

The Book of OHDSI

Observational Health Data Science and Informatics

2019-05-28

Contents

Welcome	7
I The OHDSI Community	9
1 Mission, vision, values	11
1.1 Our Mission	11
1.2 Our Vision	11
1.3 Our Objectives	11
2 Collaborators	13
3 Open Science	15
4 Where to begin	17
II Uniform Data Representation	19
5 The Common Data Model	21
5.1 Design Principles	21
5.2 Data Model Conventions	22
5.3 OMOP CDM Standardized Tables	27
6 Standardized Vocabularies	43
7 Extract Transform Load	45
III Data Analytics	47
8 Data Analytics Use Cases	49
9 OHDSI Analytics Tools	51

10 SQL and R	53
11 Building the building blocks: cohorts	55
12 Characterization	57
13 Population-level estimation	59
13.1 Study designs	60
13.2 Cohort study implementation	64
13.3 Advanced topics	64
13.4 Excercises	64
14 Patient Level Prediction	65
14.1 Study specification	68
14.2 Study implementation	76
14.3 Internal validation	89
14.4 External validation	101
14.5 Journal paper generation	102
14.6 Other functionality	104
14.7 Demos	104
14.8 Acknowledgments	105
Appendix 1: Study population settings details	106
IV Evidence Quality	109
15 Evidence Quality	111
16 Data Quality	113
16.1 Introduction	113
16.2 Achilles Heel tool	114
16.3 Study-specific checks	115
17 Clinical Validity	117
18 Software Validity	119
19 Method Validity	121
19.1 Design-specific diagnostics	121
19.2 Diagnostics for all estimation	124
19.3 Diagnostics for all prediction	127
19.4 Method validation in practice	127
V OHDSI Studies	129
20 Study steps	131

CONTENTS	5
21 OHDSI Network Research	133
21.1 OHDSI Network Study Examples	134
21.2 Excercises	134
A Glossary	135

Welcome

This is a book about OHDSI, and is currently very much under development.
The book is written in RMarkdown with bookdown. It is automatically rebuilt
from source by travis.

Part I

The OHDSI Community

Chapter 1

Mission, vision, values

1.1 Our Mission

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

1.2 Our Vision

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

1.3 Our Objectives

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

- **Openness:** We strive to make all our community's proceeds open and publicly accessible, including the methods, tools and the evidence that we generate.
- **Beneficence:** We seek to protect the rights of individuals and organizations within our community at all times.

Chapter 2

Collaborators

History of OHDSI

Map of collaborators Forums Wiki Workgroups and chapters Symposia and
hack-a-thons

Governance at local sites

Chapter 3

Open Science

Mention FAIR principles?

Chapter 4

Where to begin

This chapter will discuss where to begin if one is new in OHDSI. For various activities, we can describe how one might get started.

For example, if interested in doing a network study, these are the steps. Same for interests in methods research, grant writing, etc.

Add a diagram that shows what tools are used for which steps?

Part II

Uniform Data Representation

Chapter 5

The Common Data Model

No single observational data source provides a comprehensive view of the clinical data a patient accumulates while receiving healthcare, and therefore none can be sufficient to meet all expected outcome analysis needs. This explains the need for assessing and analyzing multiple data sources concurrently using a common data standard. This standard is provided by the OMOP Common Data Model (CDM).

The CDM is designed to support the conduct of research to identify and evaluate associations between interventions (drug exposure, procedures, healthcare policy changes etc.) and outcomes caused by these interventions (condition occurrences, procedures, drug exposure etc.). Outcomes can be efficacious (benefit) or adverse (safety risk). Often times, specific patient cohorts (e.g., those taking a certain drug or suffering from a certain disease) may be defined for treatments or outcomes, using clinical events (diagnoses, observations, procedures, etc.) that occur in predefined temporal relationships to each other. The CDM, combined with its standardized content (via the Standardized Vocabularies), will ensure that research methods can be systematically applied to produce meaningfully comparable and reproducible results.

5.1 Design Principles

The CDM is designed to include all observational health data elements (experiences of the patient receiving health care) that are relevant for analysis use cases to support the generation of reliable scientific evidence about disease natural history, healthcare delivery, effects of medical interventions, the identification of demographic information, health care interventions and outcomes.

Therefore, the CDM is designed to store observational data to allow for research, under the following principles:

- **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.

5.2 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.

5.2.1 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.

5.2.2 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.

5.2.3 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.

5.2.4 General conventions of fields

Variable names across all tables follow one convention:

Notation	Description
<code>_SOURCE_VALUE</code>	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.
<code>_ID</code>	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.
<code>_CONCEPT_ID</code>	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'
<code>_SOURCE_CONCEPT_ID</code>	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.
<code>_TYPE_CONCEPT_ID</code>	Denotes 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'

5.2.5 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.

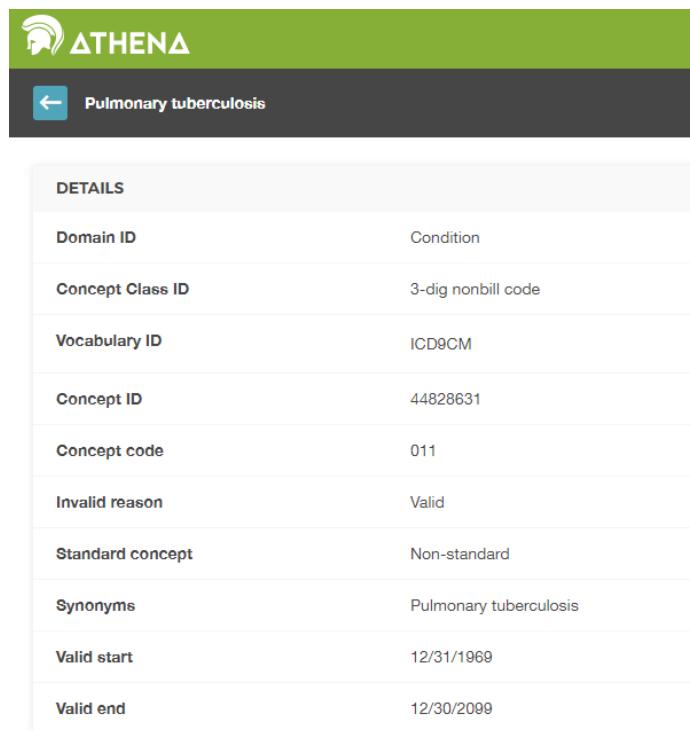
5.2.6 Difference between Concept IDs and Source Values

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 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 5.1 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



The image shows a screenshot of the ATHENA interface. At the top, there is a green header bar with the ATHENA logo on the left and a search bar containing the text "Pulmonary tuberculosis". Below the header is a dark grey navigation bar with a back arrow icon and the text "Pulmonary tuberculosis". The main content area is a table titled "DETAILS" with the following rows:

DETAILS	
Domain ID	Condition
Concept Class ID	3-dig nonbill code
Vocabulary ID	ICD9CM
Concept ID	44828631
Concept code	011
Invalid reason	Valid
Standard concept	Non-standard
Synonyms	Pulmonary tuberculosis
Valid start	12/31/1969
Valid end	12/30/2009

Figure 5.1: ICD9CM code for Pulmonary Tuberculosis

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 5.2: SNOMED code for Pulmonary Tuberculosis

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 5.2 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.

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

5.3 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 Chapter 6.

5.3.1 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

Lauren had been experiencing endometriosis symptoms for many years; however, it took a ruptured cyst in her ovary before she was diagnosed.

*"Every step of this painful journey
I've had to convince everyone how
much pain I was in."*

Figure 5.3: Read more about Lauren and endometriosis at <https://www.endometriosis-uk.org/laurens-story>

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.



5.3.2 PERSON table

As the Common Data Model is a person-centric model (see section 5.2.1) 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

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.

Column Name	Value	Explanation
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		Same principle as gender_source_concept_id.

5.3.3 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

Encounter_ID	Start_Date	Stop_Date	EncounterClass
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_d	2010-01-06	This is the start date of her earliest encounter on record.
observation_period_end_da	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’.

5.3.4 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?

Reviewing 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
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

Column Name	Value	Explanation
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		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		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).
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’.

Column Name	Value	Explanation
discharge_to_concept0_id		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_0value		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		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)

5.3.5 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		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
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		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.

Column	Value	Explanation
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 dysmenorrhea 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.

5.3.6 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_time	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.

Column	Value	Explanation
drug_exposure_end_date	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.
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.

Column	Value	Explanation
route_concept_id	4132161	This concept is meant to represent the route of the drug the patient 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.

Column	Value	Explanation
dose_unit_source_value	NULL	The information about the dose unit as detailed in the source.

5.3.7 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.

Column	Value	Explanation
procedure_type_concept_38000275	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.

Column	Value	Explanation
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	4127451	This is the concept_id that represents the procedure source value.
modifier_source_value	NULL	The source code for the modifier as it appears in the source data.

Chapter 6

Standardized Vocabularies

The OMOP Standardized Vocabulary: Christian's (almost) finished paper +
<http://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary>

Chapter 7

Extract Transform Load

Leads: Mui van Zandt & Clair Blacketer

Business Rules and Conventions: From the CDM Wiki + Themis

Conversion to OMOP CDM (ETL - Extract, Transform, Load): http://www.ohdsi.org/web/wiki/doku.php?id=documentation:etl_best_practices

- WhiteRabbit and Rabbit-in-a-Hat: <http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:whiterabbit>
- Usagi: <http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:usagi>
- Achilles: <http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:achilles>
- Athena: http://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary_etl

Mapping and QA of codes to Standard Concepts

- Mapping codes locally versus through the OHDSI Standard Vocabularies
- Usagi
- Systematic mapping of Drug codes
- Systematic mapping of Condition codes
- Systematic mapping of Procedure codes
- Systematic mapping of other codes

Part III

Data Analytics

Chapter 8

Data Analytics Use Cases

- Introduction

The OHDSI collaboration focuses on generating reliable evidence from real-world healthcare data, typically in the form of claims databases or electronic health record databases. The use cases that OHDSI focuses on fall into three major buckets and we describe these below. Note, for all the use cases, the evidence we generate inherits the limitations of the data; we discuss these limitations at length in Chapters X, Y, and Z.

- Theory

1. Population characterization

We can use the data to provide answers to questions about the characteristics of the patients in each database, the practice of healthcare, and study how these things change over time.

The data can provide answers to questions like: - for patients newly diagnosed with atrial fibrillation, how many receive a prescription for warfarin? - what is the average age of patients who undergo hip arthroplasty?

2. Population-level estimation

To a limited extent, the data can support causal inferences about the effects of healthcare interventions.

The data can provide answers to questions like: - for patients newly diagnosed with atrial fibrillation, in the first year after therapy initiation, does warfarin cause more major bleeds than dabigatran? - Does the causal effect of metformin on diarrhea vary by age?

3. Patient-Level prediction

Based on the collected patient health histories in the database, we can make patient-level predictions about future health events. - for a specific patient newly diagnosed with atrial fibrillation, in the first year after therapy initiation with warfarin, what is the probability the patient suffers an ischemic stroke?

These tasks overlap to a certain extent. For example, an important use-case for prediction is to predict an outcome for a specific patient had drug A been prescribed and also predict the same outcome had drug B been prescribed. Let's assume that in reality only one of these drugs is prescribed (say drug A) so we get to see whether the outcome following treatment with A actually occurs. Since drug B was not prescribed, the outcome following treatment B, while predictable, is "counterfactual" since it is not ever observed. Each of these prediction tasks falls under patient-level prediction. However, the difference between (or ratio of) the two outcomes is a unit-level *causal* effect.

There are many important healthcare questions for which OHDSI databases cannot provide answers. These include:

- Causal effects of interventions compared to placebo. Sometimes it is possible to consider the causal effect of a treatment as compared with non-treatment but not placebo treatment.
- Anything related to over-the-counter medications
- Many outcomes are sparsely recorded if at all. These include mortality, behavioral outcomes, lifestyle, and socioeconomic status.
- Since patients tend to encounter the healthcare system when they are unwell, measurement of the benefits of treatments can prove elusive.

Missingness in OHDSI databases presents subtle challenges. A health event (e.g., prescription, laboratory value, etc.) that should be recorded in a database, but isn't, is "missing." The statistics literature distinguishes between types of missingness such as "missing completely at random," "missing at random," and "missing not at random" and methods of increasing complexity attempt to address these types. Perkins et al. (2017) provide a use introduction to this topic.

What use cases are often observed? Drug safety, Drug utilization, etc.

- Practice

Chapter 9

OHDSI Analytics Tools

ATLAS: <http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:atlas>

ARACHNE: Network Research

Methods Library: <https://ohdsi.github.io/MethodsLibrary/>

Best practices enforced in all OHDSI methods.

Ethical consideration: e.g. should always communicate uncertainty. Prespecification of research questions, etc.

Analytic use cases

What is the difference between characterization, population-level estimation, patient-level prediction?

Case study: Perhaps on how to install the tools?

Chapter 10

SQL and R

DatabaseConnector and SqlRender

Querying the CDM

Probably borrow heavily from <https://github.com/OHDSI/QueryLibrary>

Chapter 11

Building the building blocks: cohorts

Introduction: a cohort is a group of people that meet a set of criteria for a particular span of time etc. Cohorts are used throughout OHDSIs analytical tools as the primary building blocks.

Using ATLAS: use material from Patrick's tutorial on cohort building

Using SQL: For advanced users, explain how cohorts can be created programmatically.

Probabilistic cohorts: Aphrodite?

Case study: some example cohort definitions

Chapter 12

Characterization

ATLAS' incidence rate calculator + cohort characterization tool

FeatureExtraction package: <https://github.com/OHDSI/FeatureExtraction>

Case study: characteristics + IRs of some cohorts

Example .. <http://www.pnas.org/content/113/27/7329>

Chapter 13

Population-level estimation

Chapter leads: Martijn Schuemie, David Madigan & Marc Suchard

Observational healthcare data, such as administrative claims and electronic health records, offer opportunities to generate real-world evidence about the effect of treatments that can meaningfully improve the lives of patients. In this chapter we focus on population-level effect estimation, that is, the estimation of average causal effects of medical interventions on specific health outcomes of interest. In what follows, we consider two different estimation tasks:

- **Direct effect estimation:** estimating the effect of an exposure on the risk of an outcome, as compared to no exposure.
- **Comparative effect estimation:** estimation the effect of one exposure (the target exposure) on the risk of an outcome, as compared to another exposure (the comparator 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 methods employ analytic devices to shed light on the counterfactual outcomes.

Use-cases for population-level effect estimation include treatment selection, safety surveillance, and comparative effectiveness. Methods can test specific hypotheses one-at-a-time (e.g. ‘signal evaluation’) or explore multiple-hypotheses-at-once (e.g. ‘signal detection’). In all cases, the objective remains the same: to produce a high-quality estimate of the causal effect.

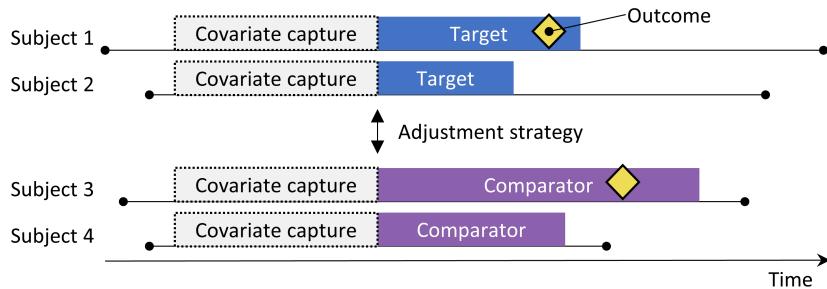


Figure 13.1: The new-user cohort design. Subjects observed to initiate the target treatment are compared to those initiating the comparator treatment. To adjust for differences between the two treatment groups several adjustment strategies can be used, such as stratification, matching, or weighting by the propensity score, or by adding baseline characteristics to the outcome model. The characteristics included in the propensity model or outcome model are captured prior to treatment initiation.

13.1 Study designs

Several different study designs can be used to estimate treatment effects. The main difference between these is how they construct the (unobserved) counterfactual. Below is a brief discussion of the most commonly used designs, all of which are implemented as R packages in the OHDSI Methods Library.

13.1.1 Cohort method

The new-user cohort method attempts to emulate a randomized clinical trial (Hernan and Robins, 2016). Subjects that are observed to initiate one treatment (the target) are compared to subjects initiating another treatment (the comparator) and are followed for a specific amount of time following treatment initiation, for example the time they stay on the treatment. We can specify the questions we wish to answer in a cohort study by making five choices:

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	What type of outcome model should be used

One example of an outcome model to choose is a logistic regression, which evaluates whether or not the outcome has occurred, and produces an odds ratio. A logistic regression assumes the time-at-risk is of the same length for both target and comparator, or irrelevant. Alternatively, we could choose a Poisson regression which estimates the incidence rate ratio, assuming a constant incidence rate. Often a Cox regression is used which considers time to first outcome to estimate the hazard ratio, assuming proportional hazards.

One crucial difference with a randomized trial is that there is no randomization, and therefore there might be systematic differences between the target

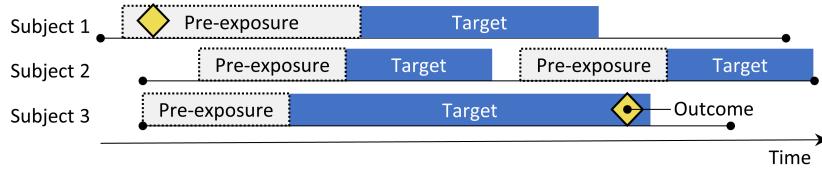


Figure 13.2: The self-controlled cohort design. The rate of outcomes during exposure to the target is compared to the rate of outcomes in the time pre-exposure.

CohortMethod package can support such practices, we prefer the use of large-scale regularized regression using many generic characteristics. (Tian et al., 2018) These characteristics include demographics, as well as all diagnoses, drug exposures, measurement, and medical procedures observed prior to treatment initiation, and exclude the target and comparator treatment. A model typically involves 10,000 to 100,000 unique characteristics. The PS can be used in several ways, for example by stratifying the study population based on the PS, by matching target subjects to comparator subjects with similar PS, or by weighting subjects using Inverse Probability of Treatment Weighting (IPTW) derived from the PS. Another strategy for adjusting for differences between the two groups is to include additional variables in the outcome model. One major limitation of this approach is that whereas there often is a wealth of data to fit a propensity model, with thousands of people in both treatment groups, the outcomes we study tend to be somewhat rare, causing a paucity of data when trying to fit elaborate models with the outcome as dependent variable. One approach is to use both a PS and add the same variables that were used in the propensity model in the outcome model, thus adjusting for the same variables twice, but in different ways. The new-user cohort method inherently is a method for comparative effect estimation, comparing one treatment to another. It is difficult to use this method to compare a treatment against no treatment, since it is hard to define a group of unexposed people that is comparable with the exposed group. If one wants to use this design for direct effect estimation, the preferred way is to select a comparator treatment for the same indication as the exposure of interest, where the comparator treatment is believed to have no effect on the outcome. Unfortunately, such a comparator might not always be available. In our gold standard, the comparators were specifically selected to have no effect, so we can also evaluate the cohort method's performance for direct effect estimation.

13.1.2 Self-controlled cohort

The self-controlled cohort (SCC) design (Ryan et al., 2013) compares the rate of outcomes during exposure to the rate of outcomes in the time just prior to the exposure. The following three choices define a self-controlled cohort question:

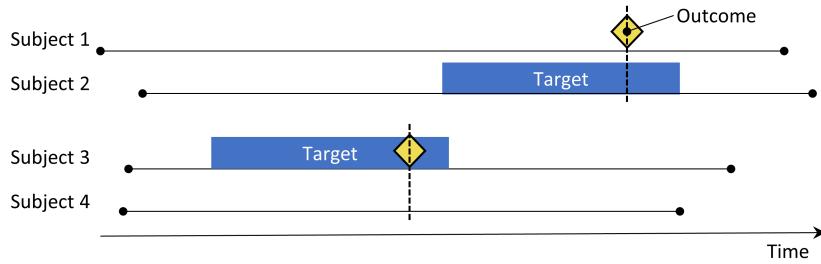


Figure 13.3: The case-control design. Subjects with the outcome ('cases') are compared to subjects without the outcome ('controls') in terms of their exposure status. Often, cases and controls are matched on various characteristics such as age and sex.

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?

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 between different time periods.

13.1.3 Case-control

Case-control (Vandenbroucke and Pearce, 2012) studies consider the question “are persons with a specific disease outcome exposed more frequently to a specific agent than those without the disease?” Thus, the central idea is to compare “cases”, i.e., subjects that experience the outcome of interest with “controls”, i.e., subjects that did not experience the outcome of interest. These choices define a case-control question:

Choice	Description
Outcome cohort	A cohort representing the cases (the outcome of interest)
Target cohort	A cohort representing the treatment
[Nesting cohort]	Optinally, 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?

Often, one matches controls to 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.

13.1.4 Case-crossover

The case-crossover (Maclure, 1991) design evaluates whether the rate of expo-

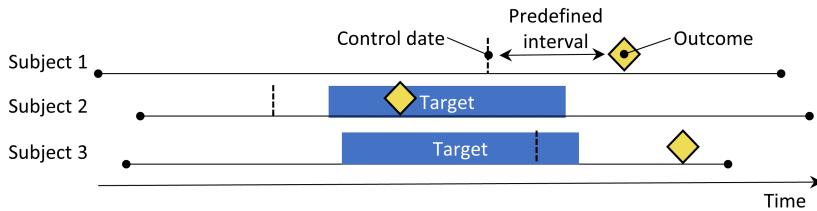


Figure 13.4: The case-crossover design. The time around the outcome is compared to a control date set at a predefined interval prior to the outcome date.

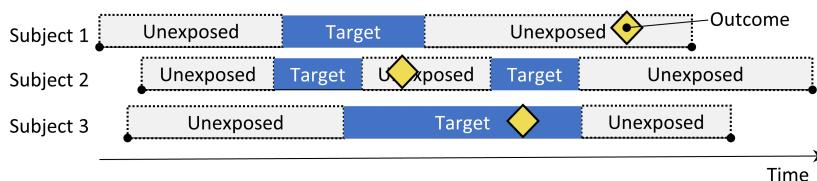


Figure 13.5: The Self-Controlled Case Series design. The rate of outcomes during exposure is compared to the rate of outcomes when not exposed.

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?

Since cases serve as their own control, it is a self-controlled design, and should therefore 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 (Süssel, 1995) was developed, which adds matched controls to the case-crossover design to adjust for exposure trends.

13.1.5 Self-controlled case series

The Self-Controlled Case Series (SCCS) design (Farrington, 1995, whitaker_2006) compares the rate of outcomes during exposure to the rate of outcomes during all unexposed time, both before, between, and after exposures. It is a Poisson regression that is conditioned on the person. Thus, it seeks to answer the question: “Given that a patient has the outcome, is the outcome more likely during exposed time compared to non-exposed time?”. These choices define an SCCS question:

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?

end is independent of the date of the outcome. Because for some outcomes, especially ones that can be fatal such as stroke, this assumption can be violated an extension to the SCCS has been developed that corrects for any such dependency. (Farrington et al., 2011)

13.2 Cohort study implementation

Describe case study: risk of angioedema and AMI in new users of ACE inhibitors compared to new users of thiazide and thiazide-like diuretics

13.2.1 Implementation using ATLAS

13.2.2 Implementation using R

13.3 Advanced topics

Best practices: http://www.ohdsi.org/web/wiki/doku.php?id=development:best_practices_estimation

Negative and positive controls, empirical calibration

13.4 Excercises

Chapter 14

Patient Level Prediction

Clinical decision making is a complicated task in which the clinician has to infer a diagnosis or treatment pathway based on the available medical history of the patient and the current clinical guidelines. Clinical prediction models have been developed to support this decision making process and are used in clinical practice in a wide spectrum of specialties. These models predict a diagnostic or prognostic outcome based on a combination of patient characteristics, e.g. demographic information, disease history, treatment history. The number of publications describing clinical prediction models has increased strongly over the last 10 years. An example is the Garvan model that predicts the 5-years and 10-years fractures risk in any elderly man or woman based on age, fracture history, fall history, bone mass density or weight (Nguyen et al., 2008). Many prediction models have been developed in patient subgroups at higher risk that need more intensive monitoring, e.g. the prediction of 30-day mortality after an acute myocardial described by Lee et al. (1995). Also, many models have been developed for asymptomatic subjects in the population, e.g. the famous Framingham risk functions for cardiovascular disease (Wilson et al., 1998), or the models for breast cancer screening (Engel and Fischer, 2015).

Surprisingly, most currently used models are estimated using small datasets and contain a limited set of patient characteristics. For example, in a review of 102 prognostic models in traumatic brain injury showed that three quarters of the models were based on samples with less than 500 patients (Perel et al., 2006). This low sample size, and thus low statistical power, forces the data analyst to make stronger modelling assumptions. The selection of the often limited set of patient characteristics is strongly guided by the expert knowledge at hand. This contrasts sharply with the reality of modern medicine wherein patients generate a rich digital trail, which is well beyond the power of any medical practitioner to fully assimilate. Presently, health care is generating huge amount of patient-specific information contained in the Electronic Health Record (EHR). This includes structured data in the form of diagnose, medication, laboratory test

results, and unstructured data contained in clinical narratives. Currently, it is unknown how much predictive accuracy can be gained by leveraging the large amount of data originating from the complete EHR of a patient.

Massive-scale, patient-specific predictive modeling has become reality due the OHDSI initiative in which the common data model (CDM) allows for uniform and transparent analysis at an unprecedented scale. These large standardized populations contain rich data to build highly predictive large-scale models and also provide immediate opportunity to serve large communities of patients who are in most need of improved quality of care. Such models can inform truly personalized medical care leading hopefully to sharply improved patient outcomes. Furthermore, these models could assist in the design and analysis of randomized controlled trials (RCT) by enabling a better patient stratification or can be utilized to adjust for confounding variables in observational research. More accurate prediction models contribute to targeting of treatment and to increasing cost-effectiveness of medical care.

Advances in machine learning for large dataset analysis have led to increased interest in applying patient-level prediction on this type of data. However, many published efforts in patient-level-prediction do not follow the model development guidelines, fail to perform extensive external validation, or provide insufficient model details that limits the ability of independent researchers to reproduce the models and perform external validation. This makes it hard to fairly evaluate the predictive performance of the models and reduces the likelihood of the model being used appropriately in clinical practice. To improve standards, several papers have been written detailing guidelines for best practices in developing and reporting prediction models.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (<https://www.equator-network.org/reporting-guidelines/tripod-statement/>) provides clear recommendations for reporting prediction model development and validation and addresses some of the concerns related to transparency. However, data structure heterogeneity and inconsistent terminologies still make collaboration and model sharing difficult as different researchers are often required to write new code to extract the data from their databases and may define variables differently.

In our paper (Reps et al., 2018), we propose a standardised framework for patient-level prediction that utilizes the OMOP Common Data Model (CDM) and standardized vocabularies, and describe the open-source software that we developed implementing the framework's pipeline. The framework is the first to support existing best practice guidelines and will enable open dissemination of models that can be extensively validated across the network of OHDSI collaborators.

Figure 14.1, illustrates the prediction problem we address. Among a population at risk, we aim to predict which patients at a defined moment in time ($t = 0$) will experience some outcome during a time-at-risk. Prediction is done using only

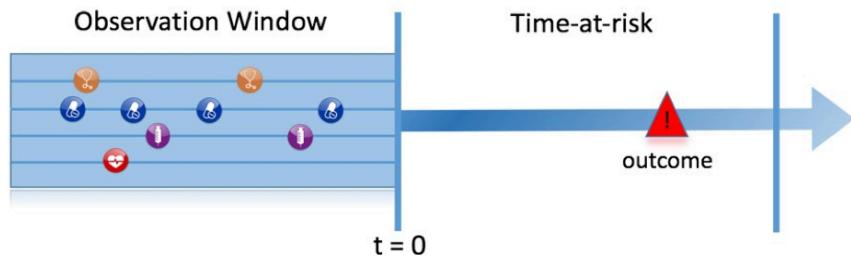


Figure 14.1: The prediction problem

Input parameter	Design choice
Target cohort (T)	We extract data for the patients in the Target Cohort (T) of which some will experience the outcome (O) in T
Outcome cohort (O)	
Time-at-risk	
Model development -which algorithm(s)? -which parameters? -which covariates?	

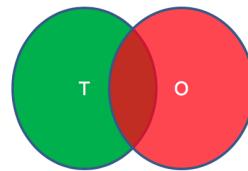


Figure 14.2: Design choices

information about the patients in an observation window prior to that moment in time.

As shown in Figure 14.2, to define a prediction problem we have to define $t=0$ by a Target Cohort (T), the outcome we like to predict by an outcome cohort (O), and the time-at-risk (TAR). Furthermore, we have to make design choices for the model we like to develop, and determine the observational datasets to perform internal and external validation. This conceptual framework works for all type of prediction problems, for example those presented in Figure 14.3.

In the next sections we will explain the best practices for model specification, implementation, and evaluation using OHDSI's Patient-Level Prediction (PLP) framework as guidance.

Type	Structure	Example
Disease onset and progression	Amongst patients who are newly diagnosed with <insert your favorite disease>, which patients will go on to have <another disease or related complication> within <time horizon from diagnosis>?	Among newly diagnosed AFib patients, which will go onto to have ischemic stroke in next 3 years?
Treatment choice	Amongst patients with <indicated disease> who are treated with either <treatment 1> or <treatment 2>, which patients were treated with <treatment 1> (on day 0)?	Among AFib patients who took either warfarin or rivaroxaban, which patients got warfarin? (as defined for propensity score model)
Treatment response	Amongst patients who are new users of <insert your favorite chronically-used drug>, which patients will <insert desired effect> in <time window>?	Which patients with T2DM who start on metformin stay on metformin after 3 years?
Treatment safety	Amongst patients who are new users of <insert your favorite drug>, which patients will experience <insert your favorite known adverse event from the drug profile> within <time horizon following exposure start>?	Among new users of warfarin, which patients will have GI bleed in 1 year?
Treatment adherence	Amongst patients who are new users of <insert your favorite chronically-used drug>, which patients will achieve <adherence metric threshold> at <time horizon>?	Which patients with T2DM who start on metformin achieve >=80% proportion of days covered at 1 year?

Figure 14.3: Examples of prediction problems

14.1 Study specification

The first step is to clearly define the prediction problem. Interestingly, in many published papers the prediction problem is poorly defined, e.g., it is unclear how the index date (start of the Target Cohort) is exactly defined. A poorly defined prediction problem does not allow for external validation by others let alone implementation in clinical practice. In the PLP framework we have enforced that we have to define the prediction problem we like to address, in which population we will build the model, which model we will build and how we will evaluate its performance. In this section we will guide you through this process and we will use a “Disease onset and progression” prediction type as an example.

14.1.1 Problem definition

Atrial fibrillation is a disease characterized by an irregular heart rate that can cause poor blood flow. Patients with atrial fibrillation are at increased risk of ischemic stroke. Anticoagulation is a recommended prophylaxis treatment strategy for patients at high risk of stroke, though the underuse of anticoagulants and persistent severity of ischemic stroke represents a substantial unmet medical need. Various strategies have been developed to predict risk of ischemic stroke in patients with atrial fibrillation. CHADS2 (Gage et al., 2001) was developed as a risk score based on history of congestive heart failure, hypertension, age ≥ 75 , diabetes and stroke. CHADS2 was initially derived using Medicare claims data, where it achieved good discrimination (AUC=0.82). However, subsequent external validation studies revealed the CHADS2 had substantially lower predictive

accuracy (Keogh et al., 2011). Subsequent stroke risk calculators have been developed and evaluated, including the extension of CHADS2Vasc. The management of atrial fibrillation has evolved substantially over the last decade, for various reasons that include the introduction of novel oral anticoagulants. With these innovations has come a renewed interest in greater precision medicine for stroke prevention.

We will apply the PLP framework to observational healthcare data to address the following patient-level prediction question:

Amongst patients who are newly diagnosed with Atrial Fibrillation, which patients will go on to have Ischemic Stroke within 1 year?

We will define ‘patients who are newly diagnosed with Atrial Fibrillation’ as the first condition record of cardiac arrhythmia, which is followed by another cardiac arrhythmia condition record, at least two drug records for a drug used to treat arrhythmias, or a procedure to treat arrhythmias. We will define ‘Ischemic stroke events’ as ischemic stroke condition records during an inpatient or ER visit; successive records with > 180 day gap are considered independent episodes.

14.1.2 Study population definition

The final study population in which we will develop our model is often a subset of the Target population, because we will e.g. apply criteria that are dependent on T and O or we want to do sensitivity analyses with subpopulations of T. For this we have to answer the following questions:

- *What is the minimum amount of observation time we require before the start of the target cohort?* This choice could depend on the available patient time in your training data, but also on the time you expect to be available in the data sources you want to apply the model on in the future. The longer the minimum observation time, the more baseline history time is available for each person to use for feature extraction, but the fewer patients will qualify for analysis. Moreover, there could be clinical reasons to choose a short or longer lookback period. For our example, we will use a prior history as lookback period (washout period).
- *Can patients enter the target cohort multiple times?* In the target cohort definition, a person may qualify for the cohort multiple times during different spans of time, for example if they had different episodes of a disease or separate periods of exposure to a medical product. The cohort definition does not necessarily apply a restriction to only let the patients enter once, but in the context of a particular patient-level prediction problem, a user may want to restrict the cohort to the first qualifying episode. In our example, a person could only enter the target cohort once since our criteria was based on first occurrence of atrial fibrillation.

- *Do we allow persons to enter the cohort if they experienced the outcome before?* Do we allow persons to enter the target cohort if they experienced the outcome before qualifying for the target cohort? Depending on the particular patient-level prediction problem, there may be a desire to predict ‘incident’ first occurrence of an outcome, in which case patients who have previously experienced the outcome are not ‘at-risk’ for having a first occurrence and therefore should be excluded from the target cohort. In other circumstances, there may be a desire to predict ‘prevalent’ episodes, whereby patients with prior outcomes can be included in the analysis and the prior outcome itself can be a predictor of future outcomes. For our prediction example, the answer to this question is ‘Yes, allow persons with prior outcomes’ because we know from the CHADS2 score that prior strokes are very predictive of future strokes. If this answer would have been ‘No’ we also have to decide how long we would look back for previous occurrences of the outcome.
- *How do we define the period in which we will predict our outcome relative to the target cohort start?* We actually have to make two decisions to answer that question. First, does the time-at-risk window start at the date of the start of the target cohort or later? Arguments to make it start later could be that you want to avoid outcomes that were entered late in the record that actually occurred before the start of the target cohort or you want to leave a gap where interventions to prevent the outcome could theoretically be implemented. Second, you need to define the time-at-risk by setting the risk window end, as some specification of days offset relative to the target cohort start or end dates. For our problem we will predict in a ‘time-at-risk’ window starting 1 day after the start of the target cohort up to 365 days later (to look for 1-year risk following atrial fibrillation diagnosis).
- *Do we require a minimum amount of time-at-risk?* We have to decide if we want to include patients that did not experience the outcome but did leave the database earlier than the end of our time-at-risk period. These patients may experience the outcome when we do not observe them. For our prediction problem we decide to answer this question with ‘Yes, require a minimum time-at-risk’ for that reason. Furthermore, we have to decide if this constraint also applies to persons who experienced the outcome or we will include all persons with the outcome irrespective of their total time at risk. For example, if the outcome is death, then persons with the outcome are likely censored before the full time-at-risk period is complete.

14.1.3 Model development settings

To develop the model we have to decide which algorithm(s) we like to train. We see the selection of the best algorithm for a certain prediction problem as an empirical question, i.e. you need to let the data speak for itself and try different

approaches to find the best one. There is no algorithm that will work best for all problems (no free lunch). In our framework we therefore aim to implement many algorithms. Furthermore, we made the system modular so you can add your own custom algorithms. This out-of-scope for this chapter but mode details can be found in the AddingCustomAlgorithms vignette (<https://github.com/OHDSI/PatientLevelPrediction/blob/master/inst/doc/AddingCustomAlgorithms.pdf>).

Our framework currently contains the following algorithms to choose from:

Algorithm	Description	Hyper-parameters
Regularized Logistic Regression	<p>Regularized Lasso logistic regression belongs to the Logistic family of generalized linear models, where a linear combination of the variables is learned and finally a logistic function maps the linear combination to a value between 0 and 1. The lasso regularization adds a cost based on model complexity to the objective function when training the model. This cost is the sum of the absolute values of the linear combination of the coefficients. The model automatically performs feature selection by minimizing this cost. We use the Cyclic coordinate descent for logistic, Poisson and survival analysis (Cyclops) package to perform large-scale regularized logistic regression:</p> <p>https://github.com/OHDSI/Cyclops</p>	var (starting variance), seed
Gradient boosting machines	<p>Gradient boosting machines is a boosting ensemble technique and in our framework it combines multiple decision trees. Boosting works by iteratively adding decision trees but adds more weight to the data-points that are misclassified by prior decision trees in the cost function when training the next tree. We use Extreme Gradient Boosting, which is an efficient implementation of the gradient boosting framework implemented in the xgboost R package available from CRAN.</p>	ntree (number of trees), max depth (max levels in tree), min rows (minimum data points in node), learning rate, seed

Algorithm	Description	Hyper-parameters
Random forest	<p>Random forest is a bagging ensemble technique that combines multiple decision trees. The idea behind bagging is to reduce the likelihood of overfitting, by using weak classifiers, but combining multiple diverse weak classifiers into a strong classifier.</p> <p>Random forest accomplishes this by training multiple decision trees but only using a subset of the variables in each tree and the subset of variables differ between trees. Our packages uses the sklearn learn implementation of Random Forest in python.</p>	mtry (number of features in each tree),ntree (number of trees), maxDepth (max levels in tree), minRows (minimum data points in in node),balance (balance class labels), seed
K-nearest neighbors	<p>K-nearest neighbors (KNN) is an algorithm that uses some metric to find the K closest labelled data-points, given the specified metric, to a new unlabelled data-point. The prediction of the new data-points is then the most prevalent class of the K-nearest labelled data-points.</p> <p>There is a sharing limitation of KNN, as the model requires labelled data to perform the prediction on new data, and it is often not possible to share this data across data sites. We included the BigKnn classifier developed in OHDSI which is a large scale k-nearest neighbor classifier using the Lucene search engine:</p> <p>https://github.com/OHDSI/BigKnn</p>	k (number of neighbours),weighted (weight by inverse frequency)
Naive Bayes	<p>The Naive Bayes algorithm applies the Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Based on the likelihood the data belongs to a class and the prior distribution of the class, a posterior distribution is obtained.</p>	none

Algorithm	Description	Hyper-parameters
AdaBoost	AdaBoost is a boosting ensemble technique. Boosting works by iteratively adding classifiers but adds more weight to the data-points that are misclassified by prior classifiers in the cost function when training the next classifier. We use the sklearn “AdaboostClassifier” implementation in Python.	nEstimators (the maximum number of estimators at which boosting is terminated), learningRate (learning rate shrinks the contribution of each classifier by learning_rate). There is a trade-off between learningRate and nEstimators)
Decision Tree	A decision tree is a classifier that partitions the variable space using individual tests selected using a greedy approach. It aims to find partitions that have the highest information gain to separate the classes. The decision tree can easily overfit by enabling a large number of partitions (tree depth) and often needs some regularization (e.g., pruning or specifying hyper-parameters that limit the complexity of the model). We use the sklearn “DecisionTreeClassifier” implementation in Python.	maxDepth (the maximum depth of the tree), minSamplesSplit,minSamplesLeaf, minImpuritySplit (threshold for early stopping in tree growth. A node will split if its impurity is above the threshold, otherwise it is a leaf.), seed,classWeight (“Balance”” or “None”)

Algorithm	Description	Hyper-parameters
Multilayer Perception	Neural networks contain multiple layers that weight their inputs using a non-linear function. The first layer is the input layer, the last layer is the output layer the between are the hidden layers. Neural networks are generally trained using feed forward back-propagation. This is when you go through the network with a data-point and calculate the error between the true label and predicted label, then go backwards through the network and update the linear function weights based on the error. This can also be performed as a batch, where multiple data-points are fed through the network at once.	size (the number of hidden nodes), alpha (the l2 regularisation), seed
Deep Learning	Deep learning such as deep nets, convolutional neural networks or recurrent neural networks are similar to a neural network but have multiple hidden layers that aim to learn latent representations useful for prediction. In the <code>BuildingDeepLearningModels</code> vignette we describe these models and their hyper-parameters in more detail	

Furthermore, we have to decide on the **covariates** that we will use to train our model. This choice can be driven by domain knowledge of available computational resources. In our example, we like to add the Gender, Age, Conditions, Drugs Groups, and Visit Count. We also have to specify in which time windows we will look and we decide to look in year before and any time prior.

14.1.4 Model evaluation

Finally, we have to define how we will train and test our model on our data, i.e. how we perform **internal validation**. For this we have to decide how we divide our dataset in a training and testing dataset and how we randomly assign patients to these two sets. Dependent on the size of the training set we can decide how much data we like to use for training, typically this is a 75%, 25% split. If you have very large datasets you can use more data for training. To randomly assign patients to the training and testing set, there are two commonly used approaches:

1. split by person. In this case a random seed is used to assign the patient to either sets.

2. split by time. In this case a time point is used to split the persons, e.g. 75% of the data is before and 25% is after this date. The advantage of this is that you take into consideration that the health care system has changed over time.

For our prediction model we decide to start with a Regularized Logistic Regression and will use the default parameters. We will do a 75%-25% split by person.

14.1.5 Study summary

We now completely defined our study:

Definition	Value
Problem Definition	
Target Cohort (T)	‘Patients who are newly diagnosed with Atrial Fibrillation’ defined as the first condition record of cardiac arrhythmia, which is followed by another cardiac arrhythmia condition record, at least two drug records for a drug used to treat arrhythmias, or a procedure to treat arrhythmias.
Outcome Cohort (O)	‘Ischemic stroke events’ defined as ischemic stroke condition records during an inpatient or ER visit; successive records with > 180 day gap are considered independent episodes.
Time-at-risk (TAR)	1 day till 365 days from cohort start
Population Definition	
Washout Period	1095
Enter the target cohort multiple times?	No
Allow prior outcomes?	Yes
Start of time-at-risk	1 day
End of time-at-risk	365 days
Require a minimum amount of time-at-risk?	Yes (364 days)
Model Development	
Algorithm	Regularized Logistic Regression
Hyper-parameters	variance = 0.01 (Default)

Definition	Value
Covariates	ender, Age, Conditions (ever before, <365), Drugs Groups (ever before, <365), and Visit Count
Data split	75% train, 25% test. Randomly assigned by person

According to the best practices we need to make a protocol that completely specifies how we plan to execute our study. This protocol will be assessed by the governance boards of the participating data sources in your network study. For this a template could be used but we prefer to automate this process as much as possible by adding functionality to automatically generate study protocol from a study specification. We will discuss this in more detail later.

14.2 Study implementation

Now we have completely design our study we have to implement the study. This will be done using the `PatientLevelPrediction` package to build patient-level predictive models. The package enables data extraction, model building, and model evaluation using data from databases that are translated into the OMOP CDM. To install the package we like to point you to the `InstallationGuide` that can be found on the GitHub website (<https://github.com/OHDSI/PatientLevelPrediction/>).

We first have to generate the target and outcome cohorts and we then need to develop the R code to run against our CDM to execute the full study. These steps will be described in the paragraphs below.

14.2.1 Cohort instantiation

For our study we need to know when a person enters the target and outcome cohorts. This is stored in a table on the server that contains the cohort start date and cohort end date for all subjects for a specific cohort definition. This cohort table has a very simple structure as shown below:

- `cohort_definition_id`, a unique identifier for distinguishing between different types of cohorts, e.g. cohorts of interest and outcome cohorts.
- `subject_id`, a unique identifier corresponding to the `person_id` in the CDM.
- `cohort_start_date`, the date the subject enters the cohort.
- `cohort_end_date`, the date the subject leaves the cohort.

The screenshot shows the ATLAS software interface for creating a cohort. The left sidebar contains navigation links like Home, Data Sources, Search, Concept Sets, Cohort Definitions, Characterizations, Cohort Pathways, Incidence Rates, Profiles, Estimation, Prediction, Jobs, Configuration, and Feedback. The main area has tabs for Definition, Concept Sets, Generation, Reporting, Export, and Warnings. The 'Definition' tab is active, showing a 'PatientLevelPrediction vignette: T: patients who are newly diagnosed with Atrial fibrillation'. Below this, there's a 'Initial Event Cohort' section with a condition 'Atrial fibrillation' and a note about it being the first time in the person's history. There are buttons for 'Add initial event', 'Add criteria attribute...', and 'Delete Criteria'. The 'Additional Qualifying Inclusion Criteria' section shows a single criterion: 'Have no prior warfarin exposure'. This is followed by a section for 'Additional Qualifying Exclusion Criteria' which is currently empty.

Figure 14.4: Target Cohort Atrial Fibrillation

How do we fill this table according to our cohort definitions? There are two options for this:

- 1) use the interactive cohort builder tool in ATLAS (www.github.com/OHDSI/ATLAS) which can be used to create cohorts based on inclusion criteria and will automatically populate this cohort table.
- 2) write your own custom SQL statements to fill the cohort table.

Both methods are described below for our example prediction problem.

14.2.1.1 ATLAS cohort builder

ATLAS allows you to define cohorts interactively by specifying cohort entry and cohort exit criteria. Cohort entry criteria involve selecting one or more initial events, which determine the start date for cohort entry, and optionally specifying additional inclusion criteria which filter to the qualifying events. Cohort exit criteria are applied to each cohort entry record to determine the end date when the person's episode no longer qualifies for the cohort. For the outcome cohort the end date is less relevant. As an example, Figure 14.4 shows how we created the Atrial Fibrillation cohort and Figure 14.5 shows how we created the stroke cohort in ATLAS.

The T and O cohorts can be found here:

- Atrial Fibrillaton (T): <http://www.ohdsi.org/web/atlas/#/cohortdefinition/1769447>
- Stroke (O) : <http://www.ohdsi.org/web/atlas/#/cohortdefinition/1769448>

In depth explanation of cohort creation in ATLAS is out of scope of this vignette but can be found on the OHDSI wiki pages (<http://www.ohdsi.org/web/wiki/>)

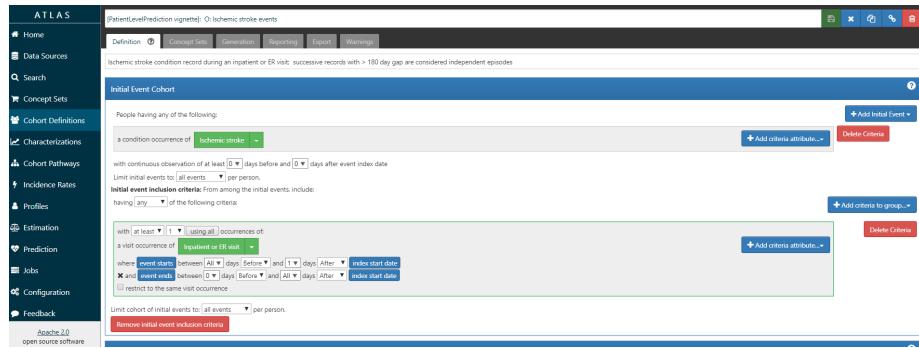


Figure 14.5: Outcome Cohort Stroke

doku.php?id=documentation:software:atlas).

Note that when a cohort is created in ATLAS the cohortid is needed to extract the data in R. The cohortid can be found at the top of the ATLAS screen.

14.2.1.2 Custom cohorts

It is also possible to create cohorts without the use of ATLAS. Using custom cohort code (SQL) you can make more advanced cohorts if needed.

For our example study, we need to create a table to hold the cohort data and we need to create SQL code to instantiate this table for both the AF and Stroke cohorts. Therefore, we create a file called *AfStrokeCohorts.sql* with the following contents:

```
*****
File AfStrokeCohorts.sql
*****
/*
   Create a table to store the persons in the T and C cohort
*/

IF OBJECT_ID(' @resultsDatabaseSchema.PLPAFibStrokeCohort', 'U') IS NOT NULL
    DROP TABLE @resultsDatabaseSchema.PLPAFibStrokeCohort;

CREATE TABLE @resultsDatabaseSchema.PLPAFibStrokeCohort
(
    cohort_definition_id INT,
    subject_id BIGINT,
    cohort_start_date DATE,
    cohort_end_date DATE
);
```

```

/*
  T cohort: [PatientLevelPrediction vignette]: T : patients who are newly
  diagnosed with Atrial fibrillation
  - persons with a condition occurrence record of 'Atrial fibrillation' or
    any descendants, indexed at the first diagnosis
  - who have >1095 days of prior observation before their first diagnosis
  - and have no warfarin exposure any time prior to first AFib diagnosis
*/
INSERT INTO @resultsDatabaseSchema.AFibStrokeCohort (cohort_definition_id,
                                                       subject_id,
                                                       cohort_start_date,
                                                       cohort_end_date)
SELECT 1 AS cohort_definition_id,
       AFib.person_id AS subject_id,
       AFib.condition_start_date AS cohort_start_date,
       observation_period.observation_period_end_date AS cohort_end_date
FROM
(
  SELECT person_id, min(condition_start_date) as condition_start_date
  FROM @cdmDatabaseSchema.condition_occurrence
  WHERE condition_concept_id IN (SELECT descendant_concept_id FROM
                                    @cdmDatabaseSchema.concept_ancestor WHERE ancestor_concept_id IN
                                    (313217 /*atrial fibrillation*/))
  GROUP BY person_id
) AFib
INNER JOIN @cdmDatabaseSchema.observation_period
  ON AFib.person_id = observation_period.person_id
  AND AFib.condition_start_date >= dateadd(dd,1095,
                                             observation_period.observation_period_start_date)
  AND AFib.condition_start_date <= observation_period.observation_period_end_date
LEFT JOIN
(
  SELECT person_id, min(drug_exposure_start_date) as drug_exposure_start_date
  FROM @cdmDatabaseSchema.drug_exposure
  WHERE drug_concept_id IN (SELECT descendant_concept_id FROM
                            @cdmDatabaseSchema.concept_ancestor WHERE ancestor_concept_id IN
                            (1310149 /*warfarin*/))
  GROUP BY person_id
) warfarin
  ON Afib.person_id = warfarin.person_id
  AND Afib.condition_start_date > warfarin.drug_exposure_start_date
WHERE warfarin.person_id IS NULL
;

```

```

/*
C cohort: [PatientLevelPrediction vignette]: 0: Ischemic stroke events
- inpatient visits that include a condition occurrence record for
  'cerebral infarction' and descendants, 'cerebral thrombosis',
  'cerebral embolism', 'cerebral artery occlusion'
*/
INSERT INTO @resultsDatabaseSchema.AFibStrokeCohort (cohort_definition_id,
                                                       subject_id,
                                                       cohort_start_date,
                                                       cohort_end_date)
SELECT 2 AS cohort_definition_id,
       visit_occurrence.person_id AS subject_id,
       visit_occurrence.visit_start_date AS cohort_start_date,
       visit_occurrence.visit_end_date AS cohort_end_date
FROM
(
  SELECT person_id, condition_start_date
  FROM @cdmDatabaseSchema.condition_occurrence
  WHERE condition_concept_id IN (SELECT DISTINCT descendant_concept_id FROM
@cdmDatabaseSchema.concept_ancestor WHERE ancestor_concept_id IN
(443454 /*cerebral infarction*/) OR descendant_concept_id IN
(441874 /*cerebral thrombosis*/, 375557 /*cerebral embolism*/,
 372924 /*cerebral artery occlusion*/))
) stroke
INNER JOIN @cdmDatabaseSchema.visit_occurrence
ON stroke.person_id = visit_occurrence.person_id
AND stroke.condition_start_date >= visit_occurrence.visit_start_date
AND stroke.condition_start_date <= visit_occurrence.visit_end_date
AND visit_occurrence.visit_concept_id IN (9201, 262 /*'Inpatient Visit' or
  'Emergency Room and Inpatient Visit'*/)
GROUP BY visit_occurrence.person_id, visit_occurrence.visit_start_date,
         visit_occurrence.visit_end_date
;

```

This is parameterized SQL which can be used by the `SqlRender` package. We use parameterized SQL so we do not have to pre-specify the names of the CDM and result schemas. That way, if we want to run the SQL on a different schema, we only need to change the parameter values; we do not have to change the SQL code. By also making use of translation functionality in `SqlRender`, we can make sure the SQL code can be run in many different environments.

To execute this SQL against our CDM we first need to tell R how to connect to the server. `PatientLevelPrediction` 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:

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

cdmDatabaseSchema <- "my_cdm_data"
cohortsDatabaseSchema <- "my_results"
cdmVersion <- "5"
```

The last three lines define the `cdmDatabaseSchema` and `cohortsDatabaseSchema` variables, as well as the CDM version. We will use these later to tell R where the data in CDM format live, where we want to create the cohorts of interest, and what version CDM is used. Note that for Microsoft SQL Server, databaseschemas need to specify both the database and the schema, so for example `cdmDatabaseSchema <- "my_cdm_data.dbo"`.

```
library(SqlRender)
sql <- readSql("AfStrokeCohorts.sql")
sql <- renderSql(sql,
                 cdmDatabaseSchema = cdmDatabaseSchema,
                 cohortsDatabaseSchema = cohortsDatabaseSchema,
                 post_time = 30,
                 pre_time = 365)$sql
sql <- translateSql(sql, targetDialect = connectionDetails$dbms)$sql

connection <- connect(connectionDetails)
executeSql(connection, sql)
```

In this code, we first read the SQL from the file into memory. In the next line, we replace four parameter names with the actual values. We then translate the SQL into the dialect appropriate for the DBMS we already specified in the `connectionDetails`. Next, we connect to the server, and submit the rendered and translated SQL.

If all went well, we now have a table with the events of interest. We can see how many events per type:

```
sql <- paste("SELECT cohort_definition_id, COUNT(*) AS count",
             "FROM @cohortsDatabaseSchema.AFibStrokeCohort",
             "GROUP BY cohort_definition_id")
sql <- renderSql(sql, cohortsDatabaseSchema = cohortsDatabaseSchema)$sql
sql <- translateSql(sql, targetDialect = connectionDetails$dbms)$sql

querySql(connection, sql)

## cohort_definition_id  count
```

```
## 1 1 527616  
## 2 2 221555
```

14.2.2 Study script creation

In this section we assume that our cohorts have been created either by using ATLAS or a custom SQL script. We will first explain how to create an R script yourself that will execute our study as we have defined earlier.

14.2.2.1 Data extraction

Now we can tell `PatientLevelPrediction` to extract all necessary data for our analysis. This is done using the `FeatureExtractionPackage` package (<https://github.com/OHDSI/FeatureExtraction>). In short the `FeatureExtractionPackage` allows you to specify which features (covariates) need to be extracted, e.g. all conditions and drug exposures, more information can be found in chapter X. It also supports the creation of custom covariates. For more detailed information on the `FeatureExtraction` package see its vignettes. For our example study we decided to use these settings:

The final step for extracting the data is to run the `getPlpData` function and input the connection details, the database schema where the cohorts are stored, the cohort definition ids for the cohort and outcome, and the `washoutPeriod` which is the minimum number of days prior to cohort index date that the person must have been observed to be included into the data, and finally input the previously constructed covariate settings.

```
    outcomeIds = 2,  
    sampleSize = 10000  
)
```

Note that if the cohorts are created in ATLAS its corresponding cohort database schema needs to be selected. There are many additional parameters for the `getPlpData` function which are all documented in the `PatientLevelPrediction` manual. The resulting `plpData` 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.

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

```
savePlpData(plpData, "stroke_in_af_data")
```

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

14.2.2.2 Additional inclusion criteria

To completely define the prediction problem the final study population is obtained by applying additional constraints on the two earlier defined cohorts, e.g., a minimum time at risk can be enforced (`requireTimeAtRisk`, `minTimeAtRisk`) and we can specify if this also applies to patients with the outcome (`includeAllOutcomes`). Here we also specify the start and end of the risk window relative to target cohort start. For example, if we like the risk window to start 30 days after the at-risk cohort start and end a year later we can set `riskWindowStart = 30` and `riskWindowEnd = 365`. In some cases the risk window needs to start at the cohort end date. This can be achieved by setting `addExposureToStart = TRUE` which adds the cohort (exposure) time to the start date.

In Appendix 1, we demonstrate the effect of these settings on the subset of the persons in the target cohort that end up in the final study population.

In the example below all the settings we defined for our study are imposed:

```

    addExposureDaysToStart = FALSE,
    addExposureDaysToEnd = FALSE,
    minTimeAtRisk = 364,
    requireTimeAtRisk = TRUE,
    includeAllOutcomes = TRUE,
    verbosity = "DEBUG"
)

```

14.2.2.3 Model Development

In the set function of an algorithm the user can specify a list of eligible values for each hyper-parameter. All possible combinations of the hyper-parameters are included in a so-called grid search using cross-validation on the training set. If a user does not specify any value then the default value is used instead.

For example, if we use the following settings for the gradientBoostingMachine: ntrees=c(100,200), maxDepth=4 the grid search will apply the gradient boosting machine algorithm with ntrees=100 and maxDepth=4 plus the default settings for other hyper-parameters and ntrees=200 and maxDepth=4 plus the default settings for other hyper-parameters. The hyper-parameters that lead to the bestcross-validation performance will then be chosen for the final model. For our problem we choose to build a logistic regression model with the default hyper-parameters

```
lrModel <- setLassoLogisticRegression()
```

The `runPlp` function uses the population, `plpData`, and model settings to train and evaluate the model. We can use the `testSplit` (person/time) and `testFraction` parameters to split the data in a 75%-25% split and run the patient-level prediction pipeline:

```
lrResults <- runPlp(population, plpData, modelSettings = lrModel, testSplit='person',
                      testFraction=0.25, nfold=2, splitSeed = 1234)
```

Under the hood the package will now use the `Cyclops` package fit a large-scale regularized regression using 75% of the data and will evaluate the model on the remaining 25%. A results data structure is returned containing information about the model, its performance etc.

In the `runPlp` function there are several parameters to save the `plpData`, `plpResults`, `plpPlots`, `evaluation`, etc. objects which are all set to `true` by default. However, there is also some functionality to this manually.

You can save the model using:

```
savePlpModel(lrResults$model, dirPath = file.path(getwd(), "model"))
```

You can load the model using:

```
plpModel <- loadPlpModel(getwd(), 'model')
```

You can also save the full results structure using:

```
savePlpResult(lrResults, location = file.path(getwd(), 'lr'))
```

To load the full results structure use:

```
lrResults <- loadPlpResult(file.path(getwd(), 'lr'))
```

14.2.3 Study package creation

The script we created manually above can also be automatically created using a powerful feature in ATLAS. By creating a new prediction study (left menu) you can select the Target and Outcome as created in ATLAS, set all the study parameters, and then you can download a R package that you can use to execute your study. What is really powerful is that you can add multiple Ts, Os, covariate settings etc. The package will then run all the combinations of automatically as separate analyses. The screenshots below explain this process.

1. Create a new prediction study and select your target and outcome cohorts.

This screenshot shows the 'Prediction Problem Settings' page in ATLAS. The left sidebar includes options like Home, Data Sources, Search, Concept Sets, Cohort Definitions, Cohort Pathways, Incidence Rates, Profiles, Estimation, Prediction, Jobs, Configuration, and Feedback. The main area has tabs for 'Specification' and 'Utilities'. Under 'Specification', there's a search bar and a table for '1-year risk of ischemic stroke among patients newly diagnosis with atrial fibrillation, using various population settings'. The 'Prediction Problem Settings' tab is active, showing 'Target Cohorts' (with one entry: 'T: patients who are newly diagnosed with Atrial fibrillation') and 'Outcome Cohorts' (with one entry: 'O: ischemic stroke events'). Below these are 'Analysis Settings' with a 'Model Settings' section containing a single entry for 'LassoLogisticRegressionSettings'.

2. Specify one or more analysis settings

This screenshot shows the 'Analysis Settings' page in ATLAS. The left sidebar is identical to the previous screenshot. The main area has tabs for 'Specification' and 'Utilities'. Under 'Specification', there's a search bar and a table for '1-year risk of ischemic stroke among patients newly diagnosis with atrial fibrillation, using various population settings'. The 'Analysis Settings' tab is active, showing 'Model Settings' (with one entry: 'LassoLogisticRegressionSettings'), 'Covariate Settings' (with one entry: 'DemographicGender, DemographicAgeGroup, ConditionGroupDiabetesType, ConditionGroupDiabetesLongTerm, DrugGroupDiabetesType, DrugGroupDiabetesLongTerm (>10 more covariate settings)'), and 'Population Settings' (with four entries for 'Risk Window Start' values: 365, 385, 365, 385, corresponding to 'Risk Window End' values of 1095, 1095, 1095, 1095, and 'Include All Outcomes' values of true, true, true, true, and 'Remove Subjects With Prior Outcome' values of false, false, false, true). Below these sections is a table showing 'Showing 1 to 4 of 4 entries'.

3. Specify the trainings settigns.

4. Specify the execution settings.

ATLAS can build a R package for you that will execute the full study against your CDM. Below the steps are explained how to do this in ATLAS.

- Under utilities you can find download. Click on the button to review the full study specification.

The screenshot shows the ATLAS web application interface. On the left is a dark sidebar with a navigation menu containing items like Home, Data Sources, Search, Concept Sets, Cohort Definitions, Cohort Pathways, Incidence Rates, Profiles, Estimation, Prediction, Jobs, Configuration, and Feedback. The 'Prediction' item is currently selected and highlighted in blue. At the top center, there's a title bar for a specific study: 'COPY CP (PatientLevelPrediction vignette) Risk of ischemic stroke in patients with Atrial fibrillation'. Below the title bar, there are two buttons: 'Download' and 'Execute'. A tooltip above the 'Download' button says: 'Please click the button below to view the full study specifications. Once reviewed, scroll down to download the study package.' To the right of the 'Download' button is a link: 'Click here to review the full study specification.' At the bottom of the sidebar, there's a footer note: 'discte 2.0 open source software provided by OHDSI join the source'.

- You now have to review that you indeed want to run all these analyses (cartesian product of all the settings for each T and O combination).

This screenshot shows the 'Step 1. Review Full Study Specification' section of the ATLAS interface. It includes a table with columns for Target Cohort Name, Outcome Cohort Name, Model Name, Model Settings, Create Settings, and Risk Window. The table lists entries such as 'Target Cohort' (patients who are newly diagnosed with Atrial fibrillation), 'Outcome Cohort' (patients who are newly diagnosed with Ischemic stroke events), and 'Model' (LogisticRegressionSettings). Below this, there's a section titled 'Step 2. Download the study package' with a button labeled 'Download Study Package'.

- If you agree, you give the package a name, and download the package as a zipfile.
- By opening the R package in R studio and building the package you can run the study using the `execute` function. There's also an example `CodeToRun.R` script available in the extras folder of the package with extra instructions.

14.3 Internal validation

Once we execute the study, the `runPlp()` function returns the trained model and the evaluation of the model on the train/test sets.

You can interactively view the results by running: `viewPlp(runPlp=lrResults)`. This will generate a Shiny App in your browser in which you can view all performance measures created by the framework as shown in the figure below.

Metric	test	train
1 AUC	0.72130	0.75348
2 AUC_lb95ci	0.70057	0.74215
3 AUC_ub95ci	0.74203	0.76482
4 AUPRC	0.10971	0.13571
5 BrierScaled	0.03755	0.04902
6 BrierScore	0.03355	0.03304
7 CalibrationIntercept.Intercept	-0.00089	-0.00813
8 CalibrationSlope.Gradient	1.02041	1.22457
9 outcomeCount	601.00000	1802.00000
10 populationSize	16685.00000	50054.00000
11 Incidence	3.60204	3.60011

Furthermore, many interactive plots are available in the Shiny App, for example the ROC curve in which you can move over the plot to see the threshold and the corresponding sensitivity and specificity values.



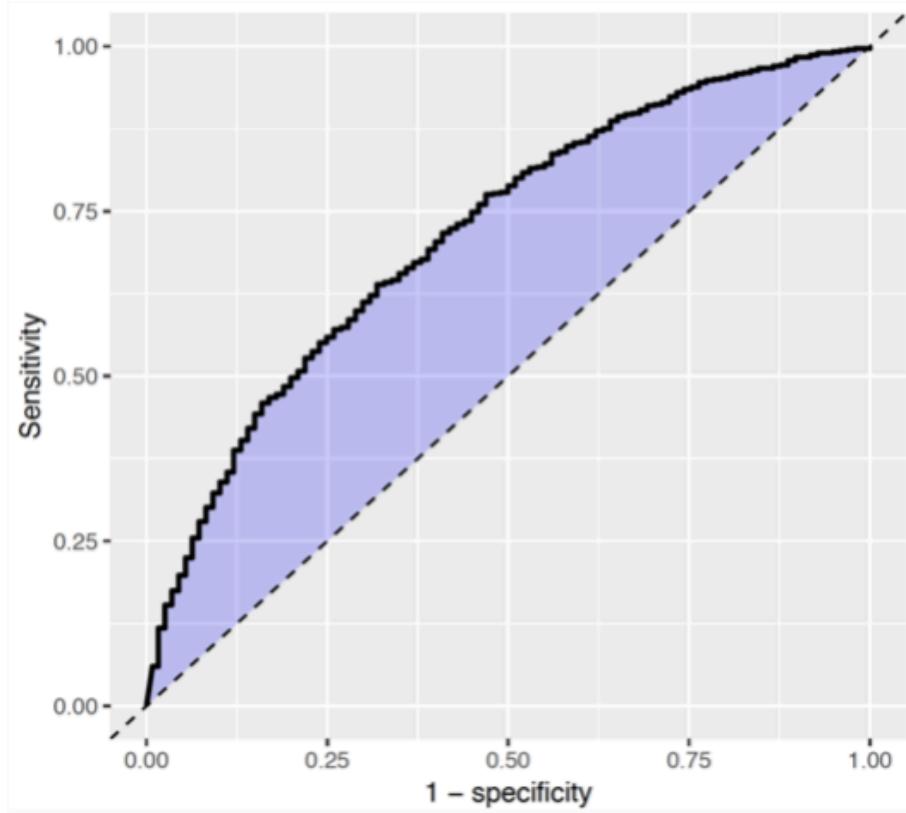
To generate and save all the evaluation plots to a folder run the following code:

```
plotPlp(lrResults, dirPath=getwd())
```

The plots are described in more detail in the next sections.

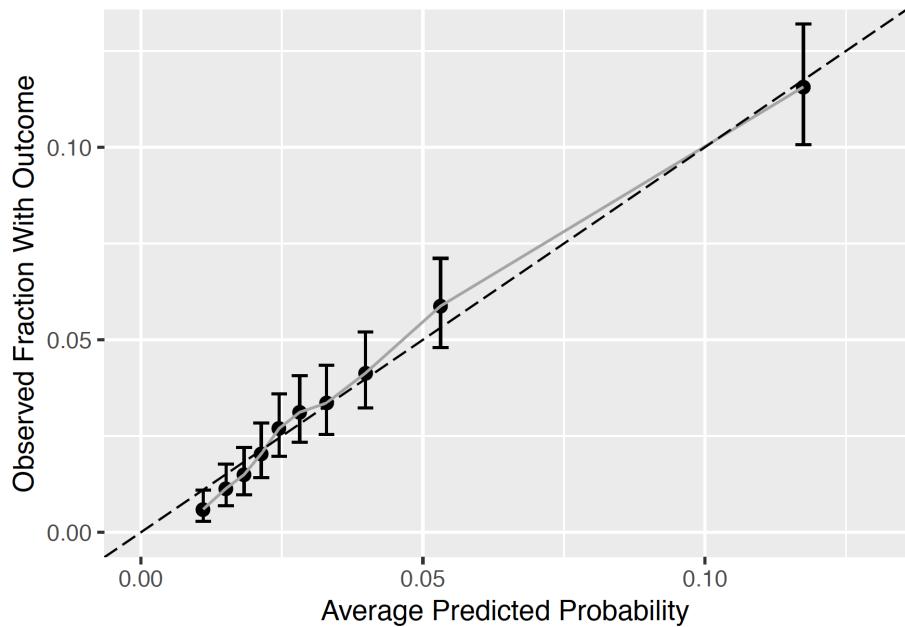
14.3.1 Discrimination

The Receiver Operating Characteristics (ROC) plot shows the sensitivity against 1-specificity on the test set. The plot illustrates how well the model is able to discriminate between the people with the outcome and those without. The dashed diagonal line is the performance of a model that randomly assigns predictions. The higher the area under the ROC plot the better the discrimination of the model. The plot is created by changing the probability threshold to assign the positive class.



14.3.2 Calibration

The calibration plot shows how close the predicted risk is to the observed risk. The diagonal dashed line thus indicates a perfectly calibrated model. The ten (or fewer) dots represent the mean predicted values for each quantile plotted against the observed fraction of people in that quantile who had the outcome (observed fraction). The straight black line is the linear regression using these 10 plotted quantile mean predicted vs observed fraction points. The straight vertical lines represented the 95% lower and upper confidence intervals of the slope of the fitted line.



14.3.3 Smooth Calibration

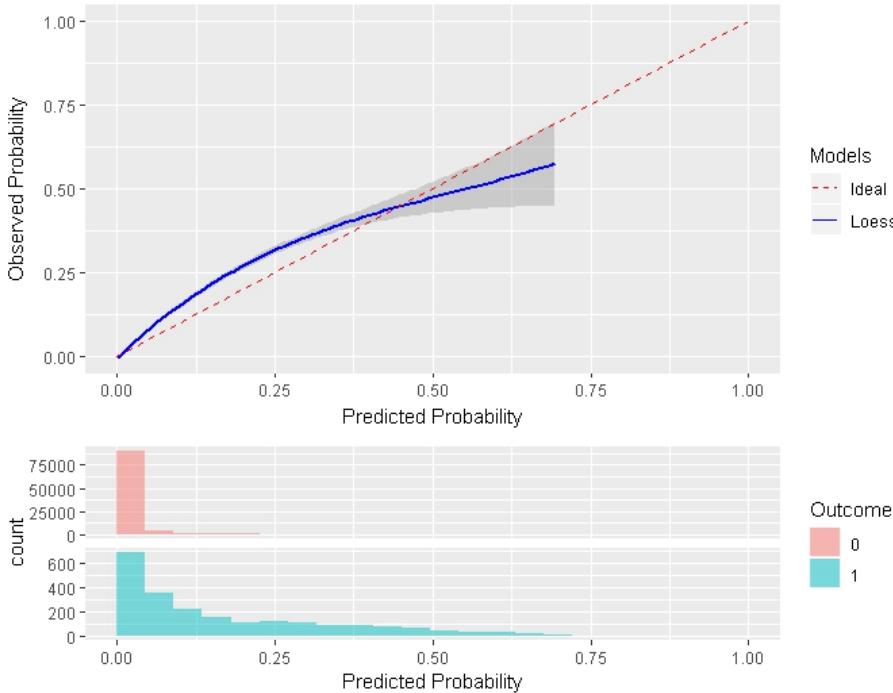
Similar to the traditional calibration shown above the Smooth Calibration plot shows the relationship between predicted and observed risk. the major difference is that the smooth fit allows for a more fine grained examination of this. Whereas the traditional plot will be heavily influenced by the areas with the highest density of data the smooth plot will provide the same information for this region as well as a more accurate interpretation of areas with lower density. the plot also contains information on the distribution of the outcomes relative to predicted risk.

However, the increased information gain comes at a computational cost. It is recommended to use the traditional plot for examination and then to produce the smooth plot for final versions. To create the smooth calibarion plot you have to run the follow command:

```
plotSmoothCalibration(lrResults)
```

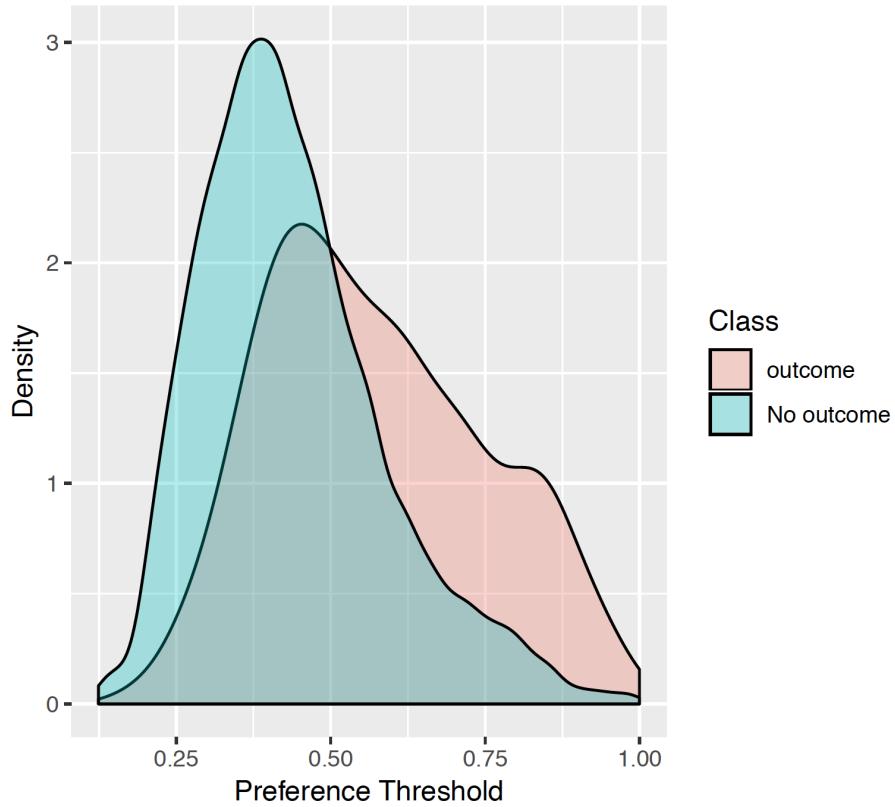
See the help function for more information, on how to set the smoothing method etc.

The example below is from another study that better demonstrates the impact of using a smooth calibration plot. The default line fit would not highlight the miss-calibration at the lower predicted probability levels that well.



14.3.4 Preference distribution

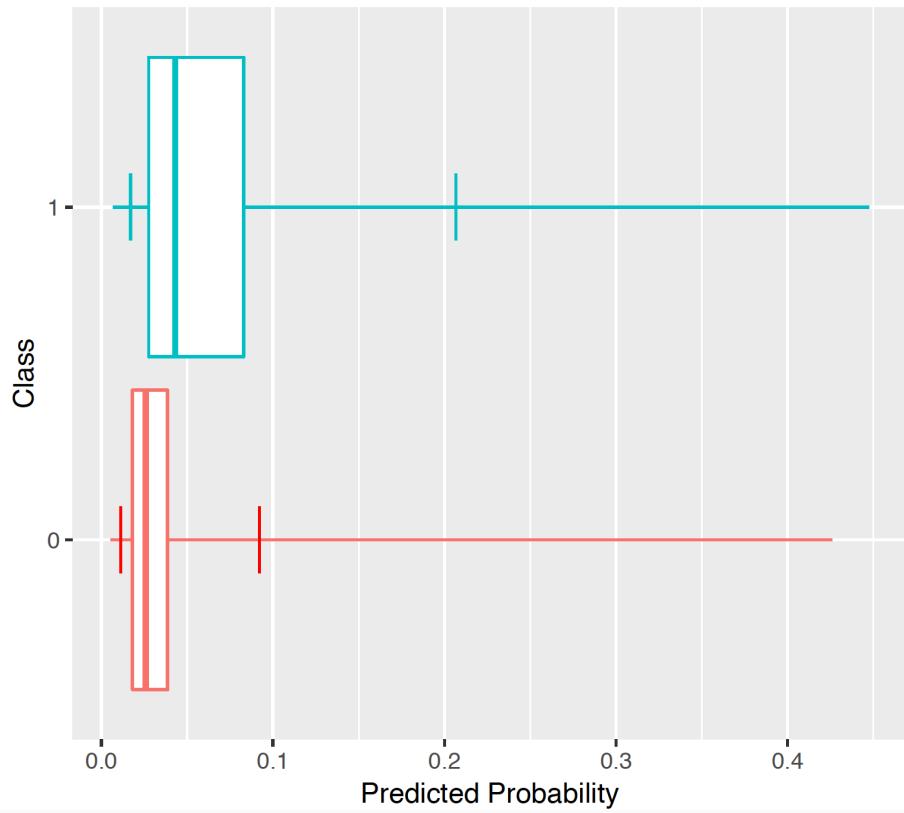
The preference distribution plots are the preference score distributions corresponding to i) people in the test set with the outcome (red) and ii) people in the test set without the outcome (blue).



14.3.5 Predicted probability distribution

The prediction distribution box plots are for the predicted risks of the people in the test set with the outcome (class 1: blue) and without the outcome (class 0: red).

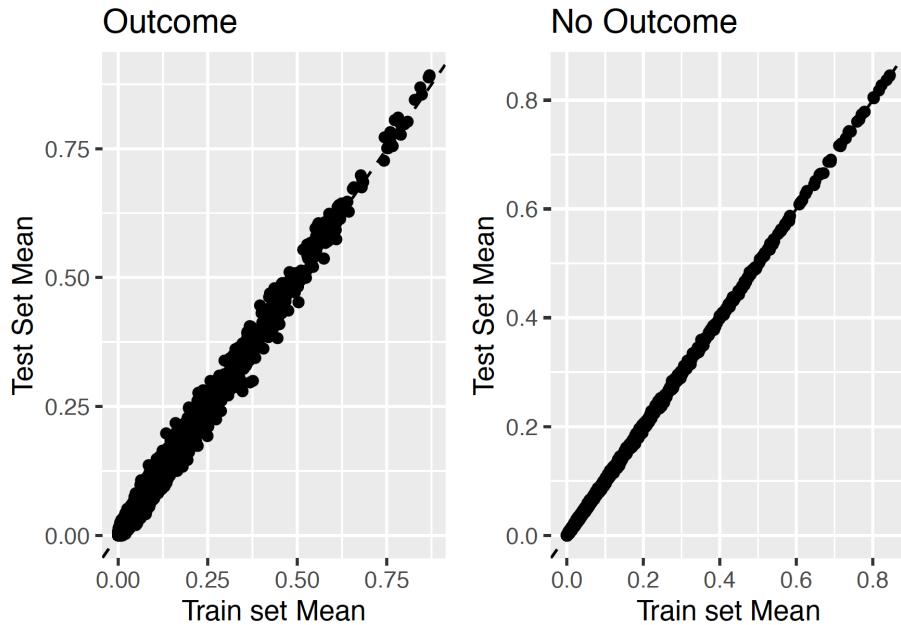
The box plots in the Figure show that the predicted probability of the outcome is indeed higher for those with the outcome but there is also overlap between the two distribution which lead to an imperfect discrimination.



14.3.6 Test-Train similarity

The test-train similarity is assessed by plotting the mean covariate values in the train set against those in the test set for people with and without the outcome.

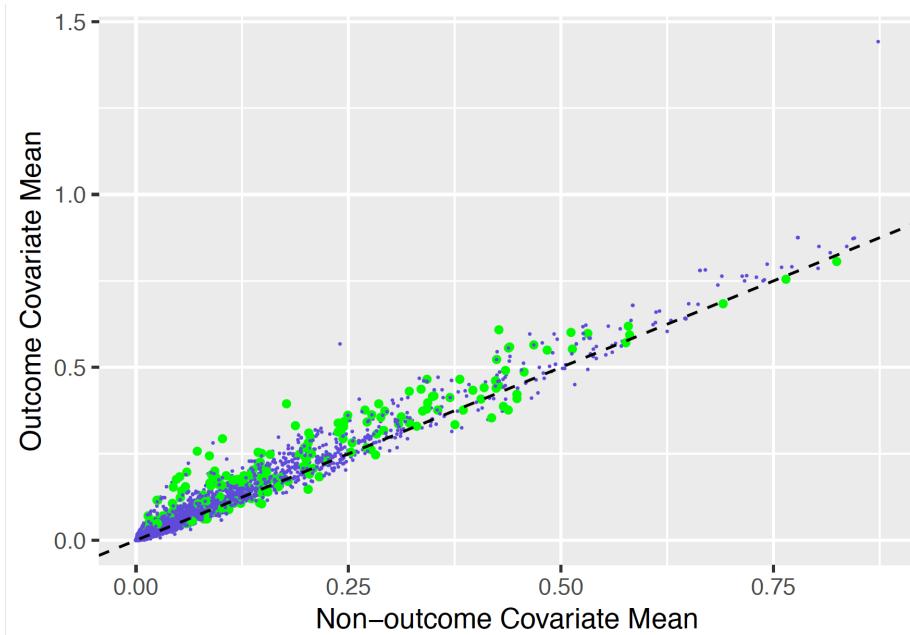
The results for our example of look very promising since the mean values of the covariates are on the diagonal.



14.3.7 Variable scatter plot

The variable scatter plot shows the mean covariate value for the people with the outcome against the mean covariate value for the people without the outcome. The color of the dots corresponds to the inclusion (green) or exclusion in the model (blue), respectively. It is highly recommended to use the Shiny App since this allows you to hoover over a covariate to show more details (name, value etc).

The plot shows that the mean of most of the covariates is higher for subjects with the outcome compared to those without.



14.3.8 Precision recall

Precision (P) is defined as the number of true positives (Tp) over the number of true positives plus the number of false positives (Fp).

```
P <- Tp/(Tp+Fp)
```

Recall (R) is defined as the number of true positives (Tp) over the number of true positives plus the number of false negatives (Fn).

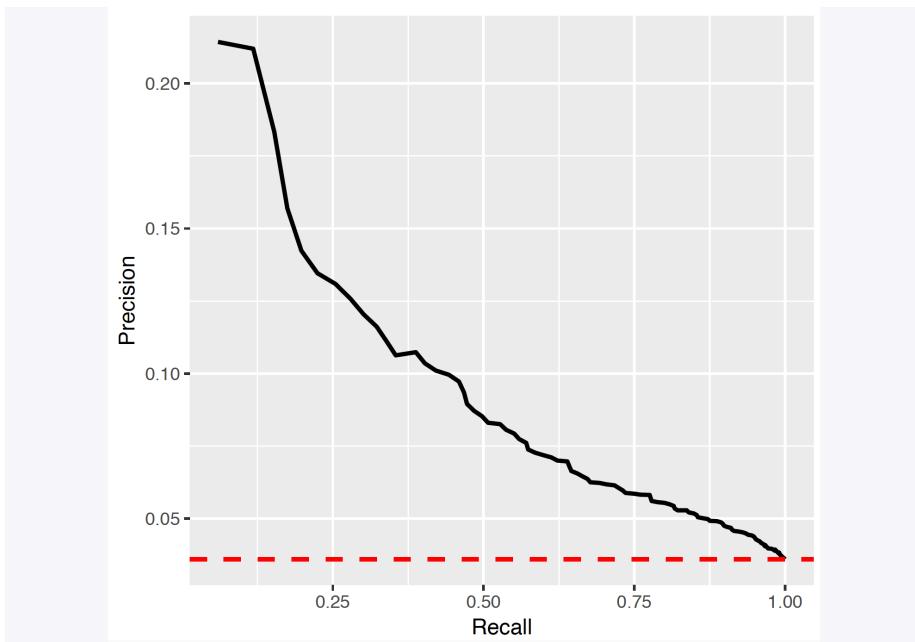
```
R <- Tp/(Tp + Fn)
```

These quantities are also related to the (F1) score, which is defined as the harmonic mean of precision and recall.

```
F1 <- 2*P*R/(P+R)
```

Note that the precision can either decrease or increase if the threshold is lowered. Lowering the threshold of a classifier may increase the denominator, by increasing the number of results returned. If the threshold was previously set too high, the new results may all be true positives, which will increase precision. If the previous threshold was about right or too low, further lowering the threshold will introduce false positives, decreasing precision.

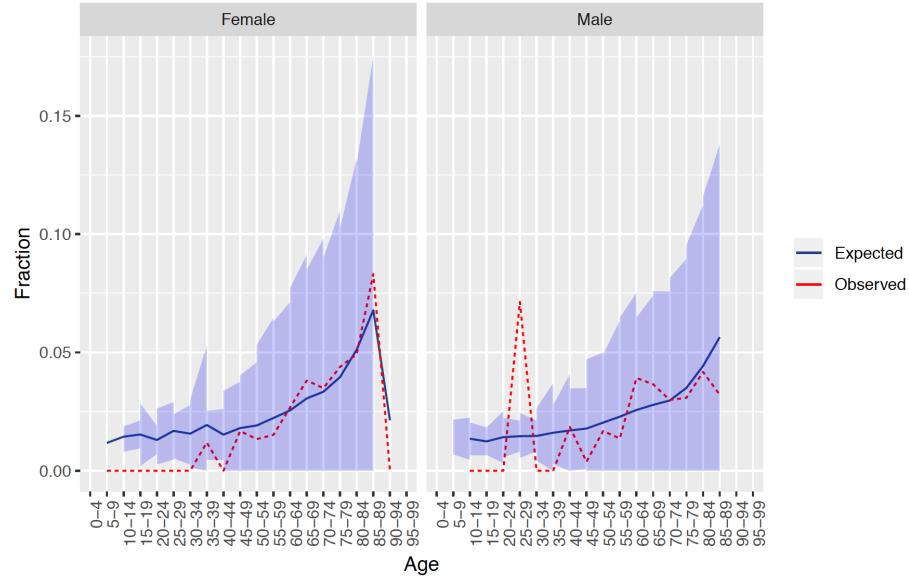
For Recall the denominator does not depend on the classifier threshold ($Tp+Fn$ is a constant). This means that lowering the classifier threshold may increase recall, by increasing the number of true positive results. It is also possible that lowering the threshold may leave recall unchanged, while the precision fluctuates.



14.3.9 Demographic summary

This plot shows for females and males the expected and observed risk in different age groups together with a confidence area.

The results show that our model is well calibrated across gender and age groups.




```
)
```

This will extract the new plpData from the specified schemas and cohort tables. It will then apply the same population settings and the trained plp model. Finally, it will evaluate the performance and return the standard output as `validation$performance` and if `keepPrediction` is TRUE then it will also return the prediction on the population as `validation$prediction`. They can be inserted into the shiny app for viewing the model and validation by running: `viewPlp(runPlp=plpResult, validatePlp=validation)`.

If you want to validate on multiple databases available you can insert the new schemas and cohort tables as a list:

```
# load the trained model
plpResult <- loadPlpResult(getwd(),'plpResult')

connectionDetails <- createConnectionDetails(dbms = "postgresql",
                                               server = "localhost/ohdsi",
                                               user = "joe",
                                               password = "supersecret")

validation <- externalValidatePlp(plpResult = plpResult,
                                    connectionDetails = connectionDetails,
                                    validationSchemaTarget = list('new_cohort_schema1',
                                                               'new_cohort_schema2'),
                                    validationSchemaOutcome = list('new_cohort_schema1',
                                                               'new_cohort_schema2'),
                                    validationSchemaCdm = list('new_cdm_schema1',
                                                               'new_cdm_schema2'),
                                    validationTableTarget = list('new_cohort_table1',
                                                               'new_cohort_table2'),
                                    validationTableOutcome = list('new_cohort_table1',
                                                               'new_cohort_table2'),
                                    validationIdTarget = 'cohort_id',
                                    validationIdOutcome = 'outcome_id',
                                    keepPrediction = T
)
```

14.5 Journal paper generation

We have added functionality to automatically generate a word document you can use as start of a journal paper. It contains many of the generated study details and results. If you have performed external validation these results will can be added as well. Optionally, you can add a “Table 1” that contains data on many covariates for the target population.

You can create the draft journal paper by running this function:

```
createPlpJournalDocument(plpResult = <your plp results>,
                         plpValidation = <your validation results>,
                         plpData = <your plp data>,
                         targetName = "<target population>",
                         outcomeName = "<outcome>",
                         table1 = F,
                         connectionDetails = NULL,
                         includeTrain = FALSE,
                         includeTest = TRUE,
                         includePredictionPicture = TRUE,
                         includeAttritionPlot = TRUE,
                         outputLocation = "<your location>")
```

For more details see the help page of the function.

14.6 Other functionality

The package has much more functionality than described in this vignette and contributions have been made by many persons in the OHDSI community. The table below provides an overview:

Functionality	Description	Vignette
Builing Multiple Models	This vignette describes how you can run multiple models automatically	Vignette
Custom algorithms	This vignette describes how you can add your own custom algorithms in the framework	Vignette
Ensemble models	This vignette describes how you can use the framework to build ensemble models, i.e combine multiple models in a super learner	Vignette
Deep Learning Models	We have added extensive functionality for Deep Learning including several architectures in both pyTorch and Keras. These algorithms can be trained using GPU power	Vignette
Learning curves	Learning curves assess the effect of training set size on model performance by training a sequence of prediction models on successively larger subsets of the training set. A learning curve plot can also help in diagnosing a bias or variance problem as explained below.	Vignette
Implementing existing models	This vignette describes how you can implement existing logistic regression models in the framework, e.g. as found in literature	Vignette

14.7 Demos

We have added several demos in the package that run on simulated data:

```
# Show all demos in our package:
demo(package = "PatientLevelPrediction")

# For example, to run the SingleModelDemo that runs Lasso and shows you how to run the
demo("SingleModelDemo", package = "PatientLevelPrediction")
```

14.8 Acknowledgments

Considerable work has been dedicated to provide the `PatientLevelPrediction` package.

```
# citation("PatientLevelPrediction")
```

Further, `PatientLevelPrediction` makes extensive use of the `Cyclops` package.

```
# citation("Cyclops")
```

Please reference this paper if you use the PLP Package in your work:

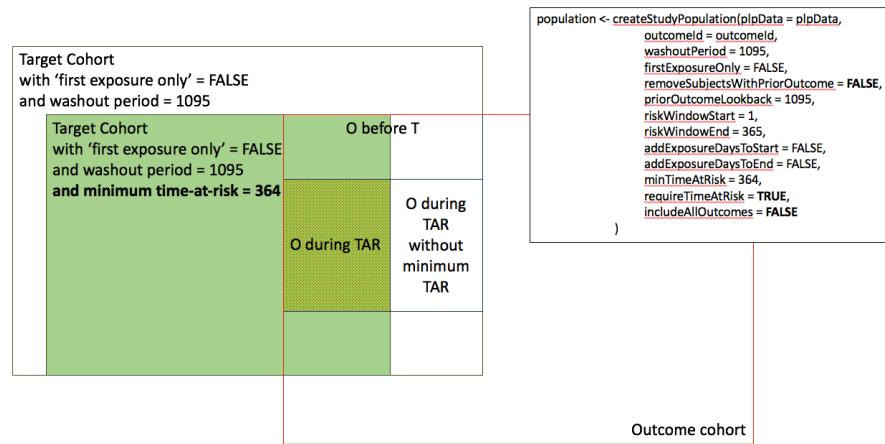
Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc.* 2018;25(8):969-975.

This work is supported in part through the National Science Foundation grant IIS 1251151 and a personalised grant to P.R. Rijnbeek from Janssen R&D.

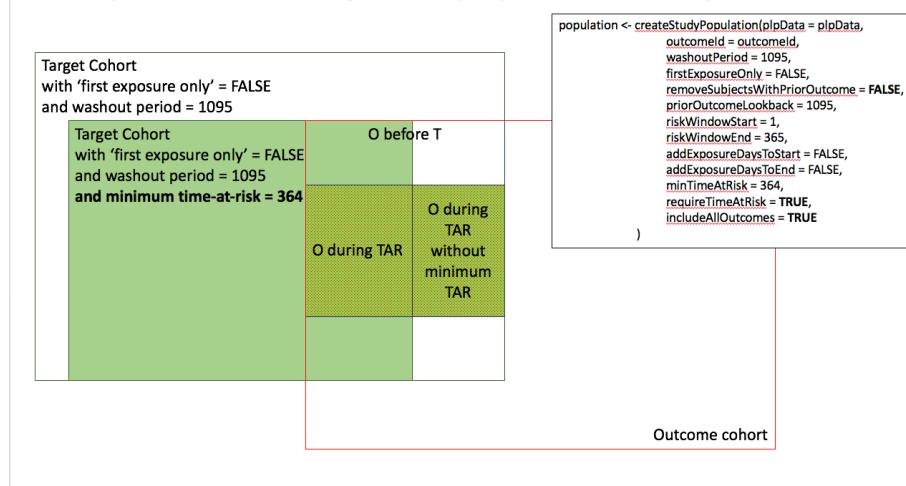
Appendix 1: Study population settings details

In the figures below the effect is shown of the removeSubjectsWithPriorOutcome, requireTimAtRisk, and includeAllOutcomes booleans on the final study population. We start with a Target Cohort with firstExposureOnly = false and we require a washout period = 1095. We then subset the target cohort based on additional constraints. The final study population in the Venn diagrams below are colored green.

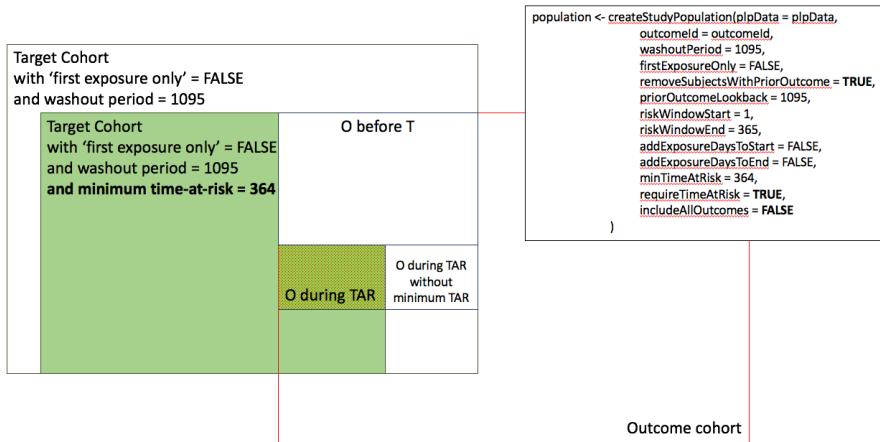
1. Require minimum time-at-risk for all person in the target cohort.



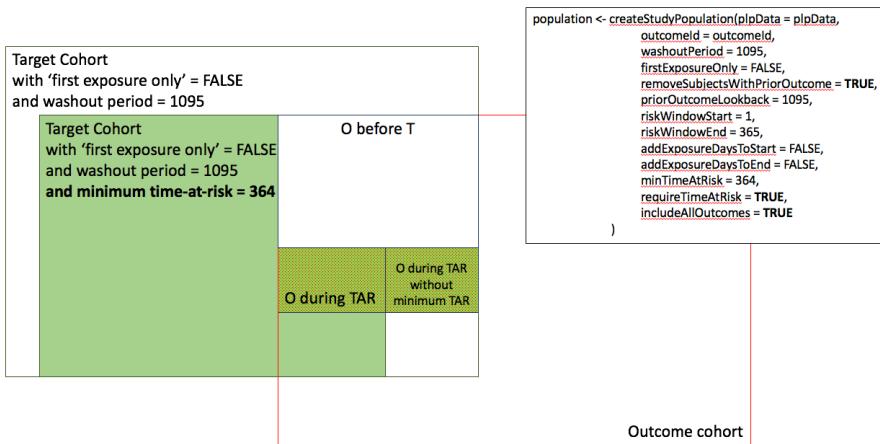
2. Require minimum time-at-risk for target cohort, except for persons with outcomes during time-at-risk.



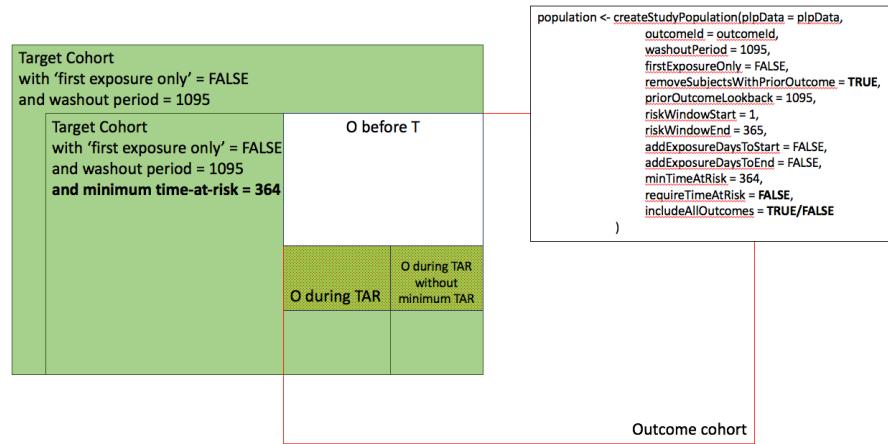
3. Include all persons in the target cohort exclude persons with prior outcomes.



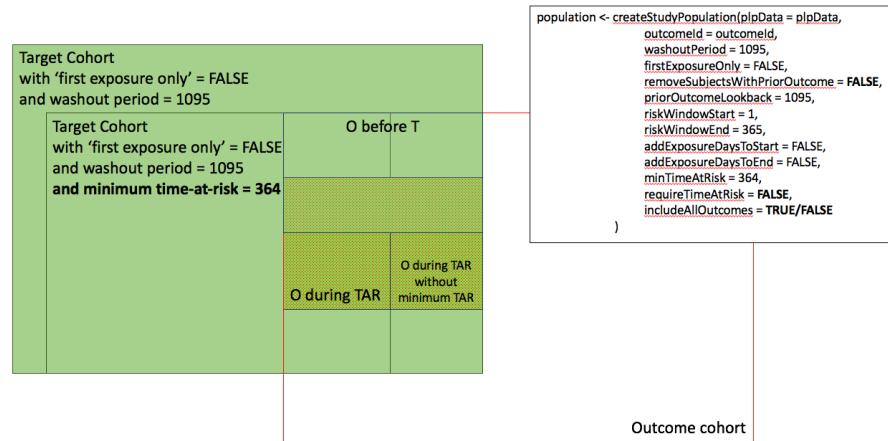
4. Require minimum time-at-risk for target cohort, except for persons with outcomes during time-at-risk, exclude persons with prior outcomes.



5. Include all persons in target cohort exclude persons with prior outcomes.



6. Include all persons in target cohort.



Part IV

Evidence Quality

Chapter 15

Evidence Quality

Loss of fidelity begins with the movement of data from the doctor's brain to the medical record.

Clem McDonald, MD Director, Lister Hill Center for Biomedical Informatics National Library of Medicine, USA

OHDSI views validation as a holistic set of processes necessary to achieve the highest quality reproducible evidence from diverse data sources.

Four components: - Data quality (data validation) - Clinical validity - Software validity - Method validity

Chapter 16

Data Quality

16.1 Introduction

Kahn et al. define data quality as consisting of three components: (1) conformance (do data values adhere to do specified standard and formats?; subtypes: value, relational and computational conformance); (2) completeness (are data values present?); and (3) plausibility (are data values believable?; subtypes uniqueness, atemporal; temporal) (Kahn et al., 2016)

Kahn additionally defines two contexts: verification and validation. Verification focuses on model and data constraints and does not rely on external reference. Validation focuses on data expectations that are derived from comparison to a relative gold standard and uses external knowledge.

Term	Subtype	Validation example
Conformance	Value	Providers are only assigned valid medical specialties.
	Relational	Prescribing provider identifier is present in drug dispensation data.
	Computational	Computed eGFR value conforms to the expected value for a test case patient scenario.
Completeness (no subtypes defined)		A drug product withdrawn from the market at a specific absolute historic date shows expected drop in dispensation.
Plausibility	Uniqueness	A zip code for a location does not refer to vastly conflicting geographical areas.
	Atemporal	Use of a medication (by age group) for a specific disease agrees with the age pattern for that disease.
	Temporal	Temporal pattern of an outbreak of a disease (e.g., Zika) agrees with external source pattern.

Kahn introduces the term *data quality check* (sometimes referred to as data quality rule) that tests whether data conform to a given requirement (e.g., implausible age of 141 of a patient (due to incorrect birth year or missing death event)). In support of checks, he also defines *data quality measure* (sometimes referred to as pre-computed analysis) as data analysis that supports evaluation of a check. For example, distribution of days of supply by drug concept.

Two types of DQ checks can be distinguished(Weiskopf and Weng, 2013)

- general checks
- study-specific checks

From the point of researcher analyzing the data, the desired situation is that data is free from errors that could have been prevented. *ETL data errors* are errors introduced during extract-tranform-load proces. A special type of ETL data error is *mapping error* that results from incorrect mapping of the data from the source terminology (e.g., Korean national drug terminology) into the target data model's standard terminology (e.g., RxNorm and RxNorm Extension). A *source data error* is an error that is already present in the source data due to various causes (e.g., human typo during data entry).

Data quality can also be seen as a component in a larger effort referred to as *evidence quality* or *evidence validation*. Data quality would fall in this framework under *data validation*.

16.2 Achilles Heel tool

Since 2014, a component of the OHDSI Achilles tool called Heel was used to check data quality.(Huser et al., 2018)

16.2.1 Precomputed Analyses

In support of data characterization, Achilles tool pre-computes number of data analyses. Each pre-computed analysis has an analysis ID and a short description of the analysis. For example, “715: Distribution of days_supply by drug_concept_id” or “506: Distribution of age at death by gender”. List of all pre-computed analyses (for Achilles version 1.6.3) as available at https://github.com/OHDSI/Achilles/blob/v1.6.3/inst/csv/achilles/achilles_analysis_details.csv

Achilles has more than 170 pre-computed analysis that support not only data quality checks but also general data characterization (outside data quality context) such as data density visualizations. The pre-computations are largely guided by the CDM relational database schema and analyze most terminology-based data columns, such as condition_concept_id or

place_of_service_concept_id. Pre-computations results are stored in table ACHILLES_RESULTS and ACHILLES_RESULTS_DIST.

16.2.2 Example DQ check

In complete data about general population, a range of services is provided by a range of providers (with many specialties). A data completeness rule with rule_id of 38 evaluates data completeness in the PROVIDER table. Checking optional fields in CDM (such as provider specialty) lead to a notification severity output. Analysis Rule 38 triggers a notification if count of distinct specialties <2. It relies on a derived measure Provider:SpecialtyCnt. The rule SQL-formulated logic can be found here: https://github.com/OHDSI/Achilles/blob/v1.6.3/inst/sql/sql_server/heels/serial/rule_38.sql

16.2.3 Overview of existing DQ Heel checks

Achilles developers maintain a list of all DQ checks in an overview file. For version 1.6.3, this overview is available here https://github.com/OHDSI/Achilles/blob/v1.6.3/inst/csv/heel/heel_rules_all.csv. Each DQ check has a rule_id.

Checks are classified into CDM conformance checks and DQ checks.

Depending on the severity of the problem, the Heel output can be error, warning or notification.

16.3 Study-specific checks

The chapter has so far focused on general DQ checks. Such checks are executed regardless of the single research question context. The assumption is that a researcher would formulate additional DQ checks that are required for a specific research question.

We use case studies to demonstrate study-specific checks.

16.3.1 Outcomes

For an international analysis, part of OHDSI study diagnostics (for a given dataset) may involve checking whether coding practices (that are country specific) affect a cohort definition. A stringent cohort definition may lead to zero cohort size in one (or multiple datasets).

16.3.2 Laboratory data

A diabetes study may utilize HbA1c measurement. A 2018 OHDSI study (<https://www.ncbi.nlm.nih.gov/pubmed/30646124>) defined a cohort ‘HbA1c8Moderate’ (see <https://github.com/rohit43/DiabetesTxPath/blob/master/inst/settings/CohortsToCreate.csv>)

Chapter 17

Clinical Validity

Chapter 18

Software Validity

Chapter 19

Method Validity

When considering method validity we aim to answer the question

Is this method valid for answering this question?

Where ‘method’ includes not only the study design, but also the data and the implementation of the design. Method validity is therefore somewhat of a catch-all; It is often not possible to observe good method validity without good data quality, clinical validity, and software validity. Those aspects of evidence quality should have already been addressed separately before we consider method validity.

The core activity when establishing method validity is evaluating whether important assumptions in the analysis have been met. For example, we assume that propensity-score matching makes two populations comparable, but we need to evaluate whether this is the case. Where possible, empirical tests should be performed to test these assumptions. We can for example generate diagnostics to show that our two populations are indeed comparable on a wide range of characteristics after matching. In OHDSI we have developed a wide range of standardized diagnostics that should be generated and evaluated whenever an analysis is performed. Some of these diagnostics are specific to certain study designs, whereas others are more generic.

19.1 Design-specific diagnostics

For each study design there are diagnostics specific to such a design. Here we review some of the standard diagnostics included in the OHDSI Methods Library R packages. This review is not exhaustive, and we recommend the reader to consult the documentation for each method to learn about all implemented diagnostics.

19.1.1 Diagnostics for cohort method

In the comparative cohort design we compare two cohorts, for example representing two treatment choices, and we want to evaluate whether the treatment choice has an effect on the risk of some outcome of interest. For the effect size estimate w to be valid, it is essential that the two groups are comparable in all relevant aspects except the treatment choice. In observational data, this comparability is by no means guaranteed, and quite often there is a reason why one group gets a treatment while the group does not, leading to fundamental differences between the groups. We often employ propensity scores to make the two groups comparable again, but that assumes there is at least some commonality between the two groups. This assumption can be tested by reviewing the preference score plot as shown in Figure 19.1 (The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups). We can evaluate whether there are patients that had some probability of receiving either treatment. In Figure 19.1 we see large numbers of people on the left and right for whom their treatment choice could have been predicted fairly accurately based on the baseline characteristics, meaning that without adjustment the two groups are incomparable. However, we also observe a substantial area of overlap, the purple area, where people were likely to get either treatment. This suggests that with some adjustment, for example using propensity score matching, the two groups can be made comparable. It is important to note that a large overlap can also be due to an unpredictable propensity model, for example because key characteristics were not included in the model. A lack of overlap can be due to including variables directly related to the exposure, such as including the procedure code for an injection if one of the treatments is an injectable. This needs to be ruled out by examining the propensity model.

Once we believe there is some hope of making the two groups comparable, we need to evaluate whether we indeed succeed by examining a large number of baseline characteristics after adjustment. Figure 19.2 shows the absolute standardized difference of the mean between the two groups for a large number of covariates, both before and after matching on the propensity score. A rule-of-thumb that is often used is to consider any variable with absolute standardized difference of the mean < 0.1 to be in balance. We see in Figure 19.2 that many covariates show imbalance before matching, but matching achieves balance on all covariates.

19.1.2 Diagnostics for SCCS

One assumption in the self-controlled case series (SCCS) design is that the end of observation is independent of the outcome. This assumption is often violated in the case of serious, potentially lethal, events such as myocardial infarction. We can evaluate whether the assumption holds by generating the

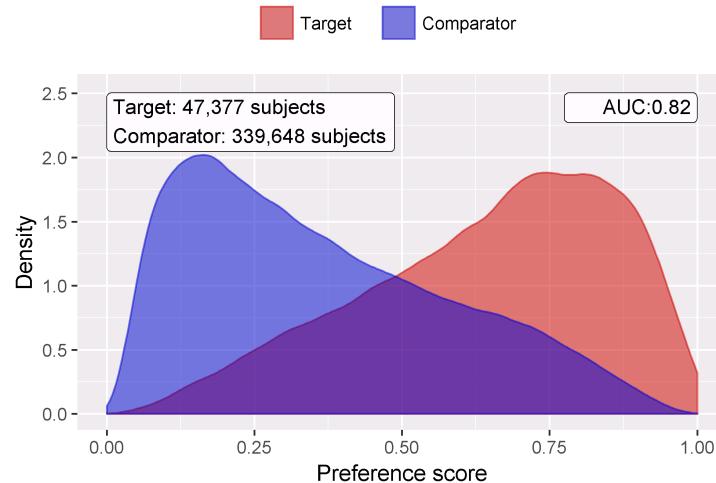


Figure 19.1: Preference score distribution. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups. A higher overlap indicates subjects in the two groups were more similar in terms of their predicted probability of receiving one treatment over the other.

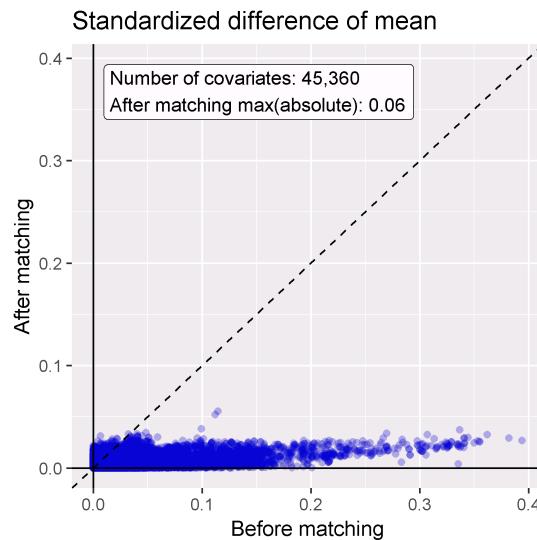


Figure 19.2: Covariate balance before and after matching. Each dot represents the standardized difference of means for a single covariate before and after matching on the propensity score.

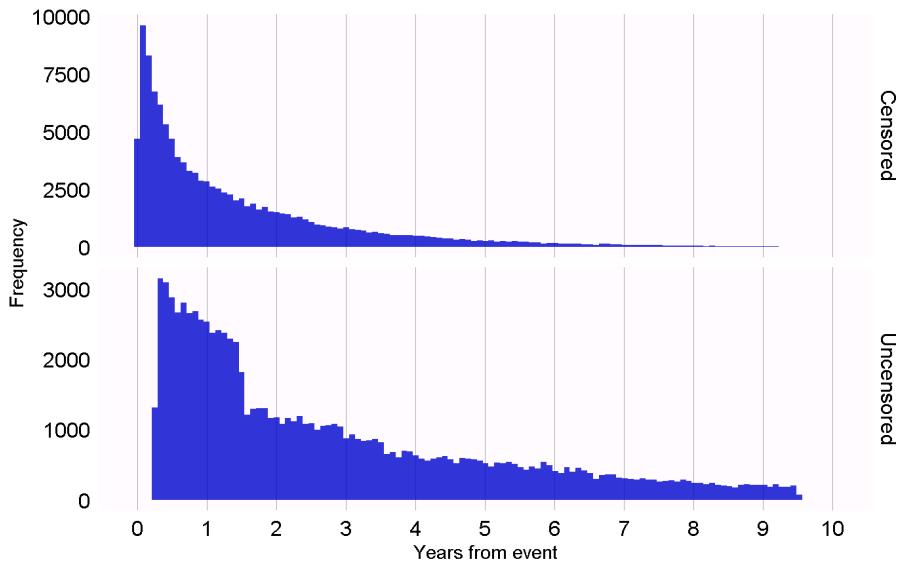


Figure 19.3: Time to observation end for those that are censored, and those that uncensored.

plot shown in Figure 19.3, which shows a histograms of the time to obsevation period end for those that are censored, and those that uncensored. In our data we consider those whose observation period ends at the end date of data capture (the date when observation stopped for the entire data base, for example the date of extraction, or the study end date) to be uncensored, and all others to be censored. In Figure 19.3 we see only minor differences between the two distributions, suggesting our assumptions holds.

19.2 Diagnostics for all estimation

Some diagnostics are applicable for all population-level estimation studies. These require the inclusion of control hypotheses, research questions where the answer is already known. We can then evaluate whether our design produces results in line with the truth. Controls can be divided into negative controls and positive controls.

19.2.1 Negative and positive controls

Negative controls are exposure-outcome pairs where one believes no causal effect exists, and including negative controls or ‘falsification endpoints’ (Prasad

and Jena, 2013) has been recommended as a means to detect confounding (Lipsitch et al., 2010), selection bias and measurement error (Arnold et al., 2016). For example, in one study (Zaadstra et al., 2008) investigating the relationship between childhood diseases and later multiple sclerosis (MS), the authors include three negative controls that are not believed to cause MS: a broken arm, concussion, and tonsillectomy. Two of these three controls produce statistically significant associations with MS, suggesting that the study may be biased. We should select negative controls that are comparable to our hypothesis of interest, which means we typically select exposure-outcome pairs that either have the same exposure as the hypothesis of interest (so-called ‘outcome controls’) or the same outcome (‘exposure controls’). In OHDSI we have developed a semi-automated procedure for selecting negative controls (Voss et al., 2016). In brief, information from literature, product labels, and spontaneous reporting is automatically extracted and synthesized to produce a candidate list of outcomes with no known links with any hypertension treatment. We rank-order this list by prevalence in an observational database and manually review these in order.

To understand the behavior of a method when the true relative risk is smaller or greater than one requires the use of positive controls, where the null is believed to not be true. Unfortunately, real positive controls for observational research tend to be problematic for three reasons. First, in most research contexts, for example when comparing the effect of two treatments, there is a paucity of positive controls relevant for that specific context. Second, even if positive controls are available, the magnitude of the effect size may not be known with great accuracy, and often depends on the population in which one measures it. Third, when treatments are widely known to cause a particular outcome, this shapes the behavior of physicians prescribing the treatment, for example by taking actions to mitigate the risk of unwanted outcomes, thereby rendering the positive controls useless as a means for evaluation (Noren et al., 2014). In OHDSI we therefore use synthetic positive controls (Schuemie et al., 2018), created by modifying a negative control through injection of additional, simulated occurrences of the outcome during the time at risk of the exposure. One issue that stands important is the preservation of confounding. The negative controls may show strong confounding, but if we inject additional outcomes randomly, these new outcomes will not be confounded, and we may therefore be optimistic in our evaluation of our capacity to deal with confounding for positive controls. To preserve confounding, we want the new outcomes to show similar associations with baseline subject-specific covariates as the original outcomes. To achieve this, we fit large-scale predictive models for each negative control using L_1 regularized survival regression (Suchard et al., 2013). We insert new outcomes by drawing from the per-subject predicted probabilities within the exposed population until we achieve the desired incidence rate ratio. Figure 19.4 depicts this process.

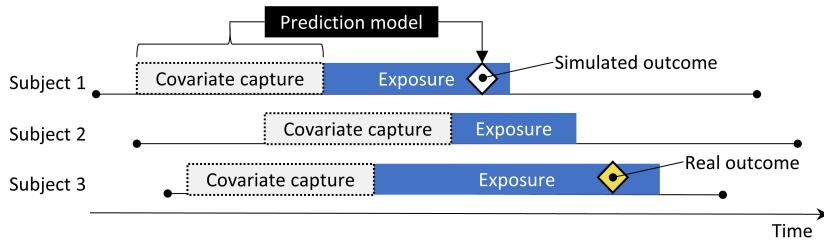


Figure 19.4: Synthesizing positive controls from negative controls.

19.2.2 Metrics

Based on the estimates of a particular method for the negative and positive controls, we can then compute a wide range of metrics, for example:

- **Area Under the receiver operator Curve (AUC):** the ability to discriminate between positive and negative controls.
- **Coverage:** how often the true effect size is within the 95% confidence interval.
- **Mean precision:** precision is computed as $1 / (\text{standard error})^2$, higher precision means narrower confidence intervals. We can use the geometric mean to account for the skewed distribution of the precision.
- **Mean squared error (MSE):** Mean squared error between the log of the effect size point-estimate and the log of the true effect size.
- **Type 1 error:** For negative controls, how often was the null rejected (at alpha = 0.05). This is equivalent to the false positive rate and 1 - specificity.
- **Type 2 error:** For positive controls, how often was the null not rejected (at alpha = 0.05). This is equivalent to the false negative rate and 1 - sensitivity.
- **Non-estimable:** For how many of the controls was the method unable to produce an estimate? There can be various reasons why an estimate cannot be produced, for example because there were no subjects left after propensity score matching, or because no subjects remained having the outcome.

19.2.3 Empirical calibration

19.2.4 OHDSI Methods Benchmark

19.2.5 Replication across sites

- Warn against ‘Christie Brinkley Bias’

19.3 Diagnostics for all prediction

19.4 Method validation in practice

Example: risk of angioedema and AMI in new users of ACE inhibitors compared to new users of thiazide and thiazide-like diuretics

How to select negative controls using ATLAS

Include negative and positive controls.

Compute metrics

Generate calibration plots

Calibrate CI and p-value

Part V

OHDSI Studies

Chapter 20

Study steps

Writing the protocol, OHDSI style: http://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=projects:workgroups:wg_study_protocols_eastern_hemisphere.pptx

Study reproducibility (Martijn has some slides that might help: http://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=projects:workgroups:wg_study_reproducability.pptx)

Chapter 21

OHDSI Network Research

Contributors: Greg Klebanov, Vojtech Huser, list others

What is OHDSI Network?

- OHDSI Community and Network Research
- International Open Science Networks
- OHDSI US
- OHDSI EU and EHDEN
- OHDSI APAC

OHDSI Network Study Process

- Goals
- Workflow Overview
- Structure of Studies
- Protocol and IRB issues
- Existing framework (de-identified [time shifted] OMOP dataset under existing IRB protocol
- Overcoming Network Study Challenges
- Data Privacy, Security and Compliance
- Data Quality
- Running OHDSI Methods in Isolated Environment
- OMOP CDM Versioning

Tools, Platforms and Study Automation * OHDSI Methods support for Network Studies * LEGEND (should we have it here?) * OHDSI ARACHNE Network Platform

Opportunities, future trends and Roadmap

21.1 OHDSI Network Study Examples

21.1.1 Endometriosis study

An endometriosis characterization study (available at <https://github.com/molliemckillip/Endometriosis-Phenotype-Characterization>) works with two cohorts. They are defined in cohorts.csv file (see here <https://github.com/molliemckillip/Endometriosis-Phenotype-Characterization/blob/master/inst/settings/cohorts.csv>).

After creating a cohort table, it is populated by executing this command here by inferring a name of a ‘.sql’ file from the previously defined cohort file. A createCohorts function is executed next. (see <https://github.com/molliemckillip/Endometriosis-Phenotype-Characterization/blob/master/R/createCohorts.R>). An SQL file that is generated by Atlas populates the cohort table with specific person_ids that fulfill the cohort definition.

21.2 Excercises

21.2.1 Defining a cohort

Q: Study the code for the x study and determine whether the cohort definition is available on the public OHDSI server. If it is, what is the cohort ID there?

A:

Appendix A

Glossary

Cohort A cohort is a list of person_ids with start and end date. It is stored in a study specific cohort table or a CDM specified cohort table can also be used. Cohort can be represented as .json file. It is used for import and export but not during an analysis. OHDSI tools use SQL so Atlas also generates a .sql file that creates the cohort during analysis.

Parametrized SQL code An SQL code that allows for use of parameters. Parameters are prefixed with @. Such code has to be “rendered”. Synonym: OHDSI SQL code.

Bibliography

- Arnold, B. F., Ercumen, A., Benjamin-Chung, J., and Colford, J. M. (2016). Brief Report: Negative Controls to Detect Selection Bias and Measurement Bias in Epidemiologic Studies. *Epidemiology*, 27(5):637–641.
- Engel, C. and Fischer, C. (2015). Breast cancer risks and risk prediction models. *Breast Care (Basel)*, 10(1):7–12.
- 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.
- Gage, B. F., Waterman, A. D., Shannon, W., Boechler, M., Rich, M. W., and Radford, M. J. (2001). Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA*, 285(22):2864–2870.
- 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.
- Huser, V., Kahn, M. G., Brown, J. S., and Gouripeddi, R. (2018). Methods for examining data quality in healthcare integrated data repositories. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 23:628–633.
- Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., and Schilling, L. (2016). A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Washington, DC)*, 4(1):1244.

- Keogh, C., Wallace, E., Dillon, C., Dimitrov, B. D., and Fahey, T. (2011). Validation of the CHADS2 clinical prediction rule to predict ischaemic stroke. A systematic review and meta-analysis. *Thromb. Haemost.*, 106(3):528–538.
- Lee, K. L., Woodlief, L. H., Topol, E. J., Weaver, W. D., Betriu, A., Col, J., Simoons, M., Aylward, P., Van de Werf, F., and Califf, R. M. (1995). Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. GUSTO-I Investigators. *Circulation*, 91(6):1659–1668.
- Lipsitch, M., Tchetgen Tchetgen, E., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388.
- 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.
- Nguyen, N. D., Frost, S. A., Center, J. R., Eisman, J. A., and Nguyen, T. V. (2008). Development of prognostic nomograms for individualizing 5-year and 10-year fracture risks. *Osteoporos Int*, 19(10):1431–1444.
- Noren, G. N., Caster, O., Juhlin, K., and Lindquist, M. (2014). Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance. *Drug Saf*, 37(9):655–659.
- Perel, P., Edwards, P., Wentz, R., and Roberts, I. (2006). Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak*, 6:38.
- Perkins, N. J., Cole, S. R., Harel, O., Tchetgen Tchetgen, E. J., Sun, B., Mitchell, E. M., and Schisterman, E. F. (2017). Principled approaches to missing data in epidemiologic studies. *American journal of epidemiology*, 187(3):568–575.
- Prasad, V. and Jena, A. B. (2013). Prespecified falsification end points: can they validate true observational associations? *JAMA*, 309(3):241–242.
- Reps, J. M., Schuemie, M. J., Suchard, M. A., Ryan, P. B., and Rijnbeek, P. R. (2018). Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association*, 25(8):969–975.
- Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.
- 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.

- Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D., and Suchard, M. A. (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc. Natl. Acad. Sci. U.S.A.*, 115(11):2571–2577.
- 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., 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.
- Vandenbroucke, J. P. and Pearce, N. (2012). Case-control studies: basic concepts. *Int J Epidemiol*, 41(5):1480–1489.
- Voss, E. A., Boyce, R. D., Ryan, P. B., van der Lei, J., Rijnbeek, P. R., and Schuemie, M. J. (2016). Accuracy of an Automated Knowledge Base for Identifying Drug Adverse Reactions. *J Biomed Inform*.
- Weiskopf, N. G. and Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association: JAMIA*, 20(1):144–151.
- Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847.
- Zaadstra, B. M., Chorus, A. M., van Buuren, S., Kalsbeek, H., and van Noort, J. M. (2008). Selective association of multiple sclerosis with infectious mononucleosis. *Mult. Scler.*, 14(3):307–313.