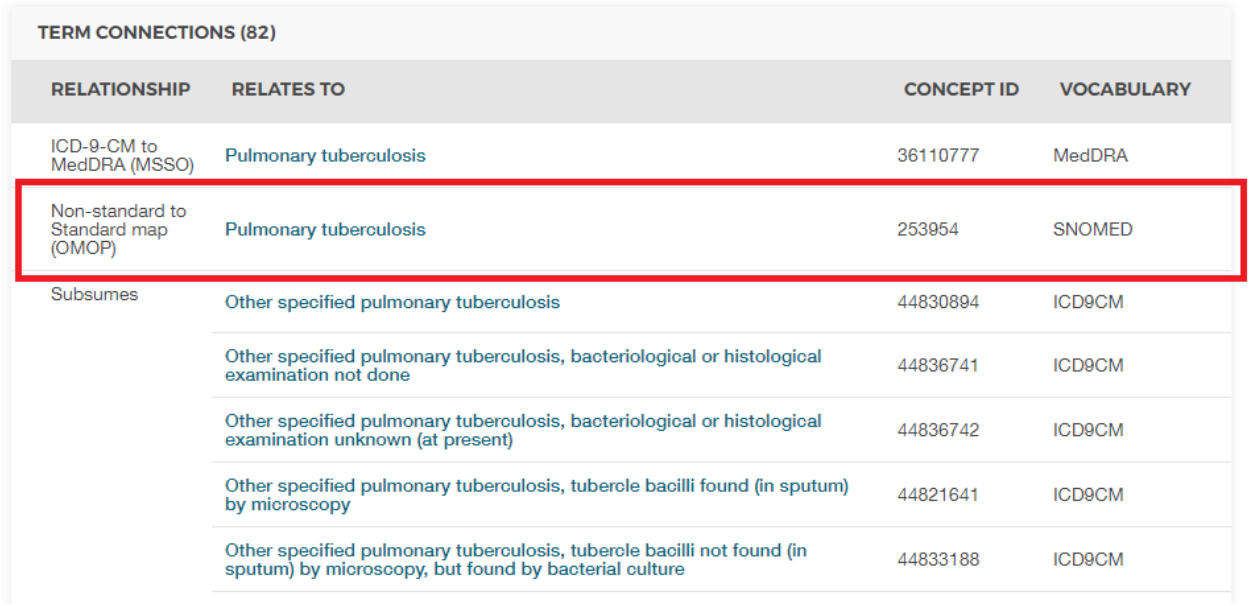


Figure 3.1: ICD9CM code for Pulmonary Tuberculosis

mapped to the Standard Concept. Standard Concepts are referred to in the _CONCEPT_ID field of the data tables.

Source Values are only provided for convenience and quality assurance (QA) purposes. Source Values and Source Concepts are optional, while **Standard Concepts are mandatory**. Source Values may contain information that is only meaningful in the context of a specific data source. This mandatory use of Standard Concepts is what allows all OHDSI collaborators to speak the same language. For example, let's look at the condition 'Pulmonary Tuberculosis' (TB). Figure 3.1 shows that the ICD9CM code for TB is 011.

Without the use of a standard way to represent TB the code 011 could be interpreted as 'Hospital Inpatient (Including Medicare Part A)' in the UB04 vocabulary, or as 'Nervous System Neoplasms without Complications, Comorbidities' in the DRG vocabulary. This is where Concept IDs, both Source and Standard, are valuable. The Concept ID that represents the 011 ICD9CM code is 44828631 (http://athena.ohdsi.org/search-terms/terms/44828631). This differentiates the ICD9CM from the UBO4 and from the DRG. The Standard Concept that ICD9CM code maps to is 253954 (http://athena.ohdsi.org/search-terms/terms/253954) as shown in figure 3.2 by the relationship 'Non-standard to Standard map (OMOP)'. This same mapping relationship exists between Read, ICD10, CIEL, and MeSH codes, among others, so that any research that references the standard SNOMED concept is sure to include all supported source codes.



TERM CONNECTIONS (82)

RELATIONSHIP	RELATES TO	CONCEPT ID	VOCABULARY
ICD-9-CM to MedDRA (MSSO)	Pulmonary tuberculosis	36110777	MedDRA
Non-standard to Standard map (OMOP)	Pulmonary tuberculosis	253954	SNOMED
Subsumes	Other specified pulmonary tuberculosis	44830894	ICD9CM
	Other specified pulmonary tuberculosis, bacteriological or histological examination not done	44836741	ICD9CM
	Other specified pulmonary tuberculosis, bacteriological or histological examination unknown (at present)	44836742	ICD9CM
	Other specified pulmonary tuberculosis, tubercle bacilli found (in sputum) by microscopy	44821641	ICD9CM
	Other specified pulmonary tuberculosis, tubercle bacilli not found (in sputum) by microscopy, but found by bacterial culture	44833188	ICD9CM

Figure 3.2: SNOMED code for Pulmonary Tuberculosis

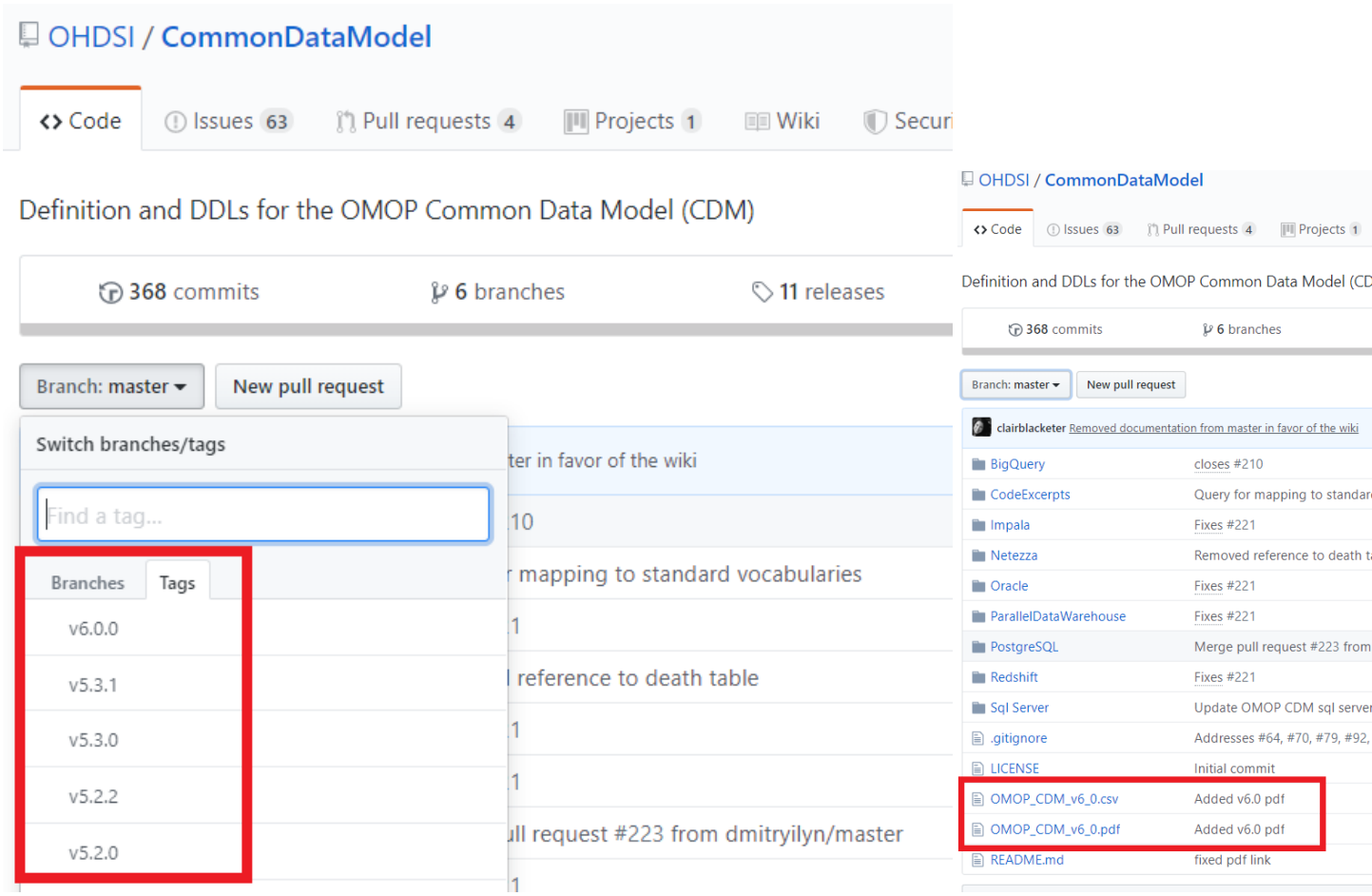
An example of how this relationship is depicted in the tables is shown in figure (**link to figure in CONDITION_OCCURRENCE**)

3.3 OMOP CDM Standardized Tables

The OMOP CDM contains 16 Clinical data tables, 10 Vocabulary tables, 2 Metadata tables, 4 Health System data tables, 2 Health Economics data tables, 3 standardized derived elements, and 2 results schema tables. To illustrate how these tables are utilized in practice the data of one person will be used as a common thread throughout the rest of the chapter. While part of the CDM the Vocabulary tables are not covered here, rather, they are detailed in depth in StandardizedVocabularies Chapter.

3.3.1 OMOP-CDM (table specification)

OMOP-CDM OHDSI Common Data Model . 2019 5 25 6.0 .
wiki page . OMOP-CDM specification Tags pdf .



3.3.2 OMOP-CDM

Open community OHDSI OMOP-CDM . , OMOP-CDM CDM . OMOP-CDM , .

OMOP-CDM , 2019 5 30 v6.0 .

3.3.3 Running Example: Endometriosis

Endometriosis is a painful condition whereby cells normally found in the lining of a woman's uterus occur elsewhere in the body. Severe cases can lead to infertility, bowel, and bladder problems. The following sections will detail one patient's experience with this disease and how her clinical experience might be represented in the Common Data Model.

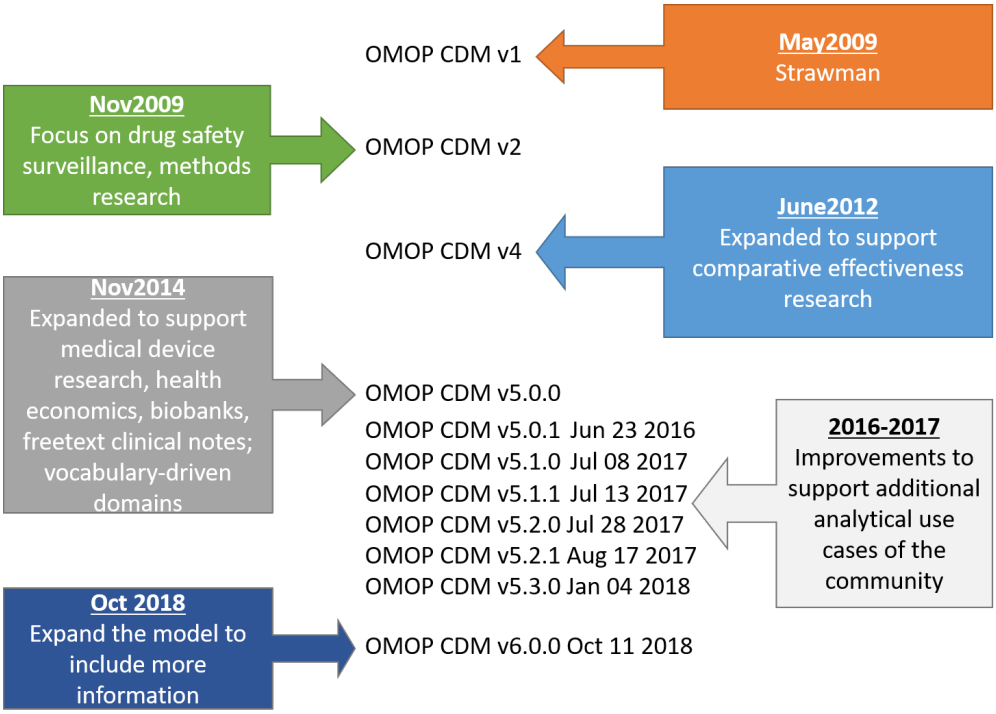


Figure 3.3: Evolution of OMOP-CDM

3.3.4 PERSON table

As the Common Data Model is a person-centric model (see section 3.2.1) let's start with how she would be represented in the PERSON table. For the full PERSON table specification please see the CDM wiki https://github.com/OHDSI/CommonDataModel/wiki/PERSON.

What do we know about Lauren?

- She is a 36-year-old woman
- Her birthday is 12-March-1982
- She is white
- She is english

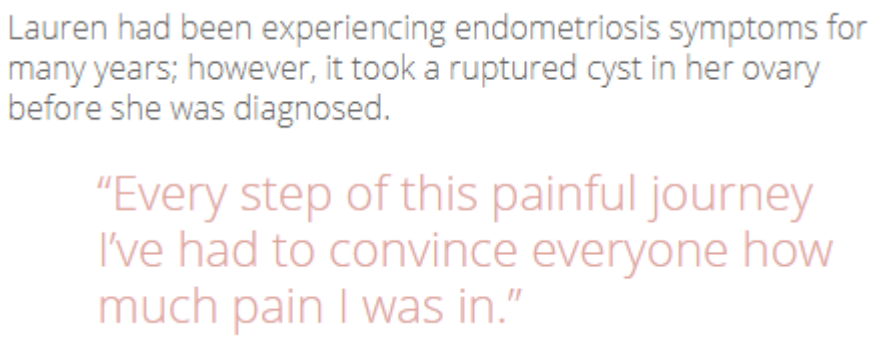


Figure 3.4: Read more about Lauren and endometriosis at https://www.endometriosis-uk.org/laurens-story

With that in mind, her PERSON table might look something like this:

Column Name	Value	Explanation
person_id	1	Person_id should be an integer, either directly from the source or generated as part of the build process.
gender_concept_id	8532	The concept_id referring to female gender is 8532 (http://athena.ohdsi.org/search-terms/terms/8532).
year_of_birth	1982	
month_of_birth	3	
day_of_birth	12	
birth_datetime	1982-03-12 00:00:00	When the time is not known midnight is used.
death_datetime		
race_concept_id	8527	The concept_id referring to white race is 8527 (http://athena.ohdsi.org/search-terms/terms/8527).
ethnicity_concept_id	38003564	Typically hispanic status is stored for ethnicity. The concept_id 38003564 (http://athena.ohdsi.org/search-terms/terms/38003564) refers to 'Not hispanic'.
location_id		Her address is not known.
provider_id		Her primary care provider is not known.
care_site_id		Her primary care site is not known.
person_source_value	1	Typically this would be her identifier in the source data, though often is it the same as the person_id.
gender_source_value	F	The gender value as it appears in the source is stored here.
gender_source_concept_id	0	If the gender value in the source was coded using a vocabulary recognized by OHDSI, that concept_id would go here. For example, if her gender was 'Sex-F' in the source and it was stated to be in the PCORNet vocabulary concept_id 44814665 (http://athena.ohdsi.org/search-terms/terms/44814665) would go in this field.
race_source_value	white	The race value as it appears in the source is stored here.
race_source_concept_id	0	Same principle as gender_source_concept_id.
ethnicity_source_value	english	The ethnicity value as it appears in the source is stored here.
ethnicity_source_concept_id	0	Same principle as gender_source_concept_id.

3.3.5 OBSERVATION_PERIOD table

The OBSERVATION_PERIOD table is designed to define the amount of time for which a patient's clinical events are recorded in the source system. For US healthcare insurance claims this is typically the enrollment period of the patient. When working with data from electronic health records (EHR) often the first record

in the system is considered the observation_period_start_date and the latest record is considered the observation_period_end_date with the understanding that only the clinical events that happened within that particular system were recorded. For the full OBSERVATION_PERIOD table specification please see the CDM wiki (https://github.com/OHDSI/CommonDataModel/wiki/OBSERVATION_PERIOD).

How can we determine Lauren's observation period?

Lauren's information is most similar to EHR data in that we only have records of her encounters from which to determine her observation period.

Encounter_ID	Start_Date	Stop_Date	EncounterClass
70	2010-01-06	2010-01-06	outpatient
80	2011-01-06	2011-01-06	outpatient
90	2012-01-06	2012-01-06	outpatient
100	2013-01-07	2013-01-07	outpatient
101	2013-01-14	2013-01-14	ambulatory
102	2013-01-17	2013-01-24	inpatient

Based on the encounter records her OBSERVATION_PERIOD table might look something like this:

Column Name	Value	Explanation
observation_period_id	1	This is typically an autogenerated field that creates a unique id number for each record in the table.
person_id	1	This comes from the PERSON table and links PERSON and OBSERVATION_PERIOD.
observation_period_start_date	2010-01-06	This is the start date of her earliest encounter on record.
observation_period_end_date	2013-01-24	This is the end date of her latest encounter on record.
period_type_concept_id	44814725	The best option in the Vocabulary with the concept class 'Obs Period Type' is 44814724 (http://athena.ohdsi.org/search-terms/terms/44814724), which stands for 'Period covering healthcare encounters'.

3.3.6 VISIT_OCCURRENCE

The VISIT_OCCURRENCE table houses information about a patient's encounters with the health care system. Within the OHDSI vernacular these are referred to as visits and are considered to be discreet events. There are 12 categories of visits though the most common are inpatient, outpatient, emergency and long term care. For the full VISIT_OCCURRENCE table specification please see the CDM wiki (https://github.com/OHDSI/CommonDataModel/wiki/VISIT_OCCURRENCE).

How do we represent Lauren's encounters as visits?

Revisting the encounters we used to determine her observation period:

Encounter_ID	Start_Date	Stop_Date	EncounterClass
70	2010-01-06	2010-01-06	outpatient
80	2011-01-06	2011-01-06	outpatient
90	2012-01-06	2012-01-06	outpatient

Encounter_ID	Start_Date	Stop_Date	EncounterClass
100	2013-01-07	2013-01-07	outpatient
101	2013-01-14	2013-01-14	ambulatory
102	**2013-01-17**	**2013-01-24**	**inpatient**

As an example let's represent the inpatient encounter as a record in the VISIT_OCCURRENCE table.

Column Name	Value	Explanation
visit_occurrence_id	514	This is typically an autogenerated field that creates a unique id number for each visit on the person's record in the converted CDM database.
person_id	1	This comes from the PERSON table and links PERSON and VISIT_OCCURRENCE.
visit_concept_id	9201	The concept_id referring to an inpatient visit is 9201 (http://athena.ohdsi.org/search-terms/terms/9201).
visit_start_date	2013-01-17	The start date of the visit.
visit_start_datetime	2013-01-17 00:00:00	The date and time of the visit started. When time is unknown midnight is used.
visit_end_date	2013-01-24	The end date of the visit. If this is a one-day visit the end date should match the start date.
visit_end_datetime	2013-01-24 00:00:00	The date and time of the visit end. If time is unknown midnight is used.
visit_type_concept_id	32034	This column is intended to provide information about the provenance of the visit record, i.e. does it come from an insurance claim, hospital billing record, EHR record, etc. For this example the concept_id 32035 (http://athena.ohdsi.org/search-terms/terms/32035) is used as the encounters are similar to electronic health records
provider_id*	NULL	If the encounter record has a provider associated, the id for that provider goes in this field. This should be the provider_id from the PROVIDER table that represents the provider on the encounter.
care_site_id	NULL	If the encounter record has a care site associated, the id for that care site goes in this field. This should be the care_site_id from the CARE_SITE table that codes for the care site on the encounter.
visit_source_value	inpatient	The visit value as it appears in the source goes here. In this context 'visit' means outpatient, inpatient, emergency, etc.
visit_source_concept_id	0	If the visit value from the source is coded using a vocabulary that is recognized by OHDSI, the concept_id that represents the visit source value would go here.
admitted_from_concept_id	0	If known, this is the concept_id that represents where the patient was admitted from. This concept should have the concept class 'Place of Service' and the domain 'Visit'. For example, if a patient was admitted to the hospital from home, the concept_id would be 8536 (http://athena.ohdsi.org/search-terms/terms/8536).

Column Name	Value	Explanation
admitted_from_source_value	NULL	This is the value from the source that represents where the patient was admitted from. Using the above example, this would be 'home'.
discharge_to_concept_id	0	If known, this is the concept_id that represents where the patient was discharged to. This concept should have the concept class 'Place of Service' and the domain 'Visit'. For example, if a patient was released to an assisted living facility, the concept_id would be 8615 (http://athena.ohdsi.org/search-terms/terms/8615).
discharge_to_source_value	0	This is the value from the source that represents where the patient was discharged to. Using the above example, this would be 'assisted living facility'.
preceding_visit_occurrence_id	NULL	The visit_occurrence_id for the visit immediately preceding the current one in time for the patient.

*A patient may interact with multiple health care providers during one visit, as is often the case with inpatient stays. These interactions can be recorded in the VISIT_DETAIL table. While not covered in depth in this chapter, you can read more about the VISIT_DETAIL table on the CDM wiki (https://github.com/OHDSI/CommonDataModel/wiki/VISIT_DETAIL)

3.3.7 CONDITION_OCCURRENCE

Records in the CONDITION_OCCURRENCE table are diagnoses, signs, or symptoms of a condition either observed by a Provider or reported by the patient.

What are Lauren's conditions?

Revisiting her account she says "About 3 years ago I noticed my periods, which had also been painful, were getting increasingly more painful. I started becoming aware of a sharp jabbing pain right by my colon and feeling tender and bloated around my tailbone and lower pelvis area. My periods had become so painful that I was missing 1-2 days of work a month. Painkillers sometimes dulled the pain, but usually they didn't do much."

The SNOMED code for painful menstruation cramps, otherwise known as dysmenorrhea, is 266599000. Let's see how that would be represented in the CONDITION_OCCURRENCE table:

Column	Value	Explanation
condition_occurrence_id	964	This is typically an autogenerated field that creates a unique id number for each condition on the person's record in the converted CDM database.
person_id	1	This comes from the PERSON table and links PERSON and CONDITION_OCCURRENCE.
condition_concept_id	194696	The concept_id that represents the SNOMED code 266599000 is 194696 (http://athena.ohdsi.org/search-terms/terms/194696)
condition_start_date	2010-01-06	The date when the instance of the Condition is recorded.
condition_start_datetime	2010-01-06 00:00:00	The date and time when the instance of the Condition is recorded. Midnight is used when the time is unknown

Column	Value	Explanation
condition_end_date	NULL	If known, this is the date when the instance of the Condition is considered to have ended.
condition_end_datetime	NULL	If known, this is the date and time when the instance of the Condition is considered to have ended.
condition_type_concept_id	32020	This column is intended to provide information about the provenance of the condition, i.e. does it come from an insurance claim, hospital billing record, EHR record, etc. For this example the concept_id 32020 (http://athena.ohdsi.org/search-terms/terms/32020) is used as the encounters are similar to electronic health records. Concept_ids in this field should be in the 'Condition Type' vocabulary.
condition_status_concept_id	0	If known, the condition_status_concept_id represents when and/or how the condition was diagnosed. For example, a condition could be an admitting diagnosis, in which case the concept_id 4203942 (http://athena.ohdsi.org/search-terms/terms/4203942) would be used.
stop_reason	NULL	If known, the reason that the Condition was no longer present, as indicated in the source data.
provider_id	NULL	If the condition record has a diagnosing provider listed, the id for that provider goes in this field. This should be the provider_id from the PROVIDER table that represents the provider on the encounter.
visit_occurrence_id	509	If known, this is the visit (represented as visit_occurrence_id taken from the VISIT_OCCURRENCE table) during which the condition was diagnosed.
visit_detail_id	NULL	If known, this is the visit detail encounter (represented as visit_detail_id from the VISIT_DETAIL table) during which the condition was diagnosed.
condition_source_value	266599000	This is the value from the source that represents the condition. In Lauren's case of dysmenorrhea the SNOMED code for that condition is stored here and the standard concept_id mapped from that code is stored in CONDITION_CONCEPT_ID.
condition_source_concept_id	194696	If the condition value from the source is coded using a vocabulary that is recognized by OHDSI, the concept_id that represents that value would go here. In the example of dysmennorhea the source value is a SNOMED code so the concept_id that represents that code is 194696. In this case it is the same as the condition_concept_id since the SNOMED vocabulary is the standard condition vocabulary
condition_status_source_value	0	If the condition status value from the source is coded using a vocabulary that is recognized by OHDSI, the concept_id that represents that source value would go here.

3.3.8 DRUG_EXPOSURE

The DRUG_EXPOSURE captures records about the utilization of a Drug when ingested or otherwise introduced into the body. Drugs include prescription and over-the-counter medicines, vaccines, and large-molecule biologic therapies. Radiological devices ingested or applied locally do not count as Drugs.

Drug Exposure is inferred from clinical events associated with orders, prescriptions written, pharmacy dispensings, procedural administrations, and other patient-reported information.

What are Lauren's drug exposures?

We know that Lauren was given 60 acetaminophen 325mg oral tablets for 30 days (NDC code 69842087651) at her visit on 2010-01-06 to help with her dysmenorrhea pain. Here's how that might look in the DRUG_EXPOSURE table:

Column	Value	Explanation
drug_exposure_id	1001	This is typically an autogenerated field that creates a unique id number for each drug_exposure on the person's record in the converted CDM database.
person_id	1	This comes from the PERSON table and links PERSON and DRUG_EXPOSURE.
drug_concept_id	1127433	The NDC code for acetaminophen maps to the RxNorm code 313782 which is represented by the concept_id 1127433 (http://athena.ohdsi.org/search-terms/terms/1127433).
drug_exposure_start_date	2010-01-06	The start date of the drug exposure
drug_exposure_start_datetime	2010-01-06 00:00:00	The start date and time of the drug exposure. Midnight is used when the time is not known.
drug_exposure_end_date	2010-02-05	The end date of the drug exposure. Depending on different sources, it could be a known or an inferred date and denotes the last day at which the patient was still exposed to the drug. In this case the end is inferred since we know Lauren had a 30 days supply.
drug_exposure_end_datetime	2010-02-05 00:00:00	The end date and time of the drug exposure. Similar rules apply as to drug_exposure_end_date. Midnight is used when time is unknown
verbatim_end_date	NULL	If the source provides an end date rather than just days supply that date goes here.
drug_type_concept_id	38000177	This column is intended to provide information about the provenance of the drug, i.e. does it come from an insurance claim, prescription record, etc. For this example the concept_id 38000177 (http://athena.ohdsi.org/search-terms/terms/38000177) is used as the drug record is from a written prescription. Concept_ids in this field should be in the 'Drug Type' vocabulary.
stop_reason	NULL	The reason the Drug was stopped. Reasons include regimen completed, changed, removed, etc.
refills	NULL	The number of refills after the initial prescription. The initial prescription is not counted, values start with null. In the case of Lauren's acetaminophen she did not have any refills so the value is NULL.

Column	Value	Explanation
quantity	60	The quantity of drug as recorded in the original prescription or dispensing record.
days_supply	30	The number of days of supply of the medication as prescribed.
sig	NULL	The directions ('signetur') on the Drug prescription as recorded in the original prescription (and printed on the container) or dispensing record.
route_concept_id	4132161	This concept is meant to represent the route of the drug the patient was was exposed to. Lauren took her acetaminophen orally so the concept_id 4132161 (http://athena.ohdsi.org/search-terms/terms/4132161) is used.
lot_number	NULL	An identifier assigned to a particular quantity or lot of Drug product from the manufacturer.
provider_id	NULL	If the drug record has a prescribing provider listed, the id for that provider goes in this field. This should be the provider_id from the PROVIDER table that represents the provider on the encounter.
visit_occurrence_id	509	If known, this is the visit (represented as visit_occurrence_id taken from the VISIT_OCCURRENCE table) during which the drug was prescribed.
visit_detail_id	NULL	If known, this is the visit detail (represented as visit_detail_id taken from the VISIT_DETAIL table) during which the drug was prescribed.
drug_source_value	69842087651	This is the source code for the Drug as it appears in the source data. In Lauren's case she was prescribed acetaminophen and the NDC code is stored here.
drug_source_concept_id	750264	This is the concept_id that represents the drug source value. In this example the concept_id is 750264 (http://athena.ohdsi.org/search-terms/terms/750264).
route_source_value	NULL	The information about the route of administration as detailed in the source.
dose_unit_source_value	NULL	The information about the dose unit as detailed in the source.

3.3.9 PROCEDURE_OCCURRENCE

The PROCEDURE_OCCURRENCE table contains records of activities or processes ordered by, or carried out by, a healthcare provider on the patient to have a diagnostic or therapeutic purpose. Procedures are present in various data sources in different forms with varying levels of standardization. For example:

- Medical Claims include procedure codes that are submitted as part of a claim for health services rendered, including procedures performed.
- Electronic Health Records that capture procedures as orders.

What procedures did Lauren have? From her description we know she had a ultrasound of her left ovary on 2013-01-14 that showed a 4x5cm cyst. Here's how that would look in the PROCEDURE_OCCURRENCE

table:

Column	Value	Explanation
procedure_occurrence_id	1277	This is typically an autogenerated field that creates a unique id number for each procedure_occurrence on the person's record in the converted CDM database.
person_id	1	This comes from the PERSON table and links PERSON and PROCEDURE_OCCURRENCE
procedure_concept_id	4127451	The SNOMED procedure code for a pelvic ultrasound is 304435002 which is represented by the concept_id 4127451 (http://athena.ohdsi.org/search-terms/terms/4127451).
procedure_date	2013-01-14	The date on which the procedure was performed.
procedure_datetime	2013-01-14 00:00:00	The date and time on which the procedure was performed. Midnight is used when time is unknown.
procedure_type_concept_id	38000275	This column is intended to provide information about the provenance of the procedure, i.e. does it come from an insurance claim, EHR order, etc. For this example the concept_id 38000275 (http://athena.ohdsi.org/search-terms/terms/38000275) is used as the procedure record is from an EHR record. Concept_ids in this field should be in the 'Procedure Type' vocabulary.
modifier_concept_id	0	This is meant for a concept_id representing the modifier on the procedure. For example, if the record indicated that a CPT4 procedure was performed bilaterally then the concept_id 42739579 (http://athena.ohdsi.org/search-terms/terms/42739579) would be used.
quantity	0	The quantity of procedures ordered or administered.
provider_id	NULL	If the procedure record has a provider listed, the id for that provider goes in this field. This should be the provider_id from the PROVIDER table that represents the provider on the encounter.
visit_occurrence_id	740	If known, this is the visit (represented as visit_occurrence_id taken from the VISIT_OCCURRENCE table) during which the procedure was performed.
visit_detail_id	NULL	If known, this is the visit detail (represented as visit_detail_id taken from the VISIT_DETAIL table) during which the procedure was performed.
procedure_source_value	304435002	The source code for the Procedure as it appears in the source data. This code is mapped to a standard procedure Concept in the Standardized Vocabularies and the original code is, stored here for reference.
procedure_source_concept_id	4127451	This is the concept_id that represents the procedure source value.

Column	Value	Explanation
modifier_source_value	NULL	The source code for the modifier as it appears in the source data.

Chapter 4

The OMOP Vocabulary

- ATHENA: ATHENA OMOP vocabulary .
- : OHDSI wiki page .
- OMOP vocabulary github : OMOP vocabulary github
- OHDSI Github : OHDSI vocabulary github
- OHDSI tutorial : OHDSI past event OMOP vocabulary tutorial .

4.1 Design Principles

Chapter 5

Extract Transform Load

ETL .

- OHDIS ETL best practice: OHDIS ETL best practice
- ETL : ETL github
- ETL sample : ETL sample page
- ETL tools : ETL tool page
- THEMIS WG: Korean THEMIS WG .

5.1 Pre-processing

5.1.1 WhiteRabbit and Rabbit-in-a-Hat

Chapter 6

SQL and R

6.1 Database Connector

Database Connnector github

- Database Connnector 2.4 version DB
  - MicrosoftSQL Server
  - Oracle
  - PostgresSql
  - Microsoft Parallel Data Warehouse (a.k.a. Analytics Platform System)
  - Amazon Redshift
  - Apache Impala
  - Google BigQuery
  - IBM Netezza
  - SQLite

6.2 SQL Render

- SQL Render github
- SQL Render
- OHDSI SQL Devloper

Chapter 7

Cohort

7.1 Using SQL

7.2 ATLAS

7.3 Phenotype Library

Chapter 8

Characterization

- Cohort definition tutorial: OHDSI past event Cohort definition .

8.1 FeatureExtraction

8.2 ATLAS

8.2.1 Baseline characteristics

8.2.2 Incidence rate calculation

Chapter 9

Population-Level Estimation

Chapter leads: Martijn Schuemie, David Madigan, Marc Suchard & Patrick Ryan : ,

Observational healthcare data(: administrative claims, electronic health records)
. (population-level effect estimation), (:)
. .

- **(direct effect estimation)**: .
- **(comparative effect estimation)**: (comparator exposure) (target exposure)
.

In both cases, the patient-level causal effect contrasts a factual outcome, i.e., what happened to the exposed patient, with a counterfactual outcome, i.e., what would have happened had the exposure not occurred (direct) or had a different exposure occurred (comparative). Since any one patient reveals only the factual outcome (the fundamental problem of causal inference), the various effect estimation designs employ different analytic devices to shed light on the counterfactual outcomes.

(use-cases) , (safety surveillance), (comparative effectiveness) . (: signal evaluation) (: signal detection) . , .

OHDSI Methods Library R **Population-Level Estimation** study design .
, ATLAS R . .

9.1 The cohort method design

(Hernan and Robins, 2016). (target) (comparator) , (:) . 13.1 5 . 9.1.

Table 9.1: Main design choices in a comparative cohort design.

Choice	Description
Target cohort	A cohort representing the target treatment
Comparator cohort	A cohort representing the comparator treatment
Outcome cohort	A cohort representing the outcome of interest
Time-at-risk	At what time (often relative to the target and comparator cohort start and end dates) do we consider the risk of the outcome?
Model	The model used to estimate the effect while adjusting for differences between the target and comparator

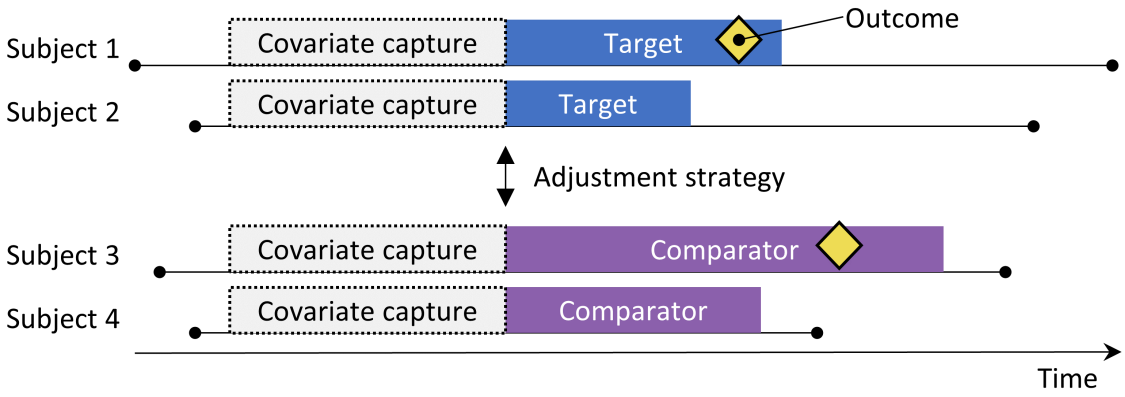


Figure 9.1: (target treatment) (comparator treatment) . (stratification), (matching), (weighting), (baseline characteristics) (adjustment strategy) . (Propensity model) (Outcome model) .

. , (odds ratio) (logistic regression) . time-at-risk target comparator . , (poisson regression) , (incidence rate) , (incidence rate ratio) . (cox regression) , target comparator (proportional hazard) , (hazard ratio) time-to-first-outcome .

New-user cohort method . . , , (exposure of interest) . , .

(systemically) . , (target cohort) 60 , (comparator cohort) 40 . (:) . . , . . . , . “confounder” .

9.1.1 Propensity scores

(randomized trial) ( ) . , . , . . trial (:) trial .

(Propensity score, PS) (Rosenbaum and Rubin, 1983). two-arm , 0.5 . , ( ) . . (:) , . , . PS . , target PS comparator (PS matching) , PS (PS stratification) , PS IPTW(Inverse Probability of Treatment Weighting) . , , variable-ratio matching . (Rassen et al., 2012)

, one-on-one PS , Jan 0.4 , . Jun (priori) 0.4 , , measured confounders Jan Jun outcome mini-randomized trial . Jan Jun . : , . .

Propensity scoring (measured confounder) . , (treatment assignment) “ ” , . “ ” , . , .

> Propensity scoring controls for measured confounders. In fact, if treatment assignment is “strongly ignorable” given measured characteristics, propensity scoring will yield an unbiased estimate of the causal effect. “Strongly ignorable” essentially means that there are no unmeasured confounders, and that the measured confounders are adjusted for appropriately. Unfortunately this is not a testable assumption.

9.1.2 Variable selection

, (manually selected characteristics) . OHDSI , (,) . (Tian et al., 2018) , , . 10,000 – 100,000 , Cyclops large-scale regularized regression(Suchard et al., 2013) .

(covariate capture window) .

. , .

We typically include the day of treatment initiation in the covariate capture window because many relevant data points such as the diagnosis leading to the treatment are recorded on that date. This does require us to explicitly exclude the target and comparator treatment from the set of covariates, because these are the things we are trying to predict.

“ ” (collider) . (Hernan et al., 2002) .(Schneeweiss, 2018) , “ ” . , , , .

Some have argued that a data-driven approach to covariate selection that does not depend on clinical expertise to specify the “right” causal structure runs the risk of erroneously including so-called instrumental variables and colliders, thus increasing variance and potentially introducing bias. (Hernan et al., 2002) However, these concerns are unlikely to have a large impact in real-world scenarios. (Schneeweiss, 2018) Furthermore, in medicine the true causal structure is rarely known, and when different researchers are asked to identify the ‘right’ covariates to include for a specific research question, each researcher invariably comes up with a different list, thus making the process irreproducible. Most importantly, our diagnostics such as inspection of the propensity model, evaluating balance on all covariates, and including negative controls would identify most problems related to colliders and instrumental variables.

9.1.3 Caliper

0 1 . , “ (caliper)” . (Austin, 2011) , 0.2 (default) .

9.1.4 Overlap: preference scores

! . OHDSI “ (preference score)” .(Walker et al., 2013) “market share” . , 10% (90%), 0.5 10% .

$$\ln\left(\frac{F}{1-F}\right) = \ln\left(\frac{S}{1-S}\right) - \ln\left(\frac{P}{1-P}\right)$$

F (preference score), S (propensity score), P (proportion of patients receiving the target treatment) .

Walker et al. (2013) “ (empirical equipoise)” . 0.3 0.7 (exposure pair) .

Walker et al. (2013) discuss the concept of “empirical equipoise.” They accept exposure pairs as emerging from empirical equipoise if at least half of the exposures are to patients with a preference score of between 0.3 and 0.7.

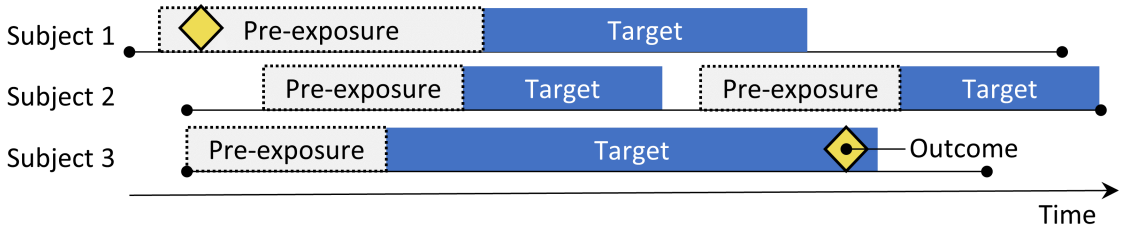


Figure 9.2: The self-controlled cohort design. The rate of outcomes during exposure to the target is compared to the rate of outcomes in the time pre-exposure.

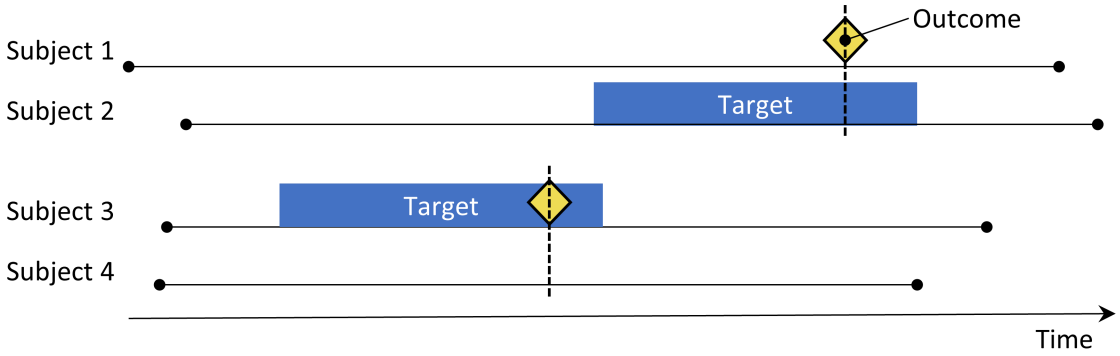


Figure 9.3: The case-control design. Subjects with the outcome ('cases') are compared to subjects without the outcome ('controls') in terms of their exposure status. Often, cases and controls are matched on various characteristics such as age and sex.

9.1.5 Balance

(good practice) PS . 9.18 OHDSI . PS . 0.1 . (Rubin, 2001)

9.2 The self-controlled cohort design

(self-controlled cohort, SCC) (Ryan et al., 2013) . 9.2 4 SCC .

Table 9.2: Main design choices in a self-controlled cohort design.

Choice	Description
Target cohort	A cohort representing the treatment
Outcome cohort	A cohort representing the outcome of interest
Time-at-risk	At what time (often relative to the target cohort start and end dates) do we consider the risk of the outcome?
Control time	The time period used as the control time

(control group) (between-person) . .

Because the same subject that make up the exposed group are also used as the control group, no adjustment for between-person differences need to be made. However, the method is vulnerable to other differences, such as differences in the baseline risk of the outcome between different time periods.

9.3 The case-control design

- (Vandenbroucke and Pearce, 2012) " (agent) ?" . ,

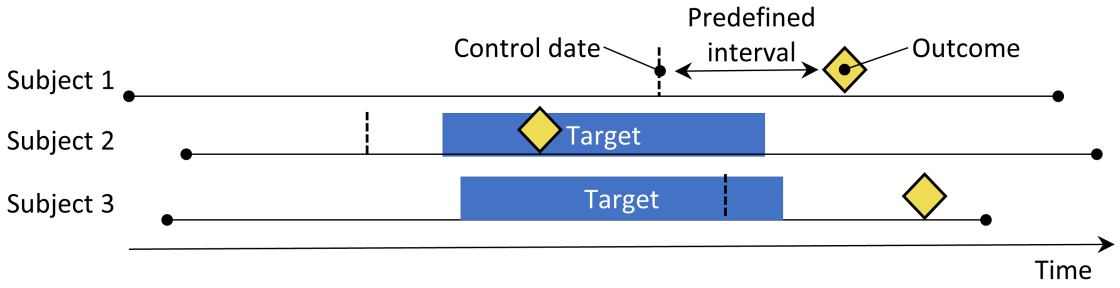


Figure 9.4: The case-crossover design. The time around the outcome is compared to a control date set at a predefined interval prior to the outcome date.

Table 9.3: Main design choices in a case-control design.

Choice	Description
Outcome cohort	A cohort representing the cases (the outcome of interest)
Control cohort	A cohort representing the controls. Typically the control cohort is automatically derived from the outcome cohort using some selection logic
Target cohort	A cohort representing the treatment
Nesting cohort	Optionally, a cohort defining the subpopulation from which cases and controls are drawn
Time-at-risk	At what time (often relative to the index date) do we consider exposure status?

. , , subgroup nested analysis .

Often, one selects controls to match cases based on characteristics such as age and sex to make them more comparable. Another widespread practice is to nest the analysis within a specific subgroup of people, for example people that have all been diagnosed with one of the indications of the exposure of interest.

9.4 The case-crossover design

The case-crossover (Maclure, 1991) design evaluates whether the rate of exposure is different at the time of the outcome than at some predefined number of days prior to the outcome. It is trying to determine whether there is something special about the day the outcome occurred. Table 9.4 shows the choices that define a case-crossover question.

Table 9.4: Main design choices in a case-crossover design.

Choice	Description
Outcome cohort	A cohort representing the cases (the outcome of interest)
Target cohort	A cohort representing the treatment
Time-at-risk	At what time (often relative to the index date) do we consider exposure status?
Control time	The time period used as the control time

Cases serve as their own controls. As self-controlled designs, they should be robust to confounding due to between-person differences. One concern is that, because the outcome date is always later than the control date, the method will be positively biased if the overall frequency of exposure increases over time (or negatively biased if there is a decrease). To address this, the case-time-control design (Suissa, 1995) was developed, which adds controls, matched for example on age and sex, to the case-crossover design to adjust for exposure trends.

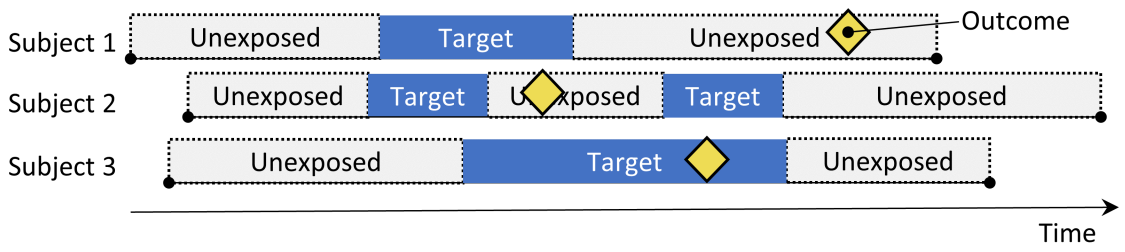


Figure 9.5: The Self-Controlled Case Series design. The rate of outcomes during exposure is compared to the rate of outcomes when not exposed.

Table 9.5: Main design choices in a self-controlled case series design.

Choice	Description
Target cohort	A cohort representing the treatment
Outcome cohort	A cohort representing the outcome of interest
Time-at-risk	At what time (often relative to the target cohort start and end dates) do we consider the risk of the outcome?
Model	The model to estimate the effect, including any adjustments for time-varying confounders

(self-controlled design) , SCCS (counfounding due to betwee-person difference) , (confounding due to time-varying effect) . , . SCCS (Simpson et al., 2013) . (cross-validation) L1-regularization .

SCCS (outcome date) . outcome , (fatal) , . SCCS .(Farrington et al., 2011)

9.6 Designing a hypertension study

9.6.1 Problem definition

ACE (ACEi) , , . (Zaman et al., 2002) , , , , . (Sabroe and Black, 1997) . (:) . (Powers et al., 2012) ACEi (Magid et al., 2010; Toh et al., 2012), 1 .(Whelton et al., 2018) thiazide thiazide-like , ACEi .

(population-level estimation) (observational healthcare data) :

Thiazide thiazide-like ACEi ?

Thiazide thiazide-like ACEi ?

(comparative effect estimation) Cohort Method Cohort Method

9.6.2 Target and comparator

ACEi THZ new-user . 7 . 1 , .

9.6.3 Outcome

, . , 180 .

9.6.5 Model

We fit a PS model using the default set of covariates, including demographics, conditions, drugs, procedures, measurements, observations, and several co-morbidity scores. We exclude ACEi and THZ from the covariates. We perform variable-ratio matching and condition the Cox regression on the matched sets.

, , , , , , PS , ACEi THZ . - (variable-ratio matching) , PS .

9.6.6 Study summary

Table 9.6: Main design choices for our comparative cohort study.

Choice	Value
Target cohort	New users of ACE inhibitors as first-line monotherapy for hypertension.
Comparator cohort	New users of thiazides or thiazide-like diuretics as first-line monotherapy for hypertension.
Outcome cohort	Angioedema or acute myocardial infarction.
Time-at-risk	Starting the day after treatment initiation, stopping when exposure stops.
Model	Cox proportional hazards model using variable-ratio matching.

9.6.7 Control questions

. (hazard ratio) 1 (negative control) 1 (positive control) . , . Method Validity .

9.7 Implementing the study using ATLAS

ATLAS (Estimation function) . ATLAS . . .

ATLAS (Estimation function) . ATLAS Estimation . . .

(estimation design function) : (comparisons), (analysis settings), (evaluation settings). , ATLAS . :

9.7.1 Comparative cohort settings

. “Add Comparison” . (target) (comparator) . “Add Outcome” , (outcome) . Cohorts .

- (target-comparator pair) . .

Negative control outcomes

(Negative Control Outcome) , 1 . . , (concept set) (However, typically, we only have a concept set, with one concept per negative control outcome, and some standard logic to turn these into outcome cohorts). Method Validity . , descendant . 9.7 .

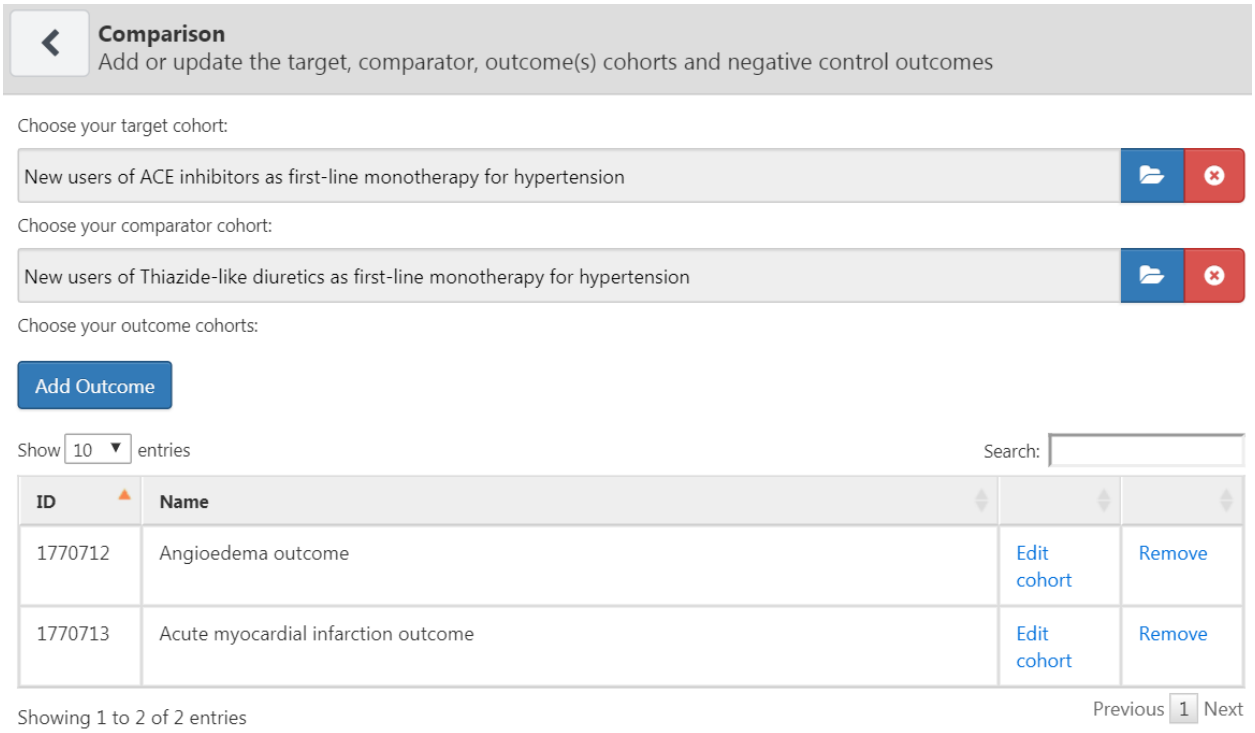


Figure 9.6: The comparison dialog

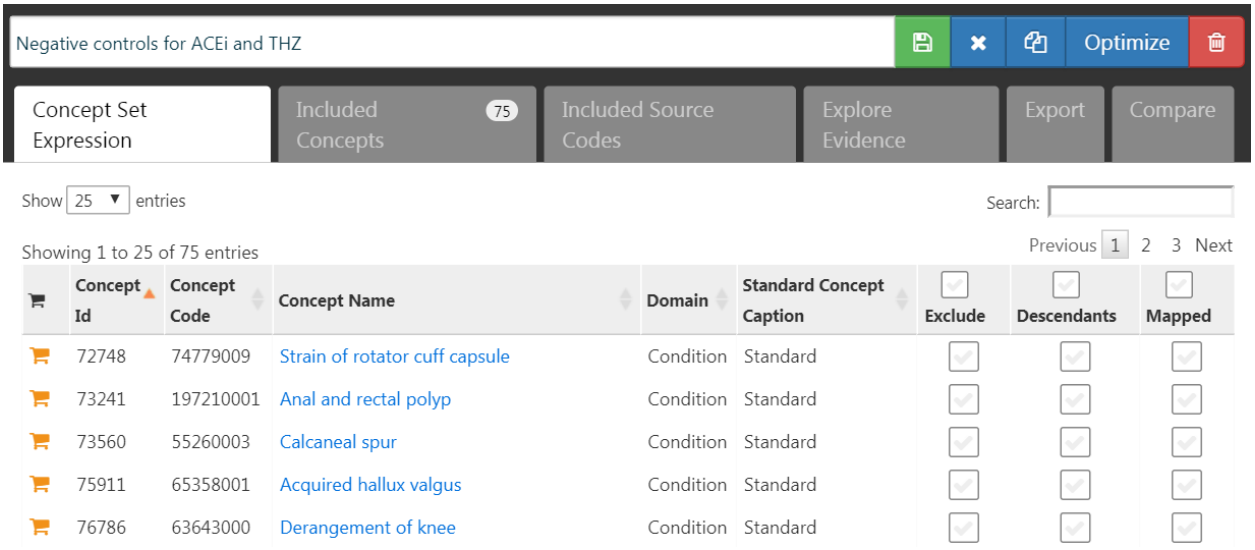


Figure 9.7: Negative Control concept set.

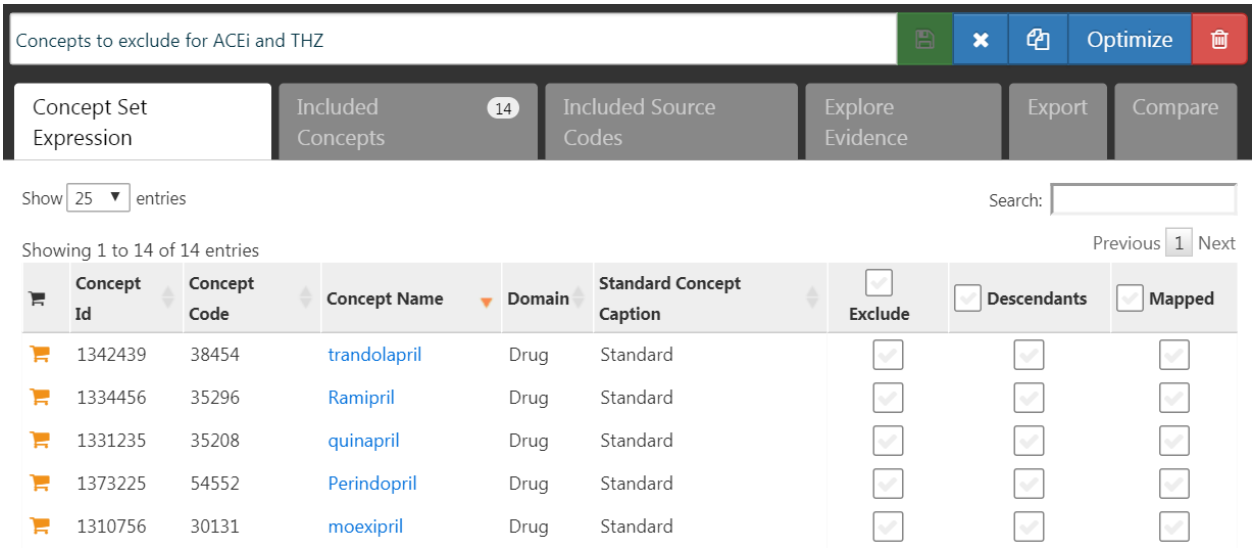


Figure 9.8: The concept set defining the concepts to exclude.

Concepts to include

When selecting concept to include, we can specify which covariates we would like to generate, for example to use in a propensity model. When specifying covariates here, all other covariates (aside from those you specified) are left out. We usually want to include all baseline covariates, letting the regularized regression build a model that balances all covariates. The only reason we might want to specify particular covariates is to replicate an existing study that manually picked covariates. These inclusions can be specified in this comparison section or in the analysis section, because sometimes they pertain to a specific comparison (e.g. know confounders in a comparison), or sometimes they pertain to an analysis (e.g. when evaluating a particular covariate selection strategy).

Concepts to exclude

Rather than specifying which concepts to include, we can instead specify concepts to *exclude*. When we submit a concept set in this field, we use every covariate except for those that we submitted. When using the default set of covariates, which includes all drugs and procedures occurring on the day of treatment initiation, we must exclude the target and comparator treatment, as well as any concepts that are directly related to these. For example, if the target exposure is an injectable, we should not only exclude the drug, but also the injection procedure from the propensity model. In this example, the covariates we want to exclude are ACEi and THZ. Figure 9.8 shows we select a concept set that includes all these concepts.

After selecting the negative controls and covariates to exclude, the lower half of the comparisons dialog should look like Figure 9.9.

9.7.2 Effect estimation analysis settings

After closing the comparisons dialog we can click on "Add Analysis Settings." In the box labeled "Analysis Name," we can give the analysis a unique name that is easy to remember and locate in the future. For example, we could set the name to "Propensity score matching."

Study population

There are a wide range of options to specify the study population, which is the set of subjects that will enter the analysis. Many of these overlap with options available when designing the target and comparator cohorts in the cohort definition tool. One reason for using the options in Estimation instead of in the cohort definition is re-usability; we can define the target, comparator, and outcome cohorts completely independently, and

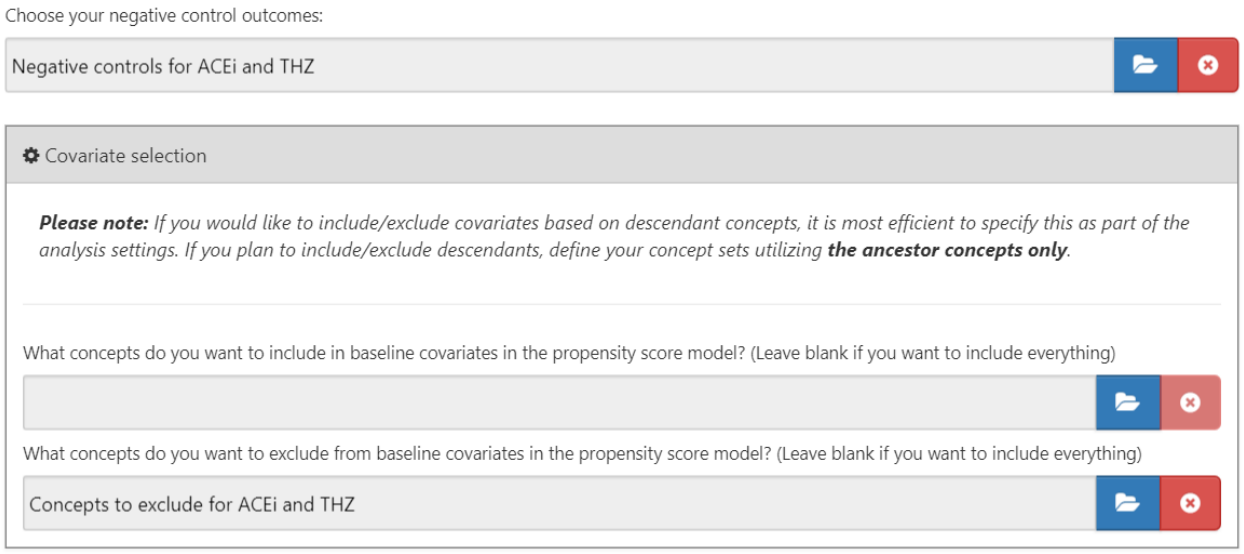


Figure 9.9: The comparison window showing concept sets for negative controls and concepts to exclude.

add dependencies between these at a later point in time. For example, if we wish to remove people who had the outcome before treatment initiation, we could do so in the definitions of the target and comparator cohort, but then we would need to create separate cohorts for every outcome! Instead, we can choose to have people with prior outcomes be removed in the analysis settings, and now we can reuse our target and comparator cohorts for our two outcomes of interest (as well as our negative control outcomes).

The **study start and end dates** can be used to limit the analyses to a specific period. The study end date also truncates risk windows, meaning no outcomes beyond the study end date will be considered. One reason for selecting a study start date might be that one of the drugs being studied is new and did not exist in an earlier time. Automatically adjusting for this can be done by answering "yes" to the question "**Restrict the analysis to the period when both exposures are observed?**". Another reason to adjust study start and end dates might be that medical practice changed over time (e.g., due to a drug warning) and we are only interested in the time where medicine was practiced a specific way.

The option "**Should only the first exposure per subject be included?**" can be used to restrict to the first exposure per patient. Often this is already done in the cohort definition, as is the case in this example. Similarly, the option "**The minimum required continuous observation time prior to index date for a person to be included in the cohort**" is often already set in the cohort definition, and can therefore be left at 0 here. Having observed time (as defined in the OBSERVATION_PERIOD table) before the index date ensures that there is sufficient information about the patient to calculate a propensity score, and is also often used to ensure the patient is truly a new user, and therefore was not exposed before.

"**Remove subjects that are in both the target and comparator cohort?**" defines, together with the option "**If a subject is in multiple cohorts, should time-at-risk be censored when the new time-at-risk starts to prevent overlap?**" what happens when a subject is in both target and comparator cohorts. The first setting has three choices:

- "**Keep All**" indicating to keep the subjects in both cohorts. With this option it might be possible to double-count subjects and outcomes.
- "**Keep First**" indicating to keep the subject in the first cohort that occurred.
- "**Remove All**" indicating to remove the subject from both cohorts.

If the options "keep all" or "keep first" are selected, we may wish to censor the time when a person is in both cohorts. This is illustrated in Figure 9.10. By default, the time-at-risk is defined relative to the cohort start

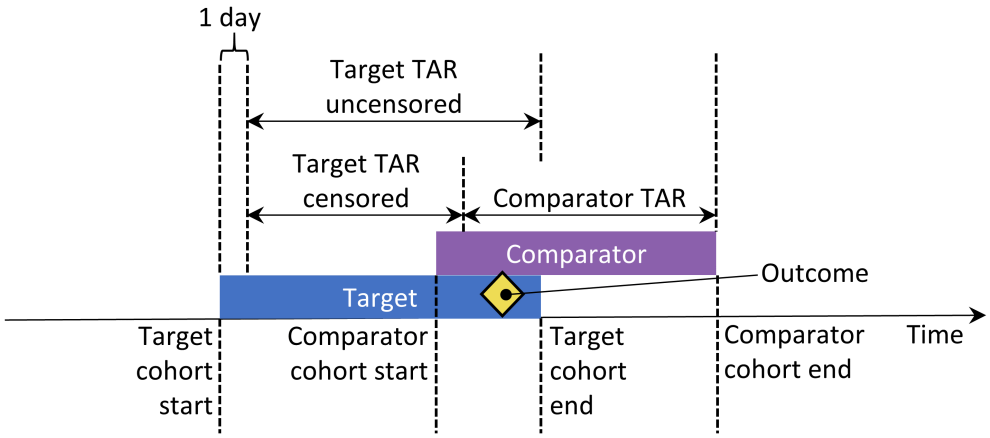


Figure 9.10: Time-at-risk (TAR) for subjects who are in both cohorts, assuming time-at-risk starts the day after treatment initiation, and stops at exposure end.

and end date. In this example, the time-at-risk starts one day after cohort entry, and stops at cohort end. Without censoring the time-at-risk for the two cohorts might overlap. This is especially problematic if we choose to keep all, because any outcome that occurs during this overlap (as shown) will be counted twice. If we choose to censor, the first cohort's time-at-risk ends when the second cohort's time-at-risk starts.

We can choose to **remove subjects that have the outcome prior to the risk window start**, because often a second outcome occurrence is the continuation of the first one. For instance, when someone develops heart failure, a second occurrence is likely, which means the heart failure probably never fully resolved in between. On the other hand, some outcomes are episodic, and it would be expected for patients to have more than one independent occurrence, like an upper respiratory infection. If we choose to remove people that had the outcome before, we can select **how many days we should look back when identifying prior outcomes**.

Our choices for our example study are shown in Figure 9.11. Because our target and comparator cohort definitions already restrict to the first exposure and require observation time prior to treatment initiation, we do not apply these criteria here.

Covariate settings

Here we specify the covariates to construct. These covariates are typically used in the propensity model, but can also be included in the outcome model (the Cox proporitional hazards model in this case). If we **click to view details** of our covariate settings, we can select which sets of covariates to construct. However, the recommendation is to use the default set, which constructs covariates for demographics, all conditions, drugs, procedures, measurements, etc.

We can modify the set of covariates by specifying concepts to **include** and/or **exclude**. These settings are the same as the ones found in Comparison Settings Section on comparison settings. The reason why they can be found in two places is because sometimes these settings are related to a specific comparison, as is the case here because we wish to exclude the drugs we are comparing, and sometimes the settings are related to a specific analysis. When executing an analysis for a specific comparison using specific analysis settings, the OHDSI tools will take the union of these sets.

The choice to **add descendants to include or exclude** affects this union of the two settings. So in this example we specified only the ingredients to exclude when defining the comparisons. Here we set "Should descendant concepts be added to the list of excluded concepts?" to "Yes" to also add all descendants.

Figure 9.12 shows our choices for this study. Note that we have selected to add descendants to the concept to exclude, which we defined in the comparison settings in Figure 9.9.

Time at risk

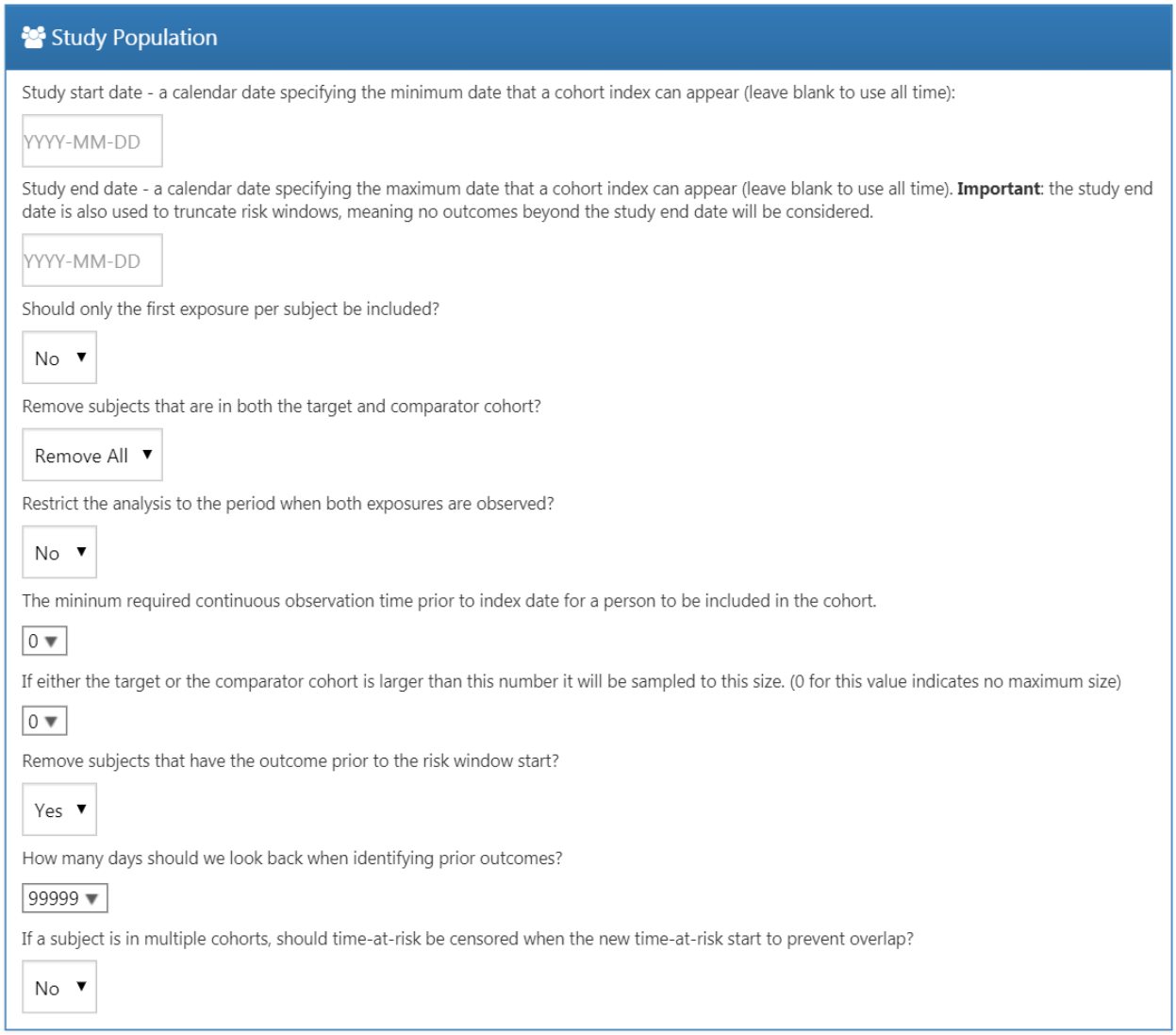


Figure 9.11: Study population settings.

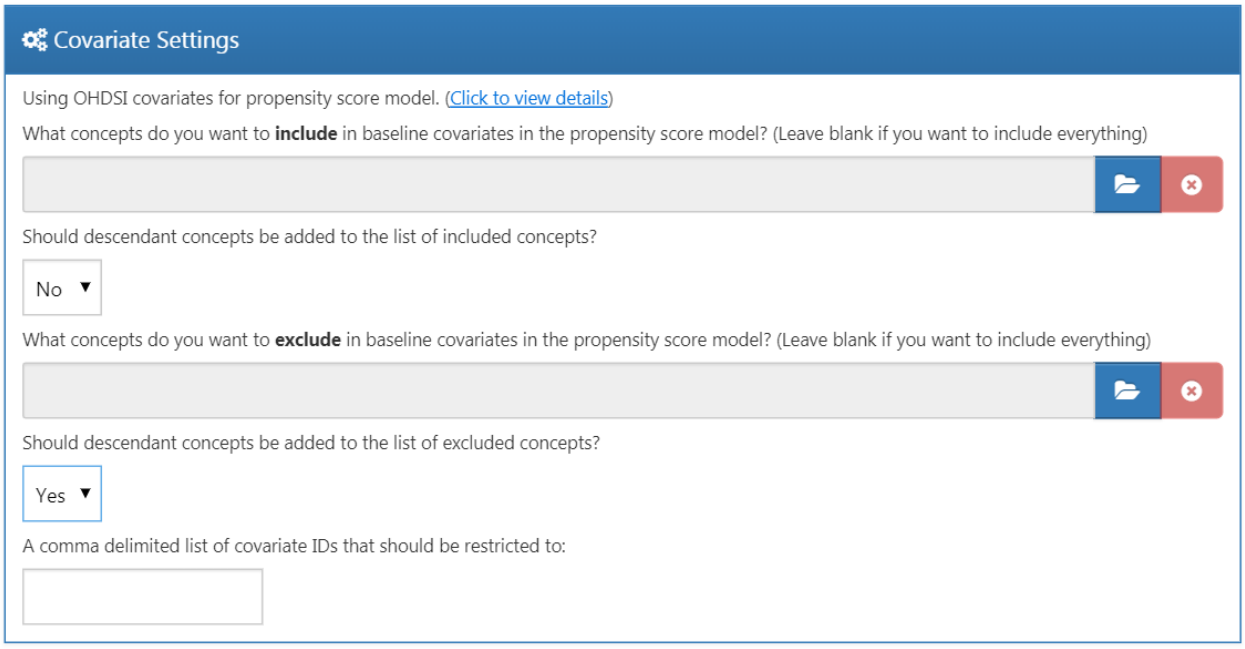


Figure 9.12: Covariate settings.

Time-at-risk is defined relative to the start and end dates of our target and comparator cohorts. In our example, we had set the cohort start date to start on treatment initiation, and cohort end date when exposure stops (for at least 30 days). We set the start of time-at-risk to one day after cohort start, so one day after treatment initiation. A reason to set the time-at-risk start to be later than the cohort start is because we may want to exclude outcome events that occur on the day of treatment initiation if we do not believe it biologically plausible they can be caused by the drug.

We set the end of the time-at-risk to the cohort end, so when exposure stops. We could choose to set the end date later if for example we believe events closely following treatment end may still be attributable to the exposure. In the extreme we could set the time-at-risk end to a large number of days (e.g. 99999) after the cohort end date, meaning we will effectively follow up subjects until observation end. Such a design is sometimes referred to as an *intent-to-treat* design.

A patient with zero days at risk adds no information, so the **minimum days at risk** is normally set at one day. If there is a known latency for the side effect, then this may be increased to get a more informative proportion. It can also be used to create a cohort more similar to that of a randomized trial it is being compared to (e.g., all the patients in the randomized trial were observed for at least N days).

A golden rule in designing a cohort study is to never use information that falls after the cohort start date to define the study population, as this may introduce bias. For example, if we require everyone to have at least a year of time-at-risk, we will likely have limited our analyses to those who tolerate the treatment well. This setting should therefore be used with extreme care.

Propensity score adjustment

We can opt to **trim** the study population, removing people with extreme PS values. We can choose to remove the top and bottom percentage, or we can remove subjects whose preference score falls outside the range we specify. Trimming the cohorts is generally not recommended because it requires discarding observations, which reduces statistical power. It may be desirable to trim in some cases, for example when using IPTW.

In addition to, or instead of trimming, we can choose to **stratify** or **match** on the propensity score. When stratifying we need to specify the **number of strata** and whether to select the strata based on the target, comparator, or entire study population. When matching we need to specify the **maximum number of**

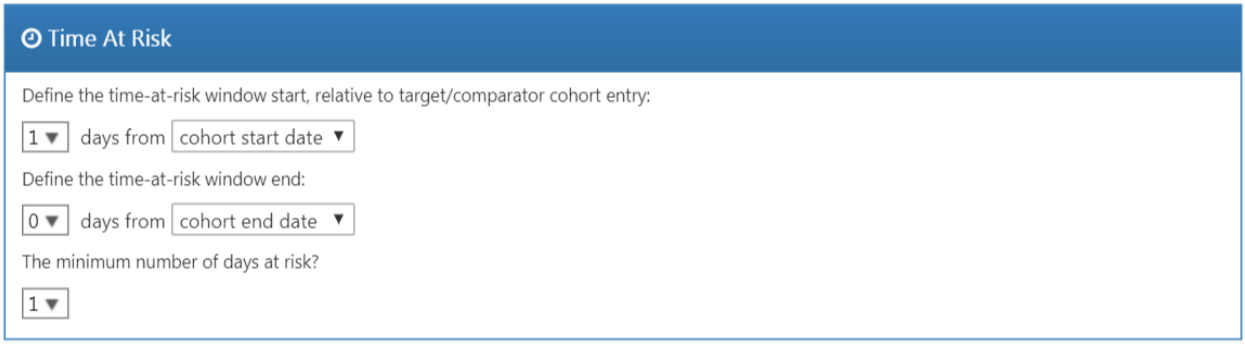


Figure 9.13: Time-at-risk settings.

people from the comparator group to match to each person in the target group. Typical values are 1 for one-on-one matching, or a large number (e.g. 100) for variable-ratio matching. We also need to specify the **caliper**: the maximum allowed difference between propensity scores to allow a match. The caliper can be defined on difference **caliper scales**:

- **The propensity score scale**: the PS itself
- **The standardized scale**: in standard deviations of the PS distributions
- **The standardized logit scale**: in standard deviations of the PS distributions after the logit transformation to make the PS more normally distributed.

In case of doubt, we suggest using the default values, or consult the work on this topic by Austin (2011).

Fitting large-scale propensity models can be computationally expensive, so we may want to restrict the data used to fit the model to just a sample of the data. By default the maximum size of the target and comparator cohort is set to 250,000. In most studies this limit will not be reached. It is also unlikely that more data will lead to a better model. Note that although a sample of the data may be used to fit the model, the model will be used to compute PS for the entire population.

Test each covariate for correlation with the target assignment? If any covariate has an unusually high correlation (either positive or negative), this will throw an error. This avoids lengthy calculation of a propensity model only to discover complete separation. Finding very high univariate correlation allows you to review the covariate to determine why it has high correlation and whether it should be dropped.

Figure 9.14 shows our choices for this study. Note that we select variable-ratio matching by setting the maximum number of people to match to 100.

Outcome model settings

First, we need to **specify the statistical model we will use to estimate the relative risk of the outcome between target and comparator cohorts**. We can choose between Cox, Poisson, and logistic regression, as discussed briefly in CohortMethod Section. For our example we choose a Cox proportional hazards model, which considers time to first event with possible censoring. Next, we need to specify **whether the regression should be conditioned on the strata**. One way to understand conditioning is to imagine a separate estimate is produced in each stratum, and then combined across strata. For one-to-one matching this is likely unnecessary and would just lose power. For stratification or variable-ratio matching it is required.

We can also choose to **add all covariates to the outcome model** to adjust the analysis. This can be done in addition or instead of using a propensity model. However, whereas there usually is ample data to fit a propensity model, with many people in both treatment groups, there is typically very little data to fit the outcome model, with only few people having the outcome. We therefore recommend keeping the outcome model as simple as possible and not include additional covariates.

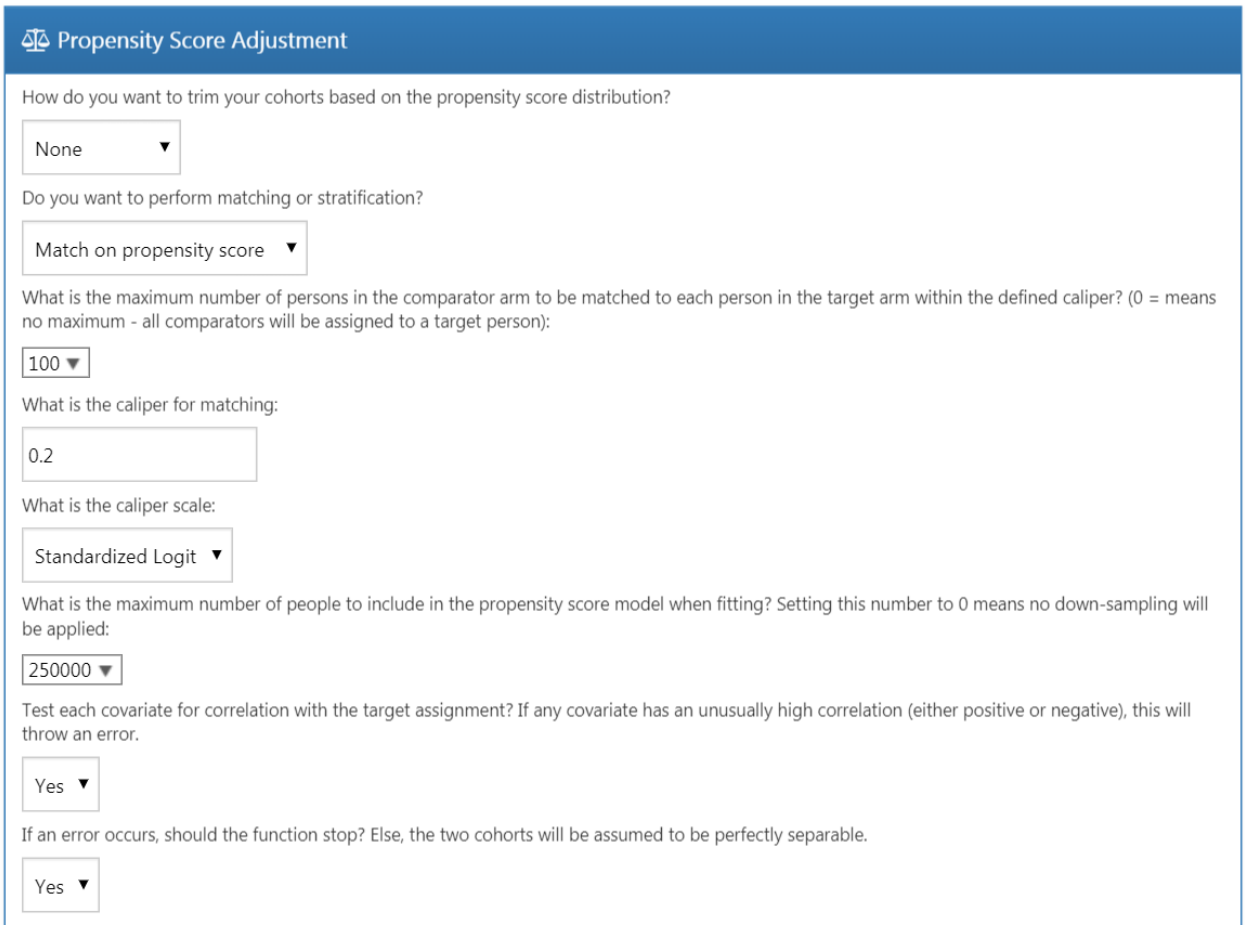


Figure 9.14: Propensity score adjustment settings.

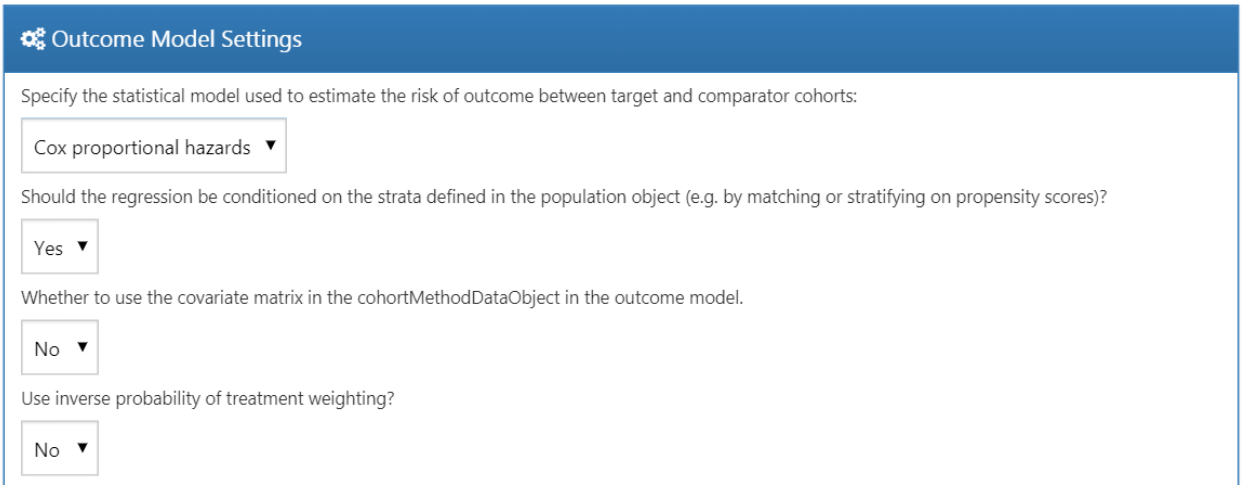


Figure 9.15: Outcome model settings.

Instead of stratifying or matching on the propensity score we can also choose to **use inverse probability of treatment weighting** (IPTW). If weighting is used it is often recommended to use some form of trimming to avoid extreme weights and therefore unstable estimates.

Figure 9.15 shows our choices for this study. Because we use variable-ratio matching, we must condition the regression on the strata (i.e. the matched sets).

9.7.3 Evaluation settings

As described in Method Validity Chapter, negative and positive controls should be included in our study to evaluate the operating characteristics, and perform empirical calibration.

Negative control outcome cohort definition

In Comparison Settings Section we selected a concept set representing the negative control outcomes. However, we need logic to convert concepts to cohorts to be used as outcomes in our analysis. ATLAS provides standard logic with three choices. The first choice is whether to **use all occurrences** or just the **first occurrence** of the concept. The second choice determines **whether occurrences of descendant concepts should be considered**. For example, occurrences of the descendant "ingrown nail of foot" can also be counted as an occurrence of the ancestor "ingrown nail." The third choice specifies which domains should be considered when looking for the concepts.

Positive control synthesis

In addition to negative controls we can also include positive controls, which are exposure-outcome pairs where a causal effect is believed to exist with known effect size. For various reasons real positive controls are problematic, so instead we rely on synthetic positive controls, derived from negative controls as described in Method Validity Chapter. Positive control synthesis is an advanced topic that we will skip for now.

9.7.4 Running the study package

Now that we have fully defined our study, we can export it as an executable R package. This package contains everything that is needed to execute the study at a site that has data in CDM. This includes the cohort definitions that can be used to instantiate the target, comparator and outcome cohorts, the negative control concept set and logic to create the negative control outcome cohorts, as well as the R code to execute the analysis. Before generating the package make sure to save your study, then click on the **Utilities** tab.

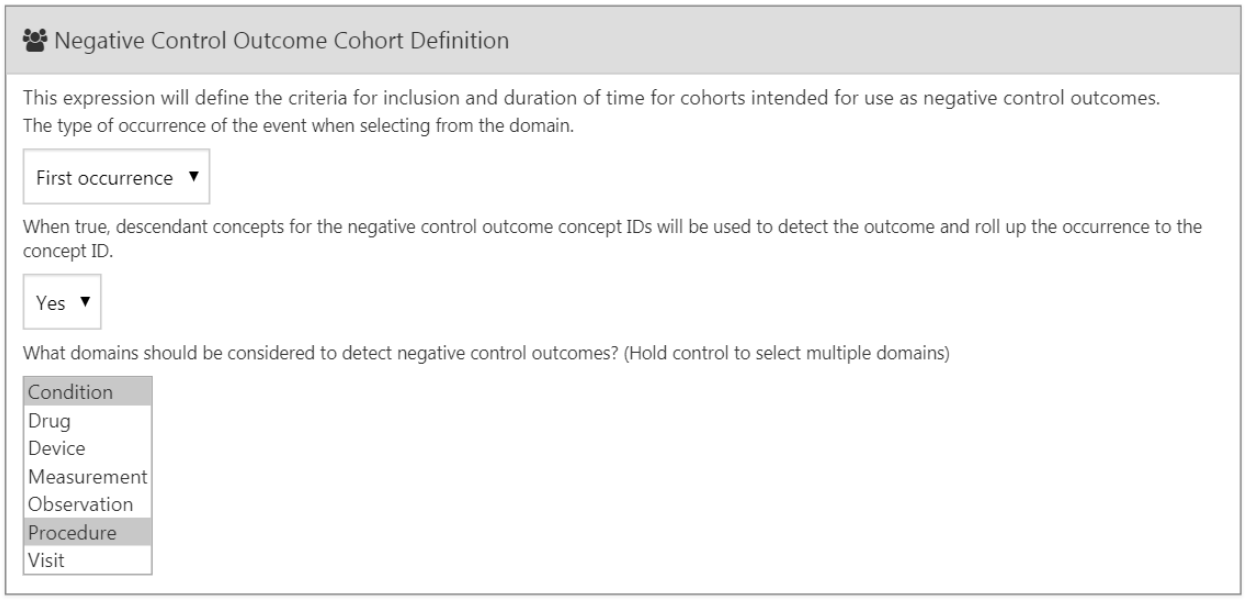
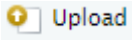


Figure 9.16: Negative control outcome cohort definition settings.

Here we can review the set of analyses that will be performed. As mentioned before, every combination of a comparison and an analysis setting will result in a separate analysis. In our example we have specified two analyses: ACEi versus THZ for AMI, and ACEi versus THZ for angioedema, both using propensity score matching.

We must provide a name for our package, after which we can click on "Download" to download the zip file. The zip file contains an R package, with the usual required folder structure for R packages. (Wickham, 2015) To use this package we recommend using R Studio. If you are running R Studio locally, unzip the file, and double click the .Rproj file to open it in R Studio. If you are running R Studio on an R studio server, click  to upload and unzip the file, then click on the .Rproj file to open the project.

Once you have opened the project in R Studio, you can open the README file, and follow the instructions. Make sure to change all file paths to existing paths on your system.

A common error message that may appear when running the study is "High correlation between covariate(s) and treatment detected." This indicates that when fitting the propensity model, some covariates were observed to be highly correlated with the exposure. Please review the covariates mentioned in the error message, and exclude them from the set of covariates if appropriate (see Variable Selection Section).

9.8 Implementing the study using R

Instead of using ATLAS to write the R code that executes the study, we can also write the R code ourselves. One reason we might want to do this is because R offers far greater flexibility than is exposed in ATLAS. If we for example wish to use custom covariates, or a linear outcome model, we will need to write some custom R code, and combine it with the functionality provided by the OHDSI R packages.

For our example study we will rely on the CohortMethod package to execute our study. CohortMethod extracts the necessary data from a database in the CDM and can use a large set of covariates for the propensity model. In the following example we first only consider angioedema as outcome. In Multiple Analyses Section we then describe how this can be extended to include AMI and the negative control outcomes.

9.8.1 Cohort instantiation

We first need to instantiate the target and outcome cohorts.

9.8.2 Data extraction

We first need to tell R how to connect to the server. CohortMethod uses the DatabaseConnector package, which provides a function called `createConnectionDetails`. Type `?createConnectionDetails` for the specific settings required for the various database management systems (DBMS). For example, one might connect to a PostgreSQL database using this code:

```
library(CohortMethod)
connDetails <- createConnectionDetails(dbms = "postgresql",
                                       server = "localhost/ohdsi",
                                       user = "joe",
                                       password = "supersecret")

cdmDbSchema <- "my_cdm_data"
cohortDbSchema <- "scratch"
cohortTable <- "my_cohorts"
cdmVersion <- "5"
```

The last four lines define the `cdmDbSchema`, `cohortDbSchema`, and `cohortTable` variables, as well as the CDM version. We will use these later to tell R where the data in CDM format live, where the cohorts of interest have been created, and what version CDM is used. Note that for Microsoft SQL Server, database schemas need to specify both the database and the schema, so for example `cdmDbSchema <- "my_cdm_data.dbo"`.

Now we can tell CohortMethod to extract the cohorts, construct covariates, and extract all necessary data for our analysis:

```
# target and comparator ingredient concepts:
aceI <- c(1335471,1340128,1341927,1363749,1308216,1310756,1373225,
          1331235,1334456,1342439)
thz <- c(1395058,974166,978555,907013)

# Define which types of covariates must be constructed:
cs <- createDefaultCovariateSettings(excludedCovariateConceptIds = c(aceI,
                                                                     thz),
                                     addDescendantsToExclude = TRUE)

#Load data:
cmData <- getDbCohortMethodData(connectionDetails = connectionDetails,
                                cdmDatabaseSchema = cdmDatabaseSchema,
                                oracleTempSchema = NULL,
                                targetId = 1,
                                comparatorId = 2,
                                outcomeIds = 3,
                                studyStartDate = "",
                                studyEndDate = "",
                                exposureDatabaseSchema = cohortDbSchema,
                                exposureTable = cohortTable,
                                outcomeDatabaseSchema = cohortDbSchema,
                                outcomeTable = cohortTable,
```

```
                                    cdmVersion = cdmVersion,
                                    firstExposureOnly = FALSE,
                                    removeDuplicateSubjects = FALSE,
                                    restrictToCommonPeriod = FALSE,
                                    washoutPeriod = 0,
                                    covariateSettings = cs)
cmData
```

```
## CohortMethodData object
##
## Treatment concept ID: 1
## Comparator concept ID: 2
## Outcome concept ID(s): 3
```

There are many parameters, but they are all documented in the CohortMethod manual. The `createDefaultCovariateSettings` function is described in the FeatureExtraction package. In short, we are pointing the function to the table containing our cohorts and specify which cohort definition IDs in that table identify the target, comparator and outcome. We instruct that the default set of covariates should be constructed, including covariates for all conditions, drug exposures, and procedures that were found on or before the index date. As mentioned in Cohort Method Section we must exclude the target and comparator treatments from the set of covariates, and here we achieve this by listing all ingredients in the two classes, and tell FeatureExtraction to also exclude all descendants, thus excluding all drugs that contain these ingredients.

All data about the cohorts, outcomes, and covariates are extracted from the server and stored in the `cohortMethodData` object. This object uses the package `ff` to store information in a way that ensures R does not run out of memory, even when the data are large, as mentioned in Section Big Data Support .

We can use the generic `summary()` function to view some more information of the data we extracted:

```
summary(cmData)
```

```
## CohortMethodData object summary
##
## Treatment concept ID: 1
## Comparator concept ID: 2
## Outcome concept ID(s): 3
##
## Treated persons: 67166
## Comparator persons: 35333
##
## Outcome counts:
##          Event count Person count
## 3                980          891
##
## Covariates:
## Number of covariates: 58349
## Number of non-zero covariate values: 24484665
```

Creating the `cohortMethodData` file can take considerable computing time, and it is probably a good idea to save it for future sessions. Because `cohortMethodData` uses `ff`, we cannot use R's regular save function. Instead, we'll have to use the `saveCohortMethodData()` function:

```
saveCohortMethodData(cmData, "AceiVsThzForAngioedema")
```

We can use the `loadCohortMethodData()` function to load the data in a future session.

Defining new users

Typically, a new user is defined as first time use of a drug (either target or comparator), and typically a washout period (a minimum number of days prior to first use) is used to increase the probability that it is truly first use. When using the CohortMethod package, you can enforce the necessary requirements for new use in three ways:

1. When defining the cohorts.
2. When loading the cohorts using the `getDbCohortMethodData` function, you can use the `firstExposureOnly`, `removeDuplicateSubjects`, `restrictToCommonPeriod`, and `washoutPeriod` arguments.
3. When defining the study population using the `createStudyPopulation` function (see below).

The advantage of option 1 is that the input cohorts are already fully defined outside of the CohortMethod package, and external cohort characterization tools can be used on the same cohorts used in this analysis. The advantage of options 2 and 3 is that they save you the trouble of limiting to first use yourself, for example allowing you to directly use the DRUG_ERA table in the CDM. Option 2 is more efficient than 3, since only data for first use will be fetched, while option 3 is less efficient but allows you to compare the original cohorts to the study population.

9.8.3 Defining the study population

Typically, the exposure cohorts and outcome cohorts will be defined independently of each other. When we want to produce an effect size estimate, we need to further restrict these cohorts and put them together, for example by removing exposed subjects that had the outcome prior to exposure, and only keeping outcomes that fall within a defined risk window. For this we can use the `createStudyPopulation` function:

```
studyPop <- createStudyPopulation(cohortMethodData = cmData,
                                  outcomeId = 3,
                                  firstExposureOnly = FALSE,
                                  restrictToCommonPeriod = FALSE,
                                  washoutPeriod = 0,
                                  removeDuplicateSubjects = "remove all",
                                  removeSubjectsWithPriorOutcome = TRUE,
                                  minDaysAtRisk = 1,
                                  riskWindowStart = 1,
                                  addExposureDaysToStart = FALSE,
                                  riskWindowEnd = 0,
                                  addExposureDaysToEnd = TRUE)
```

Note that we've set `firstExposureOnly` and `removeDuplicateSubjects` to FALSE, and `washoutPeriod` to 0 because we already applied those criteria in the cohort definitions. We specify the outcome ID we will use, and that people with outcomes prior to the risk window start date will be removed. The risk window is defined as starting on the day after the cohort start date (`riskWindowStart = 1` and `addExposureDaysToStart = FALSE`), and the risk windows ends when the cohort exposure ends (`riskWindowEnd = 0` and `addExposureDaysToEnd = TRUE`), which was defined as the end of exposure in the cohort definition. Note that the risk windows are automatically truncated at the end of observation or the study end date. We also remove subjects who have no time at risk. To see how many people are left in the study population we can always use the `getAttritionTable` function:

```
getAttritionTable(studyPop)
```

```
##                    description targetPersons comparatorPersons   ...
## 1             Original cohorts         67212             35379   ...
## 2 Removed subs in both cohorts         67166             35333   ...
## 3             No prior outcome         67061             35238   ...
## 4 Have at least 1 days at risk         66780             35086   ...
```

9.8.4 Propensity scores

We can fit a propensity model using the covariates constructed by the `getDbcohortMethodData()` function, and compute a PS for each person:

```
ps <- createPs(cohortMethodData = cmData, population = studyPop)
```

The `createPs` function uses the Cyclops package to fit a large-scale regularized logistic regression. To fit the propensity model, Cyclops needs to know the hyperparameter value which specifies the variance of the prior. By default Cyclops will use cross-validation to estimate the optimal hyperparameter. However, be aware that this can take a really long time. You can use the `prior` and `control` parameters of the `createPs` function to specify Cyclops' behavior, including using multiple CPUs to speed-up the cross-validation.

Here we use the PS to perform variable-ratio matching:

```
matchedPop <- matchOnPs(population = ps, caliper = 0.2,
                        caliperScale = "standardized logit", maxRatio = 100)
```

Alternatively, we could have used the PS in the `trimByPs`, `trimByPsToEquipoise`, or `stratifyByPs` functions.

9.8.5 Outcome models

The outcome model is a model describing which variables are associated with the outcome. Under strict assumptions, the coefficient for the treatment variable can be interpreted as the causal effect. In this case we fit a Cox proportional hazards model, conditioned (stratified) on the matched sets:

```
outcomeModel <- fitOutcomeModel(population = matchedPop,
                                modelType = "cox",
                                stratified = TRUE)
outcomeModel
```

```
## Model type: cox
## Stratified: TRUE
## Use covariates: FALSE
## Use inverse probability of treatment weighting: FALSE
## Status: OK
##
##           Estimate lower .95 upper .95   logRr seLogRr
## treatment   4.3203    2.4531    8.0771 1.4633   0.304
```

9.8.6 Running multiple analyses

Often we want to perform more than one analysis, for example for multiple outcomes including negative controls. The CohortMethod offers functions for performing such studies efficiently. This is described in detail in the package vignette on running multiple analyses. Briefly, assuming the outcome of interest and negative control cohorts have already been created, we can specify all target-comparator-outcome combinations we wish to analyze:

```
# Outcomes of interest:
ois <- c(3, 4) # Angioedema, AMI

# Negative controls:
ncs <- c(434165,436409,199192,4088290,4092879,44783954,75911,137951,77965,
         376707,4103640,73241,133655,73560,434327,4213540,140842,81378,432303,
         4201390,46269889,134438,78619,201606,76786,4115402,45757370,433111
         433527,4170770,4092896,259995,40481632,4166231,433577,4231770,440329,
         4012570,4012934,441788,4201717,374375,4344500,139099,444132,196168,
         432593,434203,438329,195873,4083487,4103703,4209423,377572,40480893,
         136368,140648,438130,4091513,4202045,373478,46286594,439790,81634,
         380706,141932,36713918,443172,81151,72748,378427,437264,194083,
         140641,440193,4115367)

tcos <- createTargetComparatorOutcomes(targetId = 1,
                                       comparatorId = 2,
                                       outcomeIds = c(ois, ncs))

tcosList <- list(tcos)
```

Next, we specify what arguments should be used when calling the various functions described previously in our example with one outcome:

```
aceI <- c(1335471,1340128,1341927,1363749,1308216,1310756,1373225,
          1331235,1334456,1342439)
thz <- c(1395058,974166,978555,907013)

cs <- createDefaultCovariateSettings(excludedCovariateConceptIds = c(aceI,
                                                                     thz),
                                     addDescendantsToExclude = TRUE)

cmdArgs <- createGetDbCohortMethodDataArgs(
  studyStartDate = "",
  studyEndDate = "",
  firstExposureOnly = FALSE,
  removeDuplicateSubjects = FALSE,
  restrictToCommonPeriod = FALSE,
  washoutPeriod = 0,
  covariateSettings = cs)

spArgs <- createCreateStudyPopulationArgs(
  firstExposureOnly = FALSE,
  restrictToCommonPeriod = FALSE,
  washoutPeriod = 0,
  removeDuplicateSubjects = "remove all",
  removeSubjectsWithPriorOutcome = TRUE,
```

```
  minDaysAtRisk = 1,
  riskWindowStart = 1,
  addExposureDaysToStart = FALSE,
  riskWindowEnd = 0,
  addExposureDaysToEnd = TRUE)

psArgs <- createCreatePsArgs()

matchArgs <- createMatchOnPsArgs(
  caliper = 0.2,
  caliperScale = "standardized logit",
  maxRatio = 100)

fomArgs <- createFitOutcomeModelArgs(
  modelType = "cox",
  stratified = TRUE)
```

We then combine these into a single analysis settings object, which we provide a unique analysis ID and some description. We can combine one or more analysis settings objects into a list:

```
cmAnalysis <- createCmAnalysis(
  analysisId = 1,
  description = "Propensity score matching",
  getDbCohortMethodDataArgs = cmdArgs,
  createStudyPopArgs = spArgs,
  createPs = TRUE,
  createPsArgs = psArgs,
  matchOnPs = TRUE,
  matchOnPsArgs = matchArgs
  fitOutcomeModel = TRUE,
  fitOutcomeModelArgs = fomArgs)

cmAnalysisList <- list(cmAnalysis)
```

We can now run the study including all comparisons and analysis settings:

```
result <- runCmAnalyses(connectionDetails = connectionDetails,
                        cdmDatabaseSchema = cdmDatabaseSchema,
                        exposureDatabaseSchema = cohortDbSchema,
                        exposureTable = cohortTable,
                        outcomeDatabaseSchema = cohortDbSchema,
                        outcomeTable = cohortTable,
                        cdmVersion = cdmVersion,
                        outputFolder = outputFolder,
                        cmAnalysisList = cmAnalysisList,
                        targetComparatorOutcomesList = tcosList)
```

The `result` object contains references to all the artifacts that were created. For example, we can retrieve the outcome model for AMI:

```
omFile <- result$outcomeModelFile[result$targetId == 1 &
                                  result$comparatorId == 2 &
```

```
                                    result$outcomeId == 4 &
                                    result$analysisId == 1]
outcomeModel <- readRDS(file.path(outputFolder, omFile))
outcomeModel
```

```
## Model type: cox
## Stratified: TRUE
## Use covariates: FALSE
## Use inverse probability of treatment weighting: FALSE
## Status: OK
##
##           Estimate lower .95 upper .95   logRr seLogRr
## treatment   1.1338    0.5921    2.1765 0.1256   0.332
```

We can also retrieve the effect size estimates for all outcomes with one command:

```
summ <- summarizeAnalyses(result, outputFolder = outputFolder)
head(summ)
```

```
##     analysisId targetId comparatorId outcomeId        rr    ci95lb  ...
## 1            1        1            2     72748 0.9734698 0.5691589  ...
## 2            1        1            2     73241 0.7067981 0.4009951  ...
## 3            1        1            2     73560 1.0623951 0.7187302  ...
## 4            1        1            2     75911 0.9952184 0.6190344  ...
## 5            1        1            2     76786 1.0861746 0.6730408  ...
## 6            1        1            2     77965 1.1439772 0.5173222  ...
```

9.9 Study outputs

Our estimates are only valid if several assumptions have been met. We use a wide set of diagnostics to evaluate whether this is the case. These are available in the results produced by the R package generated by ATLAS, or can be generated on the fly using specific R functions.

9.9.1 Propensity scores and model

We first need to evaluate whether the target and comparator cohort are to some extent comparable. For this we can compute the Area Under the Receiver Operator Curve (AUC) statistic for the propensity model. An AUC of 1 indicates the treatment assignment was completely predictable based on baseline covariates, and that the two groups are therefore incomparable. We can use the `computePsAuc` function to compute the AUC, which in our example is 0.79. Using the `plotPs` function, we can also generate the preference score distribution as shown in Figure 9.17. Here we see that for many people the treatment they received was predictable, but there is also a large amount of overlap, indicating that adjustment can be used to select comparable groups.

In general it is a good idea to also inspect the propensity model itself, and especially so if the model is very predictive. That way we may discover which variables are most predictive. Table 9.7 shows the top predictors in our propensity model. Note that if a variable is too predictive, the CohortMethod package will throw an informative error rather than attempt to fit a model that is already known to be perfectly predictive.

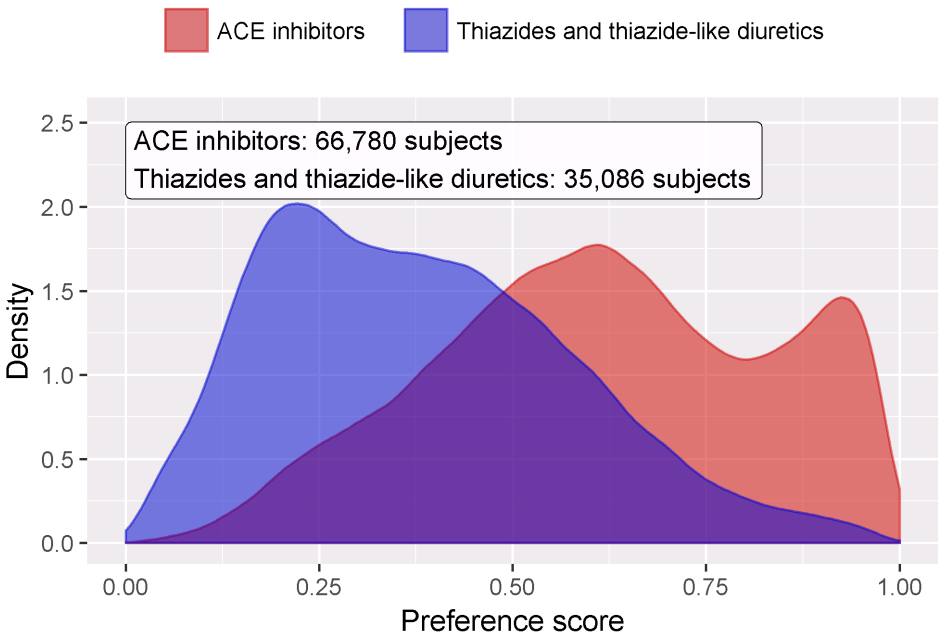


Figure 9.17: Preference score distribution.

Table 9.7: Top 10 predictors in the propensity model for ACEi and THZ. Positive values mean subjects with the covariate are more likely to receive the target treatment.

Beta	Covariate
-1.42	condition_era group during day -30 through 0 days relative to index: Edema
-1.11	drug_era group during day 0 through 0 days relative to index: Potassium Chloride
0.68	age group: 05-09
0.64	measurement during day -365 through 0 days relative to index: Renin
0.63	condition_era group during day -30 through 0 days relative to index: Urticaria
0.57	condition_era group during day -30 through 0 days relative to index: Proteinuria
0.55	drug_era group during day -365 through 0 days relative to index: INSULINS AND ANALOGUES
-0.54	race = Black or African American
0.52	(Intercept)
0.50	gender = MALE

If a variable is found to be highly predictive, there are two possible conclusions: Either we find that the variable is clearly part of the exposure itself and should be removed before fitting the model, or else we must conclude that the two populations are truly incomparable, and the analysis must be stopped.“‘

9.9.2 Covariate balance

The goal of using PS is to make the two groups comparable (or at least to select comparable groups). We must verify whether this is achieved, for example by checking whether the baseline covariates are indeed balanced after adjustment. We can use the `computeCovariateBalance` and `plotCovariateBalanceScatterPlot` functions to generate Figure 9.18. One rule-of-thumb to use is that no covariate may have an absolute standardized difference of means greater than 0.1 after propensity score adjustment. Here we see that although there was substantial imbalance before matching, after matching we meet this criterion.

9.9.3 Follow up and power

Before fitting an outcome model, we might be interested to know whether we have sufficient power to detect

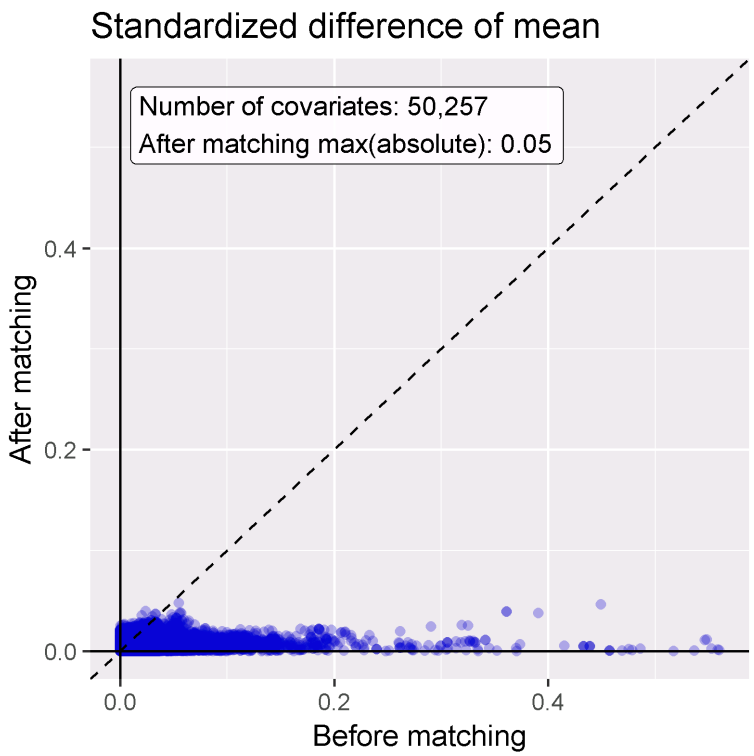


Figure 9.18: Covariate balance, showing the absolute standardized difference of mean before and after propensity score matching. Each blue dot represents a covariate.

proportionality of hazards holds. The Kaplan-Meier plot automatically adjusts for stratification or weighting by PS. In this case, because variable-ratio matching is used, the survival curve for the comparator groups is adjusted to mimick what the curve had looked like for the target group had they been exposued to the comparator instead.

9.9.5 Effect size estimate

We observe a hazard ratio of 4.32 (95% confidence interval: 2.45 - 8.08) for angioedema, which tells us that ACEi appears to increase the risk of angioedema compared to THZ. Similarly, we observe a hazard ratio of 1.13 (95% confidence interval: 0.59 - 2.18) for AMI, suggesting little or no effect for AMI. Our diagnostics, as reviewed earlier, give no reason for doubt. However, ultimately the quality of this evidence, and whether we choose to trust it, depends on many factors that are not covererd by the study diagnostics as described in Evidence Quality Chapter.

9.10 Summary

- Population-level estimation .
- , , .
- (counterfactual) .
- OHDSI Methods Library counterfactual .

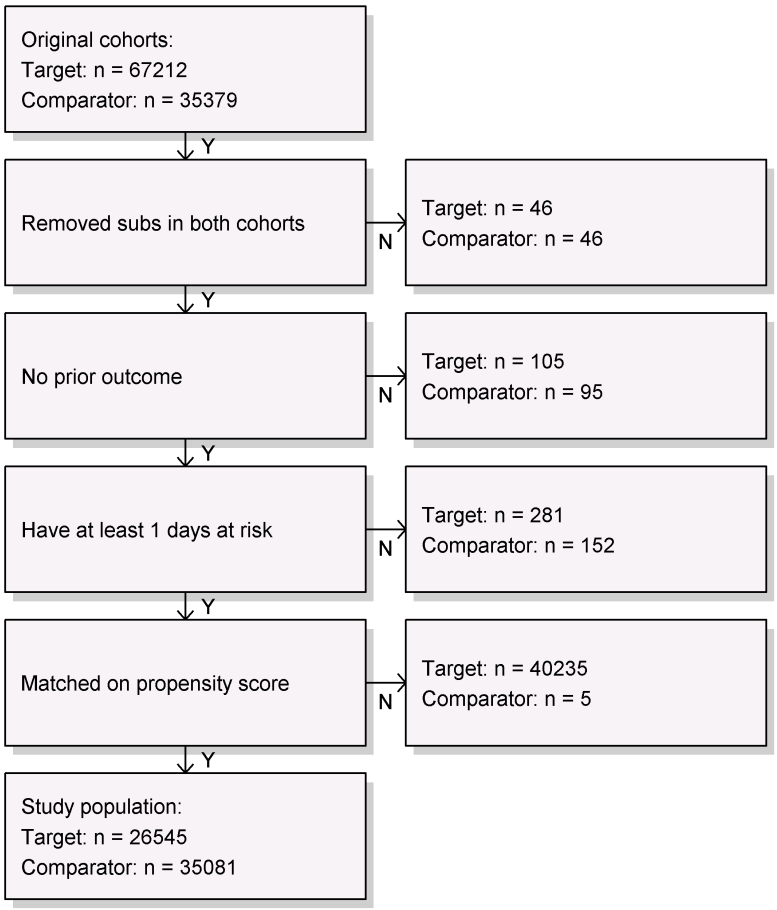


Figure 9.19: Attrition diagram. The counts shown at the top are those that meet our target and comparator cohort definitions. The counts at the bottom are those that enter our outcome model, in this case a Cox regression.

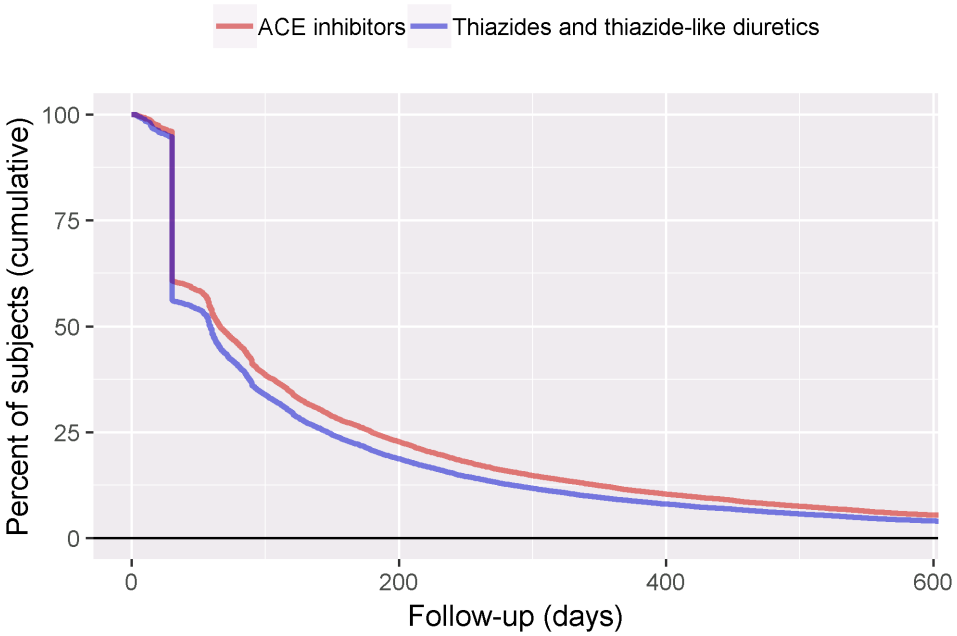


Figure 9.20: Distribution of follow-up time for the target and comparator cohorts.

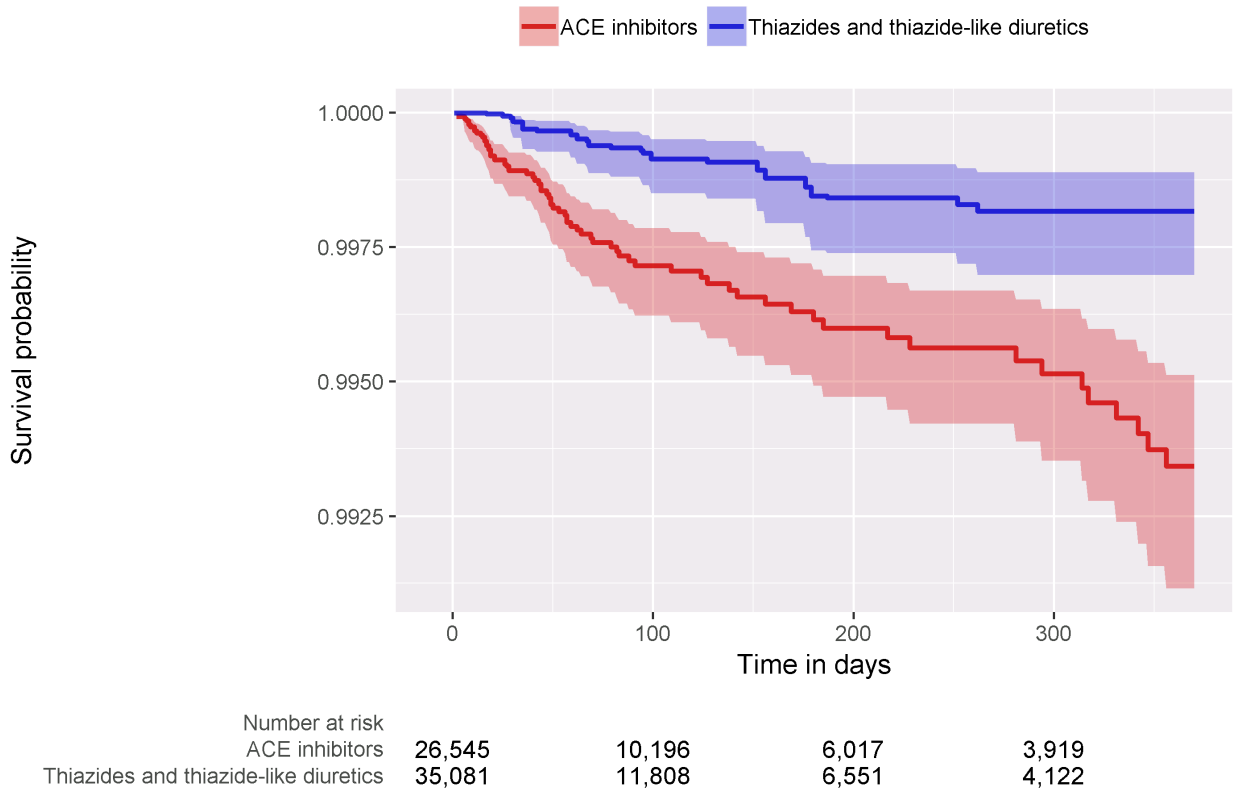


Figure 9.21: Kaplan-Meier plot.

Population-level estimation aims to infer causal effects from observational data. The **counterfactual**, what would have happened if the subject had received an alternative exposure or no exposure, cannot be observed. Different designs aim to construct the counterfactual in different ways. The various designs as implemented in the OHDSI Methods Library provide diagnostics to evaluate whether the assumptions for creating an appropriate counterfactual have been met.

9.11 Excercises

Note: The exercises still have to be defined. The idea is to require readers to define a study that estimates the effect of celecoxib on GI bleed, compared to diclofenac. For this they must use the Eunomia package, which is still under development.

Chapter 10

Patient-Level Prediction

- Patient-Level Prediction : OHDSI past event Patient-Level Prediction .

Chapter 11

Extension of CDM

11.1 Genomic CDM

11.2 Radiology CDM

11.3 AEGIS

Bibliography

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161.

Farrington, C. P. (1995). Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*, 51(1):228–235.

Farrington, C. P., Anaya-Izquierdo, K., Whitaker, H. J., Hocine, M. N., Douglas, I., and Smeeth, L. (2011). Self-controlled case series analysis with event-dependent observation periods. *Journal of the American Statistical Association*, 106(494):417–426.

Hernan, M. A., Hernandez-Diaz, S., Werler, M. M., and Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am. J. Epidemiol.*, 155(2):176–184.

Hernan, M. A. and Robins, J. M. (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am. J. Epidemiol.*, 183(8):758–764.

Maclure, M. (1991). The case-crossover design: a method for studying transient effects on the risk of acute events. *Am. J. Epidemiol.*, 133(2):144–153.

Magid, D. J., Shetterly, S. M., Margolis, K. L., Tavel, H. M., O'Connor, P. J., Selby, J. V., and Ho, P. M. (2010). Comparative effectiveness of angiotensin-converting enzyme inhibitors versus beta-blockers as second-line therapy for hypertension. *Circ Cardiovasc Qual Outcomes*, 3(5):453–458.

Powers, B. J., Coeytaux, R. R., Dolor, R. J., Hasselblad, V., Patel, U. D., Yancy, W. S., Gray, R. N., Irvine, R. J., Kendrick, A. S., and Sanders, G. D. (2012). Updated report on comparative effectiveness of ACE inhibitors, ARBs, and direct renin inhibitors for patients with essential hypertension: much more data, little new information. *J Gen Intern Med*, 27(6):716–729.

Rassen, J. A., Shelat, A. A., Myers, J., Glynn, R. J., Rothman, K. J., and Schneeweiss, S. (2012). One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol Drug Saf*, 21 Suppl 2:69–80.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188.

Ryan, P. B., Schuemie, M. J., and Madigan, D. (2013). Empirical performance of a self-controlled cohort method: lessons for developing a risk identification and analysis system. *Drug Saf*, 36 Suppl 1:95–106.

Sabroe, R. A. and Black, A. K. (1997). Angiotensin-converting enzyme (ACE) inhibitors and angio-oedema. *Br. J. Dermatol.*, 136(2):153–158.

Schneeweiss, S. (2018). Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clin Epidemiol*, 10:771–788.

Simpson, S. E., Madigan, D., Zorych, I., Schuemie, M. J., Ryan, P. B., and Suchard, M. A. (2013). Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*, 69(4):893–902.

Suchard, M. A., Simpson, S. E., Zorych, I., Ryan, P. B., and Madigan, D. (2013). Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Trans. Model. Comput. Simul.*, 23(1):10:1–10:17.

Suissa, S. (1995). The case-time-control design. *Epidemiology*, 6(3):248–253.

Tian, Y., Schuemie, M. J., and Suchard, M. A. (2018). Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int J Epidemiol*, 47(6):2005–2014.

Toh, S., Reichman, M. E., Houstoun, M., Ross Southworth, M., Ding, X., Hernandez, A. F., Levenson, M., Li, L., McCloskey, C., Shoaibi, A., Wu, E., Zornberg, G., and Hennessy, S. (2012). Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. *Arch. Intern. Med.*, 172(20):1582–1589.

Vandenbroucke, J. P. and Pearce, N. (2012). Case-control studies: basic concepts. *Int J Epidemiol*, 41(5):1480–1489.

Walker, A. M., Patrick, A. R., Lauer, M. S., Hornbrook, M. C., Marin, M. G., Platt, R., Roger, V. L., Stang, P., and Schneeweiss, S. (2013). A tool for assessing the feasibility of comparative effectiveness research. *Comp Eff Res*, 3:11–20.

Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Dennison Himmelfarb, C., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E. J., Muntner, P., Ovbiagele, B., Smith, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., Thomas, R. J., Williams, K. A., Williamson, J. D., and Wright, J. T. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*, 138(17):e426–e483.

Whitaker, H. J., Farrington, C. P., Spiessens, B., and Musonda, P. (2006). Tutorial in biostatistics: the self-controlled case series method. *Stat Med*, 25(10):1768–1797.

Wickham, H. (2015). *R Packages*. O'Reilly Media, Inc., 1st edition.

Zaman, M. A., Oparil, S., and Calhoun, D. A. (2002). Drugs targeting the renin-angiotensin-aldosterone system. *Nat Rev Drug Discov*, 1(8):621–636.