The Book of OHDSI 2nd Edition

Observational Health Data Sciences and Informatics

2025-06-21

Table of contents

Welcome

Welcome to 2nd edition of *The Book of OHDSI*. Here you will find resources for conducting research with the Observational Health Data Sciences Initiative (OHDSI) and the Observational Medical Outcomes Partnership CDM (OMOP).

Sincerely,

The BOO Editorial Committee

How to Contribute

We welcome suggestions, edits, and larger contributions to this guide. This book is typeset in Markdown, rendered with Quarto, and hosted on GitHub. For errors or requests, please submit an Issue to the book's issue tracker. To make larger or direct contributions, please make a pull request using the standard GitHub workflow.

If you would like to contribute but are unfamiliar with any of these technologies, please feel free to email QQQ with comments and suggestions for changes.

How to Cite this Work

O'Neil ST, Beasley W, Loomba J, Patrick S, Wilkins KJ, Crowley KM., Anzalone, AJ (Eds.) (2023). *The Researcher's Guide to N3C: A National Resource for Analyzing Real-World Health Data*. DOI: 10.5281/zenodo.7749367

Editorial Committee:

- Christian Reich: March 2024 - present

- William H. Beasley: March 2024 - present
- (Add more names & probably remove me/Will)

Licensing

Funding

This content is solely the responsibility of the authors and does not necessarily represent the official views of QQQ.

Part I.

OHDSI

1. The OHDSI Community

TODO- write abstract

Chapter Leads: George Hripcsak, Patrick Ryan

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

1.1. The Journey from Data to Evidence

1.2. OMOP

1.3. Adoption

1.4. Stakeholders

1.5. Global Diversity

1.6. Summary

2. OHDSI Principles

TODO- write abstract

Chapter Leads: tbc

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

2.1. Open Science

2.2. Open Standards

2.3. Open Source

2.4. Open Data

2.5. Open Discourse

2.6. Collaboration

3. Where to Begin

TODO- write abstract

Chapter Leads: Kristin Kostka

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

3.1. Join the Journey

3.2. Where You Fit In

3.3. Summary

Part II.

Uniform Data Representation

4. The Common Data Model

TODO- write abstract

Chapter Leads: Clair Blacketer

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter 4 Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

5. Standardized Vocabularies

TODO- write abstract

Chapter Leads: Anna Ostropolets

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter 5 Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

The OHDSI Standardized Vocabularies are a foundational part of the OHDSI research network, and an integral part of the Common Data Model (CDM) (1). They allow standardization of methods, definitions, and results by defining the content of the data, paving the way for true remote (behind the firewall) network research and analytics. Usually, finding and interpreting the content of observational healthcare data, whether it is structured data using coding schemes or laid down in free text, is passed all the way through to the researcher, who is faced with a myriad of different ways to describe clinical

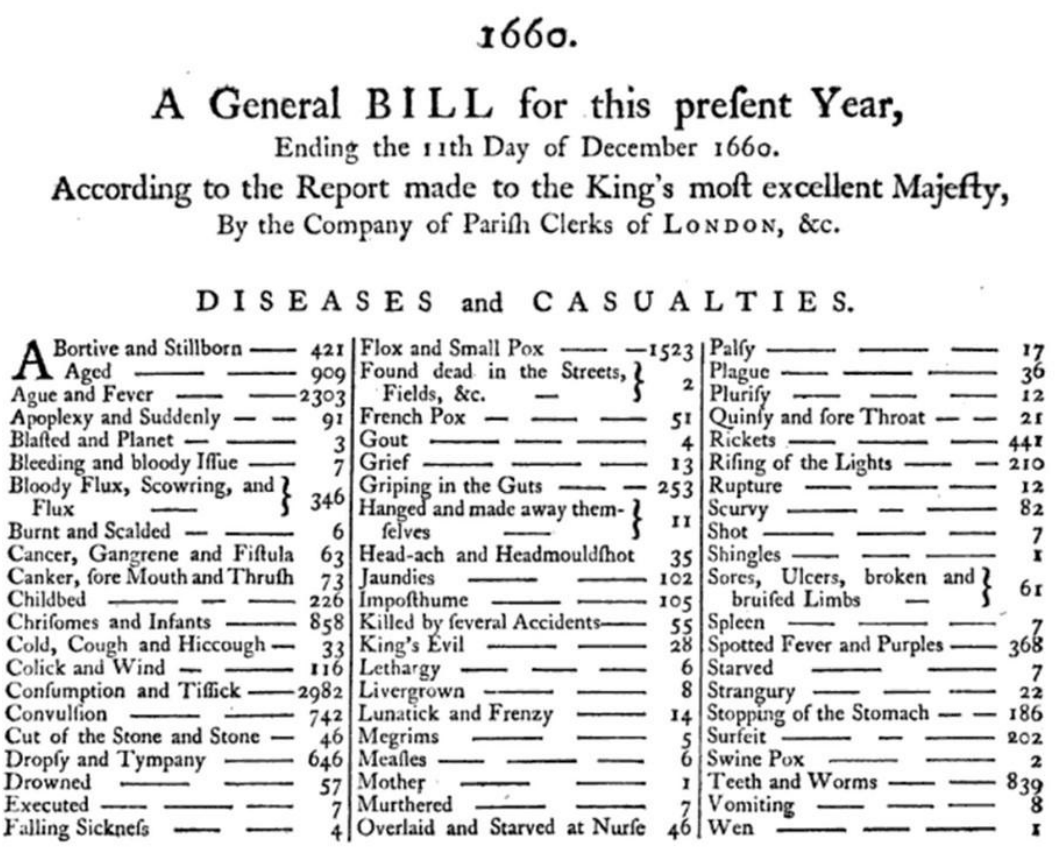
events. OHDSI requires harmonization not only to a standardized format, but also to a rigorous standard content.

The Vocabularies are a collection of public standard vocabularies and terminologies used in the network, which we consolidate from their different original formats and life-cycle conventions into the CDM table structure. The system is dynamic, it evolves with frequent source vocabulary updates, deprecations, and concept replacements, all of which are version-controlled and available via ATHENA - OHDSI's vocabulary distribution platform (2). Vocabularies support phenotyping, covariate construction, large-scale analytics and result reporting and is a product of community effort: Vocabulary Team maintains and improves vocabularies according to the roadmap (3), vocabulary stewards maintain individual vocabularies (4), various Workgroups coordinate efforts for special use cases such as device or vaccine harmonization and contributors across the community add their vocabulary content and improve existing, particularly through enhancing mappings (5).

In this chapter we first describe the main principles of the Standardized Vocabularies and processes. We will walk through Vocabularies components and some typical situations, all of which are necessary to understand and utilize this foundational resource. We also point out where the support of the community is required to continuously improve it and describe special situations that require further development.

5.1. Why Vocabularies, and Why Standardizing

Medical vocabularies go back to the Bills of Mortality in medieval London to manage outbreaks of the plague and other diseases (Figure 5.1).



1660.

A General BILL for this preſent Year,

Ending the 11th Day of December 1660.

According to the Report made to the King's moſt excellent Majeſty,

By the Company of Pariſh Clerks of LONDON, &c.

DISEASES and CASUALTIES.

Abortive and Stillborn	421	Flox and Small Pox	1523	Palſy	17
Aged	909	Found dead in the Streets, Fields, &c.	2	Plague	36
Ague and Fever	2303	French Pox	51	Pluriſy	12
Apoplexy and Suddenly	91	Gout	4	Quinſy and ſore Throat	21
Blaſted and Planet	3	Grief	13	Rickets	441
Bleeding and bloody Iſſue	7	Griping in the Guts	253	Riſing of the Lights	210
Bloody Flux, Scowring, and Flux	346	Hanged and made away themſelves	11	Rupture	12
Burnt and Scalded	6	Head-ach and Headmouldſhot	35	Scurvy	82
Cancer, Gangrene and Fiſtula	63	Jaundies	102	Shot	7
Canker, ſore Mouth and Thruſh	73	Impoſthume	105	Shingles	1
Childbed	226	Killed by ſeveral Accidents	55	Sores, Ulcers, broken and bruiſed Limbs	61
Chriſomes and Infants	858	King's Evil	28	Spleen	7
Cold, Cough and Hiccough	33	Lethargy	6	Spotted Fever and Purples	368
Colick and Wind	116	Livergrown	8	Starved	7
Conſumption and Tiſſick	2982	Lunatick and Frenzy	14	Strangury	22
Convulſion	742	Megrims	5	Stopping of the Stomach	186
Cut of the Stone and Stone	46	Meaſles	6	Surfeit	202
Dropſy and Tympany	646	Mother	1	Swine Pox	2
Drowned	57	Murthered	7	Teeth and Worms	839
Executed	7	Overlaid and Starved at Nurſe	46	Vomiting	8
Falling Sickneſs	4			Wen	1

Figure 5.1.: 1660 London Bill of Mortality, showing the cause of death for deceased inhabitants using a classification system of 62 diseases known at the time.

Since then, the classifications have greatly expanded in size and complexity and spread into other aspects of healthcare, such as procedures and services, drugs, medical devices, etc. The main principles have remained the same: they are controlled vocabularies, terminologies, hierarchies, or ontologies that some healthcare communities agree upon for the purpose of capturing, classifying, and analyzing patient data. Many of these vocabularies are maintained by public and government agencies with a long-term mandate for doing so. For example, the World Health Organization (WHO) produces the International Classification of Disease (ICD) with the recent addition of its 11th revision (ICD11). Local governments create country-specific versions, such as ICD10CM (USA), ICD10GM (Germany), etc. Governments also control the

marketing and sale of drugs and maintain national repositories of such certified drugs. Vocabularies are also used in the private sector, either as commercial products or for internal use, such as electronic health record (EHR) systems or for medical insurance claim reporting.

As a result, each country, region, healthcare system and institution tend to have their own classifications that would most likely only be relevant where it is used. This myriad of vocabularies prevents interoperability of the systems they are used in. Standardization is the key that enables patient data exchange, unlocks health data analysis on a global level, and allows systematic and standardized research, including performance characterization and quality assessment. To address the interoperability problem, multinational organizations have sprung up and started creating broad standards, such as the Standard Nomenclature of Medicine (SNOMED) and Logical Observation Identifiers Names and Codes (LOINC). In the US, the Health IT Standards Committee (HITAC) recommends the use of SNOMED, LOINC, and the drug vocabulary RxNorm as standards to the National Coordinator for Health IT (ONC) for use in a common platform for nationwide health information exchange across diverse entities.

OHDSI developed the OMOP CDM, a global standard for observational research. As part of the CDM, the OHDSI Standardized Vocabularies are available for two main purposes:

- Common repository of all vocabularies used in the community
- Standardization and mapping for use in research

The Standardized Vocabularies are available to the community free of charge and **must be used** for OMOP CDM instance **as its mandatory reference table**. It is crucial to use the most recent version of the Vocabularies and continuously incorporate new versions in the ETL as Vocabularies changes and shifts impact common research tasks (6).

5.1.1. Vocabularies Use Cases and Users

OHDSI Vocabularies are different from other ontology systems, such as the Unified Medical Language System (UMLS) (7) and the difference stems from the main use case of evidence generation that OHDSI supports. Both UMLS and OHDSI aggregate relationships from source vocabularies. UMLS provides

crosswalks among vocabularies with various degrees of fidelity and such crosswalks can be incomplete or ambiguous. OHDSI curates mappings and selects high-quality ones for the official "Maps to" relationships from source vocabularies to a single reference standard, ensuring that all data sources speak the same language. For example, "Atrial fibrillation" from ICD-9, Read, MeSH, etc., are all mapped to a single SNOMED concept in the Condition domain. UMLS, in contrast, groups synonymous terms under a CUI but does not designate one as "the code to use" serving as a translation table and not enforcing a single vocabulary for data encoding. UMLS contains many international vocabularies, but historically it has had a strong U.S. focus, and some content can lag in updates. OHDSI Vocabularies explicitly integrate both US and non-US coding systems and even create new standard concepts for non-US use cases to achieve global coverage. For example, US drugs are covered in RxNorm that we import, international drugs are covered in RxNorm Extension that we create de-novo and both of them are integrated with Anatomic Therapeutic Classification (ATC) (8,9). OHDSI Vocabularies are optimized for standardized analytics, offering open-access, harmonized coverage of both U.S. and international terminologies to enable consistent, reproducible studies across institutions and countries.

Vocabularies are centered around generating real-world evidence from observational studies and are mostly used for two groups of tasks: ETL of data to OMOP CDM and subsequent research on converted data. If you are a data engineer/ETL developer, the most relevant information is how to use correct source-to-standard mappings and populate both standard and source concept ID fields appropriately. Additionally, you need to know how to track vocabulary changes adopt ETL accordingly. If you are a researcher, the most relevant information is how to use vocabularies to find relevant codes for concept sets and features, use hierarchies, and examine mappings.

5.1.2. Access to the Standardized Vocabularies

The OHDSI Standardized Vocabularies are distributed via ATHENA (2), a web-based platform for browsing and downloading vocabulary data. You can use it to search, explore, and filter vocabularies by domain, concept class, vocabulary source, standard status and validity. You can select relevant vocabularies and download a pre-packaged vocabulary bundle, ready for loading into a local OMOP CDM instance.

To download a zip file with all Standardized Vocabularies tables, select all the vocabularies you need for your OMOP CDM. Vocabularies with Standard Concepts and very common usage are preselected. Add vocabularies that are used in your source data. Vocabularies that are proprietary have no select button. Click on the "License required" button to incorporate a licensed-required vocabulary into your list. The Vocabulary Team will contact you and request you demonstrate your license or help you connect to the right body to obtain one.

Each vocabulary download includes a ZIP file containing a standard set of CSV files, which can be loaded into your database using standard SQL scripts or programmatically. You will also need to re-constitute names of CPT4 codes as per our use agreement (10).

The VOCABULARY.csv file contains the version and release date metadata for each vocabulary, which should be recorded to ensure reproducibility in analyses and network studies. When updating to a newer vocabulary version, we recommend reviewing the changes in concept definitions, domain assignments, mappings, and deprecated concepts to ensure that downstream data and cohort definitions remain valid (6).

You can also select a specific vocabulary release different from the current release or download a file that contains the delta between two given releases.

5.2. Vocabularies Process and Governance

5.2.1. Building the Standardized Vocabularies and Vocabularies Principles

All vocabularies of the Standardized Vocabularies are consolidated into the same common format: `CONCEPT`, `CONCEPT_RELATIONSHIP`, `CONCEPT_ANCESTOR`, `CONCEPT_SYNONYM`, and supporting reference files such as VOCABULARY, DOMAIN, CONCEPT_CLASS, and RELATIONSHIP. This relieves the researchers from having to understand and handle multiple different formats and life-cycle conventions of the originating vocabularies.

OHDSI generally prefers adopting existing vocabularies, rather than de-novo construction, because (i) many vocabularies have already been utilized in observational data in the community, and (ii) construction and maintenance of vocabularies is complex and requires the input of many stakeholders over long periods of time to mature. For this reason, dedicated organizations provide vocabularies, which undergo a life cycle of generation, deprecation, merging, and splitting. Currently, OHDSI only produces internal administrative vocabularies like Type Concepts (for example, condition type concepts) as well as several other vocabularies to cover areas with existing gaps: RxNorm Extension to cover drugs that are only used outside the United States, OMOP Investigational Drugs for investigational drugs, Cancer Modifier for cancer measurements, and OMOP Extension for miscellaneous gaps. There are other community-driven efforts, such as GIS Vocabulary Package (11).

All vocabularies go through several common stages upon refresh: staging or harmonization to a common table structure, normalization and creation of crosswalks, integration with other vocabularies and release (12). All steps are accompanied by a set of quality assurance and control procedures, both automated and human-curated (13).

OHDSI Vocabularies follow twelve principles (14). Among others, Vocabularies focus on and support OHDSI **use case** of generating new evidence. They meant to be **comprehensive,** that is there are enough concepts to cover any event relevant to the patient's healthcare experience (e.g., conditions, procedures, exposures to drug, etc.) as well as some of the administrative information of the healthcare system (e.g., visits, care sites, etc.). They strive to have **unique standard concept**, where for each Clinical Entity there is only one concept representing it, called the Standard Concept. Other equivalent or similar concepts are designated non-Standard and mapped to the Standard ones. Moreover, such concepts should be stated as fact, no negations of facts, no reference to the past, and no flavors of null (unknown, not reported, etc.).

5.2.2. Vocabularies Governance, Roadmap and Role of the Community

OHDSI Vocabularies work and processes are governed by the OHDSI Central Coordinating Center's body, Vocabulary Committee, which includes representatives from across the OHDSI community and helps set priorities for main-

tenance, content expansion, and quality improvement. Committee's work is based on the landscape assessment conducted in 2023 (15). As a part of this work, it approves the roadmap for bi-annual releases of the Vocabularies.

Release happens in February and August and is accompanied by detailed roadmap updates and release notes describing the changes (16). Each release note describes (1) newly added vocabularies, (2) concepts and/or mappings newly added to the existing vocabularies, (4) changes in mappings, domains, status of the concepts as well as detailed description of the actions performed for specific vocabularies. Additionally, the release notes contain the artifacts not available through Athena: pack content, SSSOM-compatible metadata (17) for concepts and relationships as well as reports for alignment with vocabulary principles (August 2024 release).

You can use information about releases and the roadmap in two ways. First, you can assess the content of each release and use open source tools to assess its impact on ETL and research (6,18). If you are a ETLer or are responsible for your institution's data/OMOP CDM, **you should update OMOP CDM instance with the latest version of the Vocabularies** to benefit from improved coverage, consistency, and corrected mappings. Vocabularies **change** a lot. If you are not updating, you are falling behind in research.

Second, you can use this information to assess if planned activities meet your needs. You assist in vocabulary maintenance if your vocabulary is not on the roadmap as a vocabulary steward (19). With each release stewards from the community refresh and improve their vocabularies, even if they are not on the roadmap (current list of stewards can be found here (4)). You can also add your vocabulary, concepts or improve existing content (mappings, domains) through community contribution (5).

5.2.2.1. Community contributions

The extensive scope of the OMOP Standard Vocabularies poses a challenge to maintenance and scalability. The OHDSI Vocabulary Team focuses on core terminologies out of necessity. However, there are plenty of opportunities for the OHDSI community to assist in vocabulary maintenance. Examples above regarding improvements in mappings, labels, synonyms, etc. are welcomed from the community. Working Groups may feel particular ownership of a domain or specialty area and wish to help manage the necessary vocabulary. When this

happens, the core terminology management system can be extended through integration of community-contributed content. Begun in earnest in 2024, this community and centrally-management vocabulary integration allows for more scalable contribution and more rapid conceptual gap-filling. A community-contribution infrastructure has been developed in phases depending on the complexity of the contribution. Small changes or additions can be provided using templates.

You can use templates to add a new standard vocabulary, add non-standard concepts, add mappings, change mappings or concept domains, or propose upgrading a non-standard concept to standard. Modification of content requires community approval through the Vocabulary Workgroup. Template submissions should be completed and ratified two months prior to the release date (end of June/end of December for August and February releases, respectively). Instructions for completed templates are described on the GitHub Wiki (5).

Larger contributions (for example, entire terminologies or drug catalogues) require staging and integration using a compatible environment to that used for managing the core terminologies. Examples of community-staged contributions include the Heme-Onc vocabulary, the Veterinary vocabulary and the CIEL terminology. For complex contributions, it is best to have a working group sponsor your request. You can use the instructions provided on Wiki under Community Contributions Part II (5). We recommend you talk to the members of the Vocabulary Workgroup or Team to discuss your specific use case.

5.3. OHDSI Vocabularies Structure: Concepts and Relationships

All clinical events in the OMOP CDM are represented as concepts, which capture the semantic notion of each event. They are the fundamental building blocks of the data records, making almost all tables fully normalized with few exceptions. Concepts are stored in the `CONCEPT` table (Figure 5.2).

CONCEPT_ID	313217
CONCEPT_NAME	Atrial fibrillation
DOMAIN_ID	Condition
VOCABULARY_ID	SNOMED
CONCEPT_CLASS_ID	Clinical Finding
STANDARD_CONCEPT	S
CONCEPT_CODE	49436004
VALID_START_DATE	01-Jan-1970
VALID_END_DATE	31-Dec-2099
INVALID_REASON	

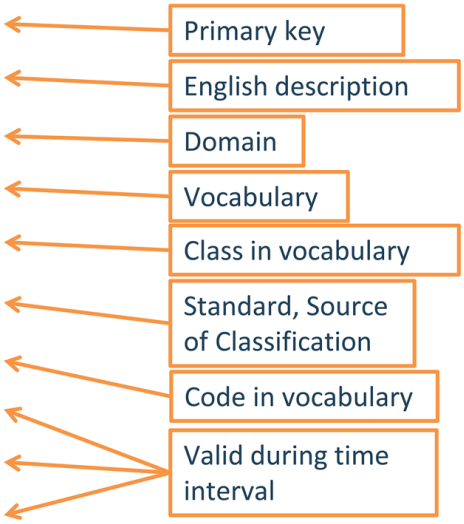


Figure 5.2.: Standard representation of vocabulary concepts in the OMOP CDM. The example provided is the `CONCEPT` table record for the SNOMED code for Atrial Fibrillation.

5.3.1. Concept IDs

Each concept is assigned a concept ID to be used as a primary key. This meaningless integer ID, rather than the original code from the source vocabulary, is used to record data in the CDM event tables via the foreign key fields. No two concepts (even from different vocabularies) share the same ID. Conversely, the same source code might appear in multiple vocabularies, but each distinct concept gets its own ID.

5.3.2. Concept Names

Each concept has one name. Names are always in English. They are imported from the source of the vocabulary. If the source vocabulary has more than one name, the most expressive (fully specified) is selected and the remaining ones are stored in the `CONCEPT_SYNONYM` table under the same `CONCEPT_ID` key. Non-English names are recorded in `CONCEPT_SYNONYM` as well, with the appropriate language concept ID in the `LANGUAGE_CONCEPT_ID` field. The

name can only be 255 characters long, which means that very long names get truncated, and the full-length version recorded as another synonym, which can hold up to 1000 characters. Tools like Athena and ATLAS use the concept names and synonyms to let users search for concepts. When doing analysis, it is often convenient to have the concept names for interpretability, but analysis logic should use the `CONCEPT_ID`.

5.3.3. Domains

Each concept is assigned a domain in the `DOMAIN_ID` field, which, in contrast to the numerical `CONCEPT_ID`, is a short, case-sensitive, unique alphanumeric ID for the domain. Domains are OMOP-specific and correspond to the OMOP CDM tables (20). Examples of such identifiers are "Condition," "Drug," "Procedure," "Visit," "Device," "Specimen," etc. Domains also direct to which CDM table and field a clinical event or event attribute is recorded. For example, "Atrial fibrillation" is a clinical finding that would be recorded in the Condition Occurrence table, so its domain is "Condition"; a concept for a lab test (for example, "Blood glucose measurement") would have domain "Measurement" and belong in the Measurement table. Domains are assigned to codes, and a vocabulary can have different domains: for example, HCPCS, while considered procedure vocabulary, also has codes with Drug and Observation domains.

The domain heuristic follows the definitions of the domains. These definitions are derived from the table and field definitions in the CDM **{Chapter 4}**. The heuristic is not perfect; there are grey zones (**{Section 5.4}** "Special Situations"), source vocabulary shifts, and occasional misassignments. Although domains of concepts may change, 95% of the concepts never changed their domain since Vocabularies' inception (for more information, see Assets in v20240830 release notes) (16).

5.3.4. Vocabularies

Each vocabulary has a short case-sensitive unique alphanumeric ID, which generally follows the abbreviated name of the vocabulary, omitting dashes. For example, ICD-9-CM has the vocabulary ID "ICD9CM". As of 2025, over 140 vocabularies are available through ATHENA and follow different cadence

of updates. The source and the version of the vocabularies is defined in the VOCABULARY reference file and documentation for individual vocabularies can be found on GitHub (4,16).

5.3.5. Concept Classes

Some vocabularies classify their codes or concepts, denoted through their case-sensitive unique alphanumerical IDs. For example, SNOMED has 33 such concept classes, which SNOMED refers to as "semantic tags": clinical finding, social context, body structure, etc. These are vertical divisions of the concepts. Others, such as MedDRA or RxNorm, have concept classes classifying horizontal levels in their stratified hierarchies. Vocabularies without any concept classes, such as HCPCS, use the vocabulary ID as the Concept Class ID.

Table 5.1.: Vocabularies with or without horizontal and vertical subclassification principles in concept class.

Concept class subdivision principle	Vocabulary
Horizontal	All drug vocabularies, CD , Episode, HCPCS, HemOnc, ICDs, MedDRA, OSM, Census
Vertical	CIEL, HES Specialty, ICDO3, MeSH, NAACCR, NDFRT, OPCS4, PCORNET, Plan, PPI, Provider, SNOMED, SPL, UCUM
Mixed	CPT4, ISBT, LOINC
None	OXMIS, Race, Revenue Code, Sponsor, Supplier, UB04s, Visit

Horizontal concept classes allow you to determine a specific hierarchical level. For example, in the drug vocabulary RxNorm, the concept class "Ingredient" defines the top level of the hierarchy. In the vertical model, members of a concept class can be of any hierarchical level from the top to the very bottom. Concept class is mostly a descriptive attribute and helps to filter concepts. For example, if you only want to select drugs with a specific Brand Name you can filter to "Branded Drug" class.

5.3.6. Standard Concepts

A Standard Concept is the community-endorsed, canonical representation of a clinical meaning within the OHDSI Vocabularies. It serves as the unified semantic identifier for a specific entity (for example, condition, drug, procedure), regardless of how that entity is expressed in source vocabularies. Only Standard Concepts are used to populate the `CONCEPT_ID` fields in the CDM, ensuring consistency across diverse datasets. Standard concepts serve as the target for mappings. For each clinical entity, one concept from one vocabulary is chosen to be standard. This becomes the "hub" to which all equivalent source codes are mapped. For example, MESH code D001281, CIEL code 148203, SNOMED code 49436004, ICD9CM code 427.31 and Read code G573000 all define "Atrial fibrillation" in the condition domain, but only the SNOMED concept is Standard and represents the condition in the data. The others are designated non-standard or source concepts and mapped to the Standard ones. Standard Concepts are indicated through an "S" in the `STANDARD_CONCEPT` field. And only these Standard Concepts are used to record data in the CDM fields ending in `_CONCEPT_ID`.

We rely on well-known reference terminologies for standard terms: SNOMED CT for conditions, RxNorm and RxNorm Extension for drugs, LOINC and SNOMED for measurements, etc. Not all concepts in those vocabularies are necessarily standard. Occasionally, a concept in a standard vocabulary might be deemed out of scope or duplicative and not used. Conversely, some concepts from typically non-standard vocabularies might be made standard if no better alternative exists.

While we strive to align with the unique standard concept principle to have one Standard concept per semantic entity, duplicates exist. For example, no deduplication of standard concepts has been performed for the Device domain. While concept mappings avoid direct collisions, this dual-standard condition can introduce ambiguity in cohort definition, concept set construction, and analytic interpretation. This phenomenon is not a flaw but rather a reflection of ontology convergence in progress, where two high-quality terminologies independently arrive at comparable representations of the same clinical reality. OHDSI addresses these cases through community review, classification logic, and long-term efforts toward consolidation via concept deprecation, reclassification, or updated mappings. Until then, such duplications must be handled with care in concept set design and ETL strategies.

5.3.7. Non-Standard Concepts

Non-standard concepts are not used in standardized analytics, but they are still part of the Standardized Vocabularies and are often found in the source data. For that reason, they are also called "source concepts". The conversion of source concepts to Standard Concepts is a process called "mapping".

Some of the non-standard concepts cannot be mapped and are not suitable for analytic use. Examples of such include terms like "Not reported", "Not specified", "Passport number" and more.

5.3.8. Classification Concepts

These concepts are not Standard and hence cannot be used to represent the data. But they are participating in the hierarchy with the Standard Concepts and can therefore be used to perform hierarchical queries. For example, querying for all descendants of ATC code prednisolone;systemic will retrieve the Standard RxNorm concept for prednisolone 5 MG Oral Tablet (Figure 5.3).

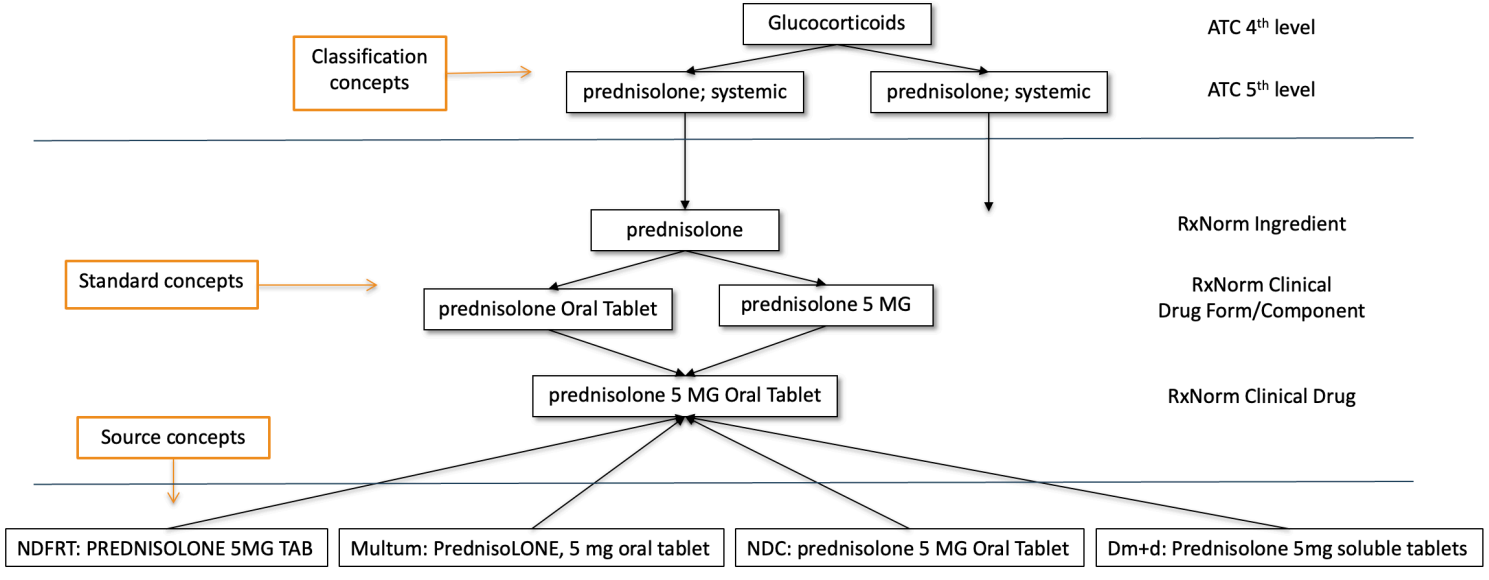


Figure 5.3.: Standard, non-standard source and classification concepts and their hierarchical relationships in the drug domain.

Classification concepts are marked with a "C" in the `STANDARD_CONCEPT` field. Most classification concepts form a hierarchy along with the standard concepts, and these relationships are stored in the `CONCEPT_ANCESTOR` table.

Classification concepts are vital in enabling concept set expansion via ancestry traversal. While they cannot be used to populate clinical event tables directly, they serve as entry points into clinically meaningful groupings (for example, drug classes or disorder categories). They are especially powerful when used in cohort definitions or phenotype algorithms that require aggregation of clinically related Standard Concepts. For example, selecting the ATC classification concept C09AA ACE inhibitors will retrieve all Standard RxNorm ingredients and products mapped as descendants.

Quality and coverage of classification hierarchies vary across domains. Currently, Drug and Condition domains have mature classification structures (for example, ATC, MedDRA), while Procedure, Measurement, and Device domains lack formal classification vocabularies. Caution should be exercised when interpreting classification-derived hierarchies, as they may not always reflect clinical practice or data granularity.

5.2.8.1 Standard/non-standard/classification Concept Designation

The choice of concept designation as Standard, non-standard, and classification is typically done for each domain separately at the vocabulary level. This is based on the quality of the concepts, the built-in hierarchy, and the declared purpose of the vocabulary. Also, not all concepts in a vocabulary are used as Standard Concepts. The designation is separate for each domain, each concept must be active, and there might be an order of precedence if more than one concept from different vocabularies compete for the same meaning. See Table 5.2 for examples.

Table 5.2.: List of vocabularies to utilize for Standard/non-standard/classification concept assignments.

Domain	for Standard Concepts	for source concepts	for classification concepts
Condition	SNOMED, ICDO3	SNOMED Veterinary	MedDRA
Procedure	SNOMED, CPT4, HCPCS, ICD10PCS, ICD9Proc, OPCS4	SNOMED Veterinary, HemOnc, NAACCR	None at this point
Measurement	SNOMED, LOINC	SNOMED Veterinary, NAACCR, CPT4, HCPCS, OPCS4, PPI	None at this point
Drug	RxNorm, RxNorm Extension, CVX	HCPCS, CPT4, HemOnc, NAACCR	ATC
Device	SNOMED	Others, currently not normalized	None at this point
Observation	SNOMED	Others	None at this point
Visit	CMS Place of Service, ABMT, NUCC	SNOMED, HCPCS, CPT4, UB04	None at this point

5.3.9. Concept Codes

Concept codes are the identifiers used in the source vocabularies. For example, ICD9CM or NDC codes are stored in this field, while the OMOP tables use the concept ID as a foreign key into the `CONCEPT` table. The reason is that the name space overlaps across vocabularies, that is the same code can exist in different vocabularies with completely different meanings (Table 5.3).

Table 5.3.: Concepts with identical concept code 1001, but different vocabularies, domains and concept classes.

Concept ID	Concept Code	ConceptName	DomainID	VocabularyID	ConceptClass
35803438	1001	Granulocyte colony-stimulating factors	Drug	HemOnc	Component Class
35942070	1001	AJCC TNM Clin T	Measurement	NAACCR	NAACCR Variable
1036059	1001	Antipyrine	Drug	RxNorm	Ingredient
38003544	1001	Residential Treatment - Psychiatric	Revenue Code	Revenue Code	Revenue Code
43228317	1001	Aceprometazine maleate	Drug	BDPM	Ingredient
45417187	1001	Brompheniramine Maleate, 10 mg/mL injectable solution	Drug	Multum	Multum
45912144	1001	Serum	Specimen	CIEL	Specimen

CONCEPT_CODE is unique only within a given vocabulary. You should not join datasets via CONCEPT_CODE unless constrained by VOCABULARY_ID.

In addition, certain vocabularies, such as HCPCS, NDC, and DRG are known to reuse codes over time, assigning new meanings to previously used codes. In such cases, Vocabularies differentiate concepts based on validity dates (VALID_START_DATE, VALID_END_DATE) and keep the most recent meaning.

Some OMOP-specific vocabularies (for example, Type Concept, Visit) contain system-generated concept codes rather than real-world codes. Finally, certain source vocabularies (such as ATC or hierarchical clinical classifications) embed structural hierarchy into their codes (ATC G03E vs. G03EK), meaning that not all CONCEPT_CODE matches imply equivalence at the clinical level.

5.3.10. Lifecycle

Vocabularies are rarely permanent corpora with a fixed set of codes. Instead, codes and concepts are added and get deprecated. The OMOP CDM is a model to support longitudinal patient data, which means it needs to support concepts that were used in the past and might no longer be active, as well as supporting new concepts and placing them into context. There are three fields in the CONCEPT table that describe the possible life-cycle statuses: VALID_START_DATE, VALID_END_DATE, and INVALID_REASON. Their values differ depending on the concept life-cycle status:

- **Active or new concept**
  - Description: Concept in use.
  - VALID_START_DATE: Day of instantiation of concept; if that is not known, day of incorporation of concept in Vocabularies; if that is not known, 1970-1-1.
  - VALID_END_DATE: Set to 2099-12-31 as a convention to indicate "Might become invalid in an undefined future, but active right now".
  - INVALID_REASON: NULL
- **Deprecated Concept with no successor**
  - Description: Concept inactive and cannot be used as Standard.

  - `VALID_START_DATE`: Day of instantiation of concept; if that is not known, day of incorporation of concept in Vocabularies; if that is not known, 1970-1-1.
  - `VALID_END_DATE`: Day in the past indicating deprecation, or if that is not known, day of vocabulary refresh where concept in vocabulary went missing or set to inactive.
  - `INVALID_REASON`: "D"

- **Upgraded Concept with successor**
  - Description: Concept inactive but has defined successor. These are typically concepts which went through de-duplication.
  - `VALID_START_DATE`: Day of instantiation of concept; if that is not known, day of incorporation of concept in Vocabularies; if that is not known, 1970-1-1.
  - `VALID_END_DATE`: Day in the past indicating an upgrade, or if that is not known day of vocabulary refresh where the upgrade was included.
  - `INVALID_REASON`: "U"

- **Reused code for another new concept**
  - Description: The vocabulary reused the concept code of this deprecated concept for a new concept.
  - `VALID_START_DATE`: Day of instantiation of concept; if that is not known, day of incorporation of concept in Vocabularies; if that is not known, 1970-1-1.
  - `VALID_END_DATE`: Day in the past indicating deprecation, or if that is not known day of vocabulary refresh where concept in vocabulary went missing or set to inactive.

In addition to concept lifecycle management, the `CONCEPT_RELATIONSHIP` table also has lifecycle fields (`VALID_START_DATE`, `VALID_END_DATE`, `INVALID_REASON`) for relationships. Relationships may change over time independently of the concepts themselves. While all relationships are versioned in the internal vocabulary system, only active mappings are included in Athena

downloads. Every OMOP CDM instance should record the vocabulary version (stored in the VOCABULARY table) used at ETL time to ensure transparency and reproducibility. Lifecycle management principles apply equally to custom extensions and community-contributed vocabularies: all new concepts and mappings must carry valid VALID_START_DATE entries and, when deprecated, clearly marked VALID_END_DATE and INVALID_REASON values.

5.3.11. Relationships

Any two concepts can have a defined relationship, regardless of whether the two concepts belong to the same domain or vocabulary. The nature of the relationships is indicated in its short case-sensitive unique alphanumeric ID in the RELATIONSHIP_ID field of the CONCEPT_RELATIONSHIP table. Relationships are symmetrical, that is for each relationship an equivalent relationship exists, where the content of the fields CONCEPT_ID_1 and CONCEPT_ID_2 are swapped, and the RELATIONHSIP_ID is changed to its opposite. For example, the "Maps to" relationship has an opposite relationship "Mapped from." Different types of relationships serve different analytic purposes. "Maps to" and "Mapped from" support source-to-standard mappings. "Is a" and "Subsumes" define hierarchical subclass relationships. "Has ingredient" and "Ingredient of" structure drug compositions. "Concept replaced by" and "Concept replaces" handle lifecycle transitions across deprecated content.

As stated in the previous section, CONCEPT_RELATIONSHIP table records also have life-cycle fields VALID_START_DATE, VALID_END_DATE and INVALID_REASON. However, only active records with INVALID_REASON = NULL are available through ATHENA. Inactive relationships are kept for internal processing only.

The RELATIONSHIP table serves as the reference with the full list of relationship IDs and their reverse counterparts. It also specifies two important flags: DEFINES_ANCESTRY, indicating whether a relationship should contribute to the CONCEPT_ANCESTOR table, and IS_HIERARCHICAL, indicating whether the relationship encodes a subsumption hierarchy. Not all relationships define ancestry; only those intended to build domain hierarchies (for example, "Is a") are used to populate CONCEPT_ANCESTOR. It is essential to distinguish between direct relationships (stored in CONCEPT_RELATIONSHIP) and inferred multi-level hierarchies (precomputed and stored in CONCEPT_ANCESTOR), especially when

writing concept set queries, building phenotypes, or exploring ontology structures.

5.3.12. Mapping Relationships

These relationships provide translations from non-standard to Standard concepts, supported by two relationship ID pairs (Table 5.4).

Table 5.4.: Type of mapping relationships.

Relationship ID pair	Purpose
"Maps to" and "Mapped from"	Mapping to Standard Concepts. Standard Concepts are mapped to themselves, non-standard concepts to Standard Concepts. Most non-standard and all Standard Concepts have this relationship to a Standard Concept. The former are stored in *_SOURCE_CONCEPT_ID, and the latter in the *_CONCEPT_ID fields. Classification concepts are not mapped.
"Maps to value" and "Value mapped from"	Mapping to a concept that represents a Value to be placed into the VALUE_AS_CONCEPT_ID fields of the MEASUREMENT and OBSERVATION tables.

The purpose of these mapping relationships is to allow a crosswalk between equivalent concepts to harmonize how clinical events are represented in the OMOP CDM. This is a main achievement of the Standardized Vocabularies.

"Equivalent concepts" means it carries the same meaning, and, importantly, the hierarchical descendants cover the same semantic space. If an equivalent concept is not available and the concept is not Standard, it is still mapped, but to a slightly broader concept (so-called "up-hill mappings" or semantic subsumption). For example, ICD10CM W61.51 "Bitten by goose" has no equivalent in the SNOMED vocabulary, which is generally used for standard condition concepts. Instead, it is mapped to SNOMED 217716004 "Peck by bird," losing the context of the bird being a goose. Up-hill mappings are only

used if the loss of information is considered irrelevant to standard research use cases.

Some mappings connect a source concept to more than one Standard Concept. For example, ICD9CM 070.43 "Hepatitis E with hepatic coma" is mapped to both SNOMED 235867002 "Acute hepatitis E" as well as SNOMED 72836002 "Hepatic Coma." The reason for this is that the original source concept is a pre-coordinated combination of two conditions, hepatitis and coma, meaning that a single code simultaneously encodes multiple clinical ideas rather than expressing them separately SNOMED does not have that combination, which results in two records written to the `CONDITION_OCCURRENCE` table for the single ICD9CM record, one with each mapped Standard Concept.

Relationships "Maps to value" have the purpose of splitting of a value for OMOP CDM tables following an entity-attribute-value (EAV) model (21). This is typically the case in the following situations:

- Measurements consisting of a test and a result value
- Personal or family disease history
- Allergy to substance
- Need for immunization

In these situations, the source concept is a combination of the attribute (test or history) and the value (test result or disease). The "Maps to" relationship maps this source to the attribute concept, and the "Maps to value" to the value concept. See Figure 5.4 for an example.

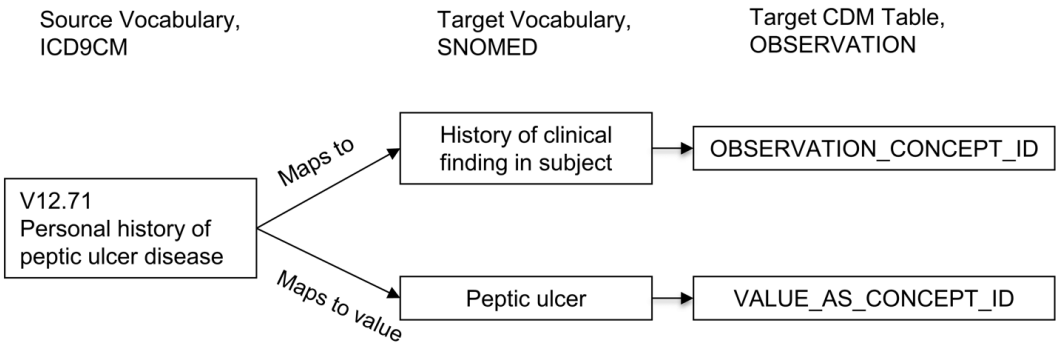


Figure 5.4.: One-to-many mapping between source concept and Standard Concepts. A pre-coordinated concept is split into two concepts, one of which is the attribute (here history of clinical finding) and the other one is the value (peptic ulcer). While "Maps to" relationship will map to concepts of the measurement or observation domains, the "Maps to value" concepts have no domain restriction.

This process represents a form of controlled **post-coordination** within OMOP vocabularies: instead of encoding every possible combination as a new standard concept, the meaning is decomposed into two (or more) standardized elements that together fully represent the clinical event. Together, they enable more flexible, semantically rich, and extensible data modeling. By post-coordinating attribute and value concepts, OHDSI Standardized Vocabularies avoid uncontrolled growth in the number of concepts while still allowing detailed, clinically meaningful data representation and analysis. Analysts must retrieve both the `CONCEPT_ID` and `VALUE_AS_CONCEPT_ID` fields together during query building to reconstruct the complete meaning.

Mapping relationships themselves are subject to lifecycle management. Deprecated mappings (mappings with an `INVALID_REASON` other than NULL) are removed from active ATHENA releases but can impact longitudinal data or historical cohort definitions if not updated. Careful management of mapping versioning is crucial during vocabulary refresh cycles.

When interpreting mappings, users must be aware that not all source-to-standard mappings imply perfect semantic equivalence. Slight loss of detail, context shift, or broader aggregation may occur, particularly in uphill mappings or when representing pre-coordinated concepts. Analysts and ETL designers should validate mappings in critical analytic contexts.

5.3.13. Hierarchical Relationships and Hierarchy

Relationships which indicate a hierarchy are defined through the "Is a" - "Subsumes" relationship pair. Hierarchical relationships are defined such that the child concept has all the attributes of the parent concept, plus one or more additional attributes or a more precisely defined attribute. For example, SNOMED 49436004 "Atrial fibrillation" is related to SNOMED 17366009 "Atrial arrhythmia" through a "Is a" relationship. Both concepts have an identical set of attributes except the type of arrhythmia, which is defined as fibrillation in one but not the other. Concepts can have more than one parent and more than one child concept. In this example, SNOMED 49436004 "Atrial fibrillation" is also an "Is a" to SNOMED 40593004 "Fibrillation."

Within a domain, and in some cases across domains, standard and classification concepts are organized in a hierarchical structure and stored in the `CONCEPT_ANCESTOR` table. This allows querying and retrieving concepts and all their hierarchical descendants. These descendants have the same attributes as their ancestor, but also additional or more defined ones.

The `CONCEPT_ANCESTOR` table is built automatically from the `CONCEPT_RELATIONSHIP` table, traversing all possible concepts connected through hierarchical relationships. These are the "Is a" - "Subsumes" pairs (Figure 5.5), and other relationships connecting hierarchies across vocabularies ("SNOMED - CPT4 equivalent", "RxNorm ingredient of"). The choice whether a relationship participates in the hierarchy constructor is defined for each relationship ID by the flag `DEFINES_ANCESTRY` in the `RELATIONSHIP` reference table. It is important to note that not all relationships with hierarchical meaning (`IS_HIERARCHICAL` = 1) are used for ancestry building; only those with `DEFINES_ANCESTRY` = 1 contribute to `CONCEPT_ANCESTOR`. Relationships such as "Has FDA approved indication" or "Consists of" are conceptually hierarchical but are excluded from ancestry paths to preserve clinical rigor.

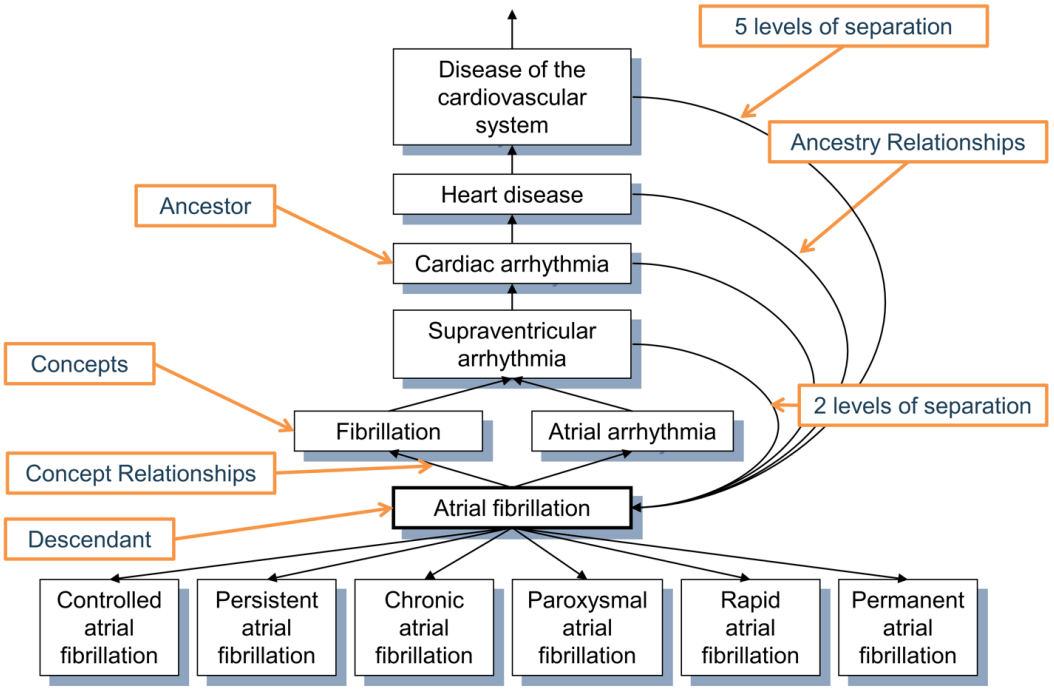


Figure 5.5.: Hierarchy of the condition "Atrial fibrillation". First degree ancestry is defined through "Is a" and "Subsumes" relationships, while all higher degree relations are inferred and stored in the `CONCEPT_ANCESTOR` table. Each concept is also its own descendant with both levels of separation equal to 0.

The ancestral degree, or the number of steps between ancestor and descendant, is captured in the `MIN_LEVELS_OF_SEPARATION` and `MAX_LEVELS_OF_SEPARATION` fields, defining the shortest or longest possible connection. Not all hierarchical relationships contribute equally to the levels-of-separation calculation. A step counted for the degree is determined by the `IS_HIERARCHICAL` flag in the `RELATIONSHIP` reference table for each relationship ID.

As of 2025, a high-quality comprehensive hierarchy exists only for two domains: Drug and Condition. Procedure, Measurement, and Observation domains are only partially covered and in the process of construction. The ancestry is particularly useful for the drug domain as it allows browsing all drugs with a given ingredient or members of drug classes irrespective of the country of origin, brand name or other attributes.

Users should also be aware that vocabulary updates can introduce changes to hierarchical structures, as relationships may be added, modified, or deprecated over time. Therefore, researchers are strongly encouraged to version-control their vocabulary snapshot to preserve analytic reproducibility.

5.3.14. Other Relationships

Relationships between two different vocabularies other than mapping and hierarchy relationships are typically of the type "Vocabulary A - Vocabulary B equivalent", which is either supplied by the original source of the vocabulary or is built de-novo. They may serve as approximate mappings but often are less precise than the better curated mapping relationships. High-quality equivalence relationships (such as "Source - RxNorm equivalent") are always duplicated by "Maps to" relationship.

Internal vocabulary relationships are usually supplied by the vocabulary provider and their quality highly depends on the vocabulary. Many of these define relationships between clinical events and can be used for information retrieval. For example, disorders of the urethra can be found by following the "Finding site of" relationship (Table 5.5):

Table 5.5.: "Finding site of" relationship of the "Urethra," indicating conditions that are situated all in this anatomical structure.

CONCEPT_ID_1	CONCEPT_ID_2
4000504 "Urethra part"	36713433 "Partial duplication of urethra"
4000504 "Urethra part"	433583 "Epispadias"
4000504 "Urethra part"	443533 "Epispadias, male"
4000504 "Urethra part"	4005956 "Epispadias, female"

Internal relationships within a vocabulary may represent hierarchical (for example, "Is a", "RxNorm ingredient of") connections or non-hierarchical semantic associations such as anatomical location, causative agent, or associated morphology. For example, within RxNorm, relationships like "Precise ingredient of" and "Has precise ingredient" enable navigation between drug products and their precise ingredients.

5.4. Special Situations

5.4.1. Device Coding

Device concepts have no standardized coding scheme that could be used to source Standard Concepts. In many source data, devices are not even coded or contained in an external coding scheme. For this same reason, there is currently no hierarchical system available. External standards like GMDN and FDA's UDI database have been considered but are not yet integrated. As a result, device concepts in OHDSI are mostly standard, same devices have multiple standard concepts across different vocabularies and there is no hierarchy to group terms. If you need help with devices or want to contribute talk to the OHDSI Device Workgroup and refer to **{Chapter 7}** of this book.

5.4.2. Coding in Oncology

Cancer data present unique modeling challenges due to the complexity of diagnoses, staging, histology, metastasis, genomic features, and treatment pathways. Please refer to the OHDSI Oncology Workgroup to learn more about conventions.

There are several mapping principles we want to cover in this chapter:

- Primary cancer diagnoses are mapped to Condition domain concepts, mostly to SNOMED CT. ICDO-3 terms are used where SNOMED coverage is insufficient.

Tumor staging, grading, and metastasis details are captured using the specialized Cancer Modifier vocabulary, which encodes structured AJCC/UICC-based elements. Mappings in Cancer Modifier are designed to ensure that cancer-related data: (1) preserve key clinical distinctions (for example, metastatic vs. localized disease), (2) support longitudinal cohort definitions (for example, new diagnosis vs. recurrence), (3) enable harmonized analytics across registries, EHRs, and claims data.

- Genomic abnormalities, when available, are mapped to concepts in the OMOP Genomic vocabulary.

- Oncology-specific measurements and observations, such as tumor dimensions or metastasis spread, often use post-coordination approaches - representing the entity and its result separately - to align with OMOP's Measurement/Observation model.
- Chemotherapy regimens are represented using the HemOnc vocabulary, while individual oncology drugs are mapped via RxNorm/RxNorm Extension.

More work is needed to refine mappings, remove duplicates, expand support for hematologic malignancies, and integrate molecular/genomic features.

5.4.3. Coding in Psychiatry

Psychiatric and neuropsychiatric data pose unique challenges for standardization due to the complexity of symptoms, variability of assessment tools, and evolving diagnostic frameworks. If you interested in this research talk to the OHDSI Psychiatry Workgroup.

In the OMOP model, psychiatric assessments are primarily captured within the Measurement and Observation domains, depending on whether the recorded information reflects a quantitative value or a qualitative clinical finding. Workgroup works on integrating and harmonizing Neuropsychiatric Assessment Tools, which include standardized psychometric scales, questionnaires, and structured interviews, into the Vocabularies, deduplicating terms and developing a hierarchy based on SNOMED structure to connect measurements to clinical concepts. They consider using Thesaurus of Psychological Index Terms and Human Phenotype Ontology (HPO), and real-world datasets (for example, MIMIC-IV) to inform this integration.

5.4.4. Coding for GIS, Exposomes and SDOH

Environmental context, exposomes, geographic location, and social conditions are not represented well in the OHDSI Vocabularies. If you are interested in research, talk to the OHDSI GIS Workgroup. One of the outputs of group is the OMOP GIS Vocabulary Package, which (22) delivers three coordinated vocabularies: OMOP GIS for geographic units and spatial relations, OMOP

Exposome for chemicals, pollutants, toxins, and their biological targets, and OMOP SDOH for structured social-determinant indicators.

To accommodate these concepts, the package adds new domain identifiers such as Geographic Feature, Environmental Feature, Socioeconomic Feature, and Behavioral Feature. Unlike the classical OMOP domains - essentially routing flags that direct ETL to a specific CDM table - these new domains act solely as semantic groupers. They organize concepts into coherent knowledge families without prescribing storage location. Events encoded with these concepts are still recorded in the appropriate CDM tables such as `EXTERNAL_EXPOSURE`, `OBSERVATION`, or `MEASUREMENT` following existing conventions.

5.4.5. Microbiology and Susceptibility Coding

There are no comprehensive conventions for microbiology coding in OHDSI. You should refer to Themis conventions for the up-to-date guidance. Generally, the most common scenarios involve (1) specimen collection with a single diagnostic result (for example, Gram stain), (2) multiple lab procedures on a single sample, (3) culture-negative findings, and (4) one or more organisms identified and tested against antibiotics.

OMOP CDM supports this complexity through the `MEASUREMENT`, `OBSERVATION`, and `SPECIMEN` tables, with event linkages (*_EVENT_ID) connecting susceptibility results to organisms and organisms to specimens. Antibiotic susceptibility results are typically stored as LOINC-coded MEASUREMENTs with quantitative values (for example, MIC) and qualitative interpretations (for example, sensitive). When coding microbiology data you should use standard concepts from Measurement domain to populate `MEASUREMENT_CONCEPT_ID` (such as susceptibility test) and Meas Value domain to populate `VALUE_AS_CONCEPT_ID` (such as detected/not detected).

5.4.6. Survey Coding

There are no comprehensive conventions for survey coding in OHDSI. You should refer to Survey Workgroup for the up-to-date guidance. Broadly, surveys can be stored as Question-Answer pairs (separate concepts) or as pre-coordinated Question-Answer (one concept). Existing survey vocabularies,

such as PPI and UK Biobank, are a mix of both. Surveys added to the Vocabularies generally should follow broad Vocabularies principles. For example, they should not contain negative information and flavors of null (not reported, not specified, etc.). If they have codes that already have standard counterparts in the Vocabularies, they should be mapped appropriately. If you want to add your survey instrument, please talk to the Survey Workgroup.

5.4.7. Flavors of NULL

Many vocabularies contain codes that represent some form of absence of information. For example, of the five gender concepts 8507 "Male," 8532 "Female," 8570 "Ambiguous," 8551 "Unknown," and 8521 "Other", only the first two are Standard, and the other three are source concepts with no mapping. In the Standardized Vocabularies, there is intentionally no distinction why a piece of information is not available; it might be because of an active withdrawal of information by the patient, a missing value, a value that is not defined or standardized in some way, or the absence of a mapping record in `CONCEPT_RELATIONSHIP`. Any such concept is not mapped, which corresponds to a default mapping to the Standard Concept with the concept ID = 0.

As per Vocabularies' principles we avoid adding new flavors of NULL to the Vocabularies and advise against using such concepts in research.

5.5. Summary

- All events and administrative facts are represented in the OHDSI Standardized Vocabularies as concepts and concept relationships.
- Most of these are adopted from existing coding schemes or vocabularies, while others are either extended (for example, RxNorm Extension, OMOP Extension) or developed de novo by OHDSI Vocabulary Team or community to cover missing areas.
- All concepts are assigned a domain, which controls where the fact represented by the concept is stored in the CDM.

- Concepts of equivalent meaning in different vocabularies are mapped to one of them, which is designated the Standard Concept. The others are source concepts. Standard concepts ("S") are the only concepts used in analytical fields.

- We strive for collaborative and transparent Vocabularies with most of the documentation located on OHDSI Vocabularies GitHub Wiki. You can get involved as a community contributor or vocabulary steward. You can contribute simple content through templates or more complex content though programmatic vocabulary development.

References

1. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI Standardized Vocabularies-a large-scale centralized reference ontology for international data harmonization. J Am Med Inform Assoc. 2024 Feb 16;31(3):583–90.

2. Athena [Internet]. [cited 2025 May 23]. Available from: https://athena.ohdsi.org/sear terms/start

3. Release planning · OHDSI/Vocabulary-v5.0 Wiki · GitHub [Internet]. [cited 2025 May 23]. Available from: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Release-planning

4. Standardized Vocabularies · OHDSI/Vocabulary-v5.0 Wiki · GitHub [Internet]. [cited 2025 May 23]. Available from: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Standardized-Vocabularies

5. Community contribution · OHDSI/Vocabulary-v5.0 Wiki · GitHub [Internet]. [cited 2025 May 23]. Available from: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Community-contribution

6. Dymshyts D. Evaluating the impact of different vocabulary versions on cohort definitions and CDM. In 2024 [cited 2025 May 23]. Available from: https://www.ohdsi.org/wp-content/uploads/2024/10/23-EvaluationConceptSets_Ddymshyts_2024_US-Dmitry-Dymshyts.pdf

7. Amos L, Anderson D, Brody S, Ripple A, Humphreys BL. UMLS users and uses: a current overview. Journal of the American Medical Informatics Association. 2020 Oct 1;27(10):1606–11.

8. De Groot R, Glaser S, Kogan A, Medlock S, Alloni A, Gabetta M, et al. ATC-to-RxNorm mappings – A comparison between OHDSI Standardized Vocabularies and UMLS Metathesaurus. International Journal of Medical Informatics. 2025 Mar;195:105777.

9. A High-Fidelity Combined ATC-Rxnorm Drug Hierarchy for Large-Scale Observational Research. In: Studies in Health Technology and Informatics [Internet]. IOS Press; 2024 [cited 2025 May 23]. Available from: https://ebooks.iospress.nl/doi/10.3233/SHTI230926

10. General Structure, Download and Use · OHDSI/Vocabulary-v5.0 Wiki · GitHub [Internet]. [cited 2025 May 23]. Available from: https://github.com/OHDSI/Vocabulary-v5.0/wiki/General-Structure,-Download-and-Use

11. Trofymenko M, Talapova P, Williams A. OMOP GIS Vocabulary Package for Observational Studies in Health Care and Public Health. In.

12. GitHub [Internet]. [cited 2025 May 26]. Vocabulary Development Process. Available from: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Vocabulary-Development-Process

13. Quality Assurance and Control · OHDSI/Vocabulary-v5.0 Wiki · GitHub [Internet]. [cited 2025 May 26]. Available from: https://github.com/OHDSI/Vocab
v5.0/wiki/Quality-assurance-and-control

14. Introduction · OHDSI/Vocabulary-v5.0 Wiki · GitHub [Internet]. [cited 2025 May 26]. Available from: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Introduction

15. Ostropolets A. OHDSI Vocabularies landscape assessment [Internet]. 2023. Available from: https://ohdsiorg.sharepoint.com/:w:/s/Workgroup-CommonDataModel/EQZxds1n62JIsywmDCknwtABnSb42q7hM5PwiyblXV9zDw?e=cyv

16. Releases · OHDSI/Vocabulary-v5.0 [Internet]. [cited 2025 May 23]. Available from: https://github.com/OHDSI/Vocabulary-v5.0/releases

17. Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, et al. A Simple Standard for Sharing Ontological Mappings (SSSOM). Database. 2022 May 25;2022:baac035.

18. OHDSI/Tantalus [Internet]. Observational Health Data Sciences and Informatics; 2024 [cited 2025 May 28]. Available from: https://github.com/OHDSI/Tantalus

19. Park Y, Yoon J, Zhuk A, Ostropolets A, You SC. Integrating Local Vocabulary into OMOP CDM: A Step-by-Step Tutorial [Internet]. 2025 [cited 2025 May 26]. Available from: http://medrxiv.org/lookup/doi/10.1101/2025.05.07.25327200

20. GitHub [Internet]. [cited 2025 May 23]. Domains. Available from: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Domains

21. Dinu V, Nadkarni P. Guidelines for the effective use of entity–attribute–value modeling for biomedical databases. International Journal of Medical Informatics. 2007 Nov;76(11–12):769–79.

22. OHDSI GIS WG [Internet]. 2025 [cited 2025 May 28]. Available from: https://ohdsi.github.io/GIS/vocabulary.html

6. Extract Transform Load

TODO- write abstract

Chapter Leads: Erica Voss

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter 6 Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

7. Data Sources

TODO- write abstract

Chapter Leads: Melanie Philofsky

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter 7 Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

Part III.

Data Analytics

8. Data Analytics Use Cases

TODO- write abstract

Chapter Leads: Rakesh Babu

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

9. OHDSI Analytics Tools

TODO- write abstract

Chapter Leads: Anthony Sena

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

10. SQL and R

TODO- write abstract

Chapter Leads: TBC

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

11. Defining Cohorts

TODO- write abstract

Chapter Leads: Azza Shoaibi

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

12. Characterization

TODO- write abstract

Chapter Leads: Gowtham Rao

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

13. Population-Level Estimation

TODO- write abstract

Chapter Leads: Martijn Schuemie

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

14. Patient-Level Prediction

TODO- write abstract

Chapter Leads: Ross Williams

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

Part IV.

Evidence Quality

15. Evidence Quality

TODO- write abstract

Chapter Leads: Patrick Ryan

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

16. Data Quality

TODO- write abstract

Chapter Leads: Clair Blacketer

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

17. Clinical Validity

TODO- write abstract

Chapter Leads: Joel Swerdel

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

> The likelihood of transforming matter into energy is something akin to shooting birds in the dark in a country where there are only a few birds. *Einstein, 1935*

The vision of OHDSI is "A world in which observational research produces a comprehensive understanding of health and disease." Retrospective designs provide a vehicle for research using existing data but can be riddled with threats to various aspects of validity as discussed in Chapter14. It is not easy

to isolate clinical validity from quality of data and statistical methodology, but here we will focus on three aspects in terms of clinical validity: Characteristics of health care databases, Cohort validation, and Generalizability of the evidence. Let's go back to the example of population-level estimation (Chapter 12). We tried to answer the question "Do ACE inhibitors cause angioedema compared to thiazide or thiazide-like diuretics?" In that example, we demonstrated that ACE inhibitors caused more angioedema than thiazide or thiazide-like diuretics. This chapter is dedicated to answer the question: "To what extent does the analysis conducted match the clinical intention?"

17.1. Characteristics of Health Care Databases

It is possible that what we found is the relationship between **prescription** of ACE inhibitor and angioedema rather than the relationship between **use** of ACE inhibitor and angioedema. We've already discussed data quality in the previous chapter (15). The quality of the converted database into the Common Data Model (CDM) cannot exceed the original database. Here we are addressing the characteristics of most healthcare utilization databases. Many databases used in OHDSI originated from administrative claims or electronic health records (EHR). Claims and EHR have different data capture processes, neither of which has research as a primary intention. Data elements from claims records are captured for the purpose of reimbursement, financial transactions between clinicians and payers whereby services provided to patients by providers are sufficiently justified to enable agreement on payments by the responsible parties. Data elements in EHR records are captured to support clinical care and administrative operations, and they commonly only reflect the information that providers within a given health system feel are necessary to document the current service and provide necessary context for anticipated follow-up care within their health system. They may not represent a patient's complete medical history and may not integrate data from across health systems.

To generate reliable evidence from observational data, it is useful for a researcher to understand the journey that the data undergoes from the moment that a patient seeks care through the moment that the data reflecting that care are used in an analysis. As an example, "drug exposure" can be inferred from various sources of observational data, including prescriptions written by

clinicians, pharmacy dispensing records, hospital procedural administrations, or patient self-reported medication history. The source of data can impact our level of confidence in the inference we draw about which patients did or did not use the drug, as well as when and for how long. The data capture process can result in under-estimation of exposure, such as if free samples or over-the counter drugs are not recorded, or over-estimation of exposure, such as if a patient doesn't fill the prescription written or doesn't adherently consume the prescription dispensed. Understanding the potential biases in exposure and outcome ascertainment, and more ideally quantifying and adjusting for these measurement errors, can improve our confidence in the validity of the evidence we draw from the data we have available.

17.2. Cohort Validation

G. Hripcsak and Albers described that "a phenotype is a specification of an observable, potentially changing state of an organism, as distinguished from the genotype, which is derived from an organism's genetic makeup". (1) The term phenotype can be applied to patient characteristics inferred from electronic health record (EHR) data. Researchers have been carrying out EHR phenotyping since the beginning of informatics, from both structured data and narrative data. The goal is to draw conclusions about a target concept based on raw EHR data, claims data, or other clinically relevant data. Phenotype algorithms – i.e., algorithms that identify or characterize phenotypes – may be generated by domain exerts and knowledge engineers, including recent research in knowledge engineering or through diverse forms of machine learning…to generate novel representations of the data."

This description highlights several attributes useful to reinforce when considering clinical validity: 1) it makes it clear that we are talking about something that is observable (and therefore possible to be captured in our observational data); 2) it includes the notion of time in the phenotype specification (since a state of a person can change); 3) it draws a distinction between the phenotype as the desired intent vs. the phenotype algorithm, which is the implementation of the desired intent.

OHDSI has adopted the term "cohort" to define the set of persons satisfying one or more inclusion criteria for a duration of time. A "cohort definition" represents the logic necessary to instantiate a cohort against an observational

database. In this regard, the cohort definition (or phenotype algorithm) is used to produce a cohort, which is intended to represent the phenotype, being the persons who belong to the observable clinical state of interest.

Most types of observational analyses, including clinical characterization, population-level effect estimation, and patient-level prediction, require one or more cohorts to be established as part of the study process. To evaluate the validity of the evidence produced by these analyses, one must consider this question for each cohort: to what extent do the persons identified in the cohort based on the cohort definition and the available observational data accurately reflect the persons who truly belong to the phenotype?

To return to the population-level estimation example (Chapter 12) "Do ACE inhibitors cause angioedema compared to thiazide or thiazide-like diuretics?", we must define three cohorts: a target cohort of persons who are new users of ACE inhibitors, a comparator cohort of persons who are new users of thiazide diuretics, and an outcome cohort of persons who develop angioedema. How confident are we that all use of ACE inhibitors or thiazide diuretics is completely captured, such that "new users" can be identified by the first observed exposure, without concern of prior (but unobserved) use? Can we comfortably infer that persons who have a drug exposure record for ACE inhibitors were in fact exposed to the drug, and those without a drug exposure were indeed unexposed? Is there uncertainty in defining the duration of time that a person is classified in the state of "ACE inhibitor use," either when inferring cohort entry at the time the drug was started or cohort exit when the drug was discontinued? Have persons with a condition occurrence record of "Angioedema" actually experienced rapid swelling beneath the skin, differentiated from other types of dermatologic allergic reactions? What proportion of patients who developed angioedema received medical attention that would give rise to the observational data used to identify these clinical cases based on the cohort definition? How well can the angioedema events which are potentially drug-induced be disambiguated from the events known to be caused by other agents, such as food allergy or viral infection? Is disease onset sufficiently well captured that we have confidence in drawing a temporal association between exposure status and outcome incidence? Answering these types of questions is at the heart of clinical validity.

In this chapter, we will discuss the methods for validating cohort definitions. We first describe the metrics used to measure the validity of a cohort definition.

Next, we describe two methods to estimate these metrics: 1) clinical adjudication through source record verification, and 2) PheValuator, a semi-automated method using diagnostic predictive modeling.

17.2.1. Cohort Evaluation Metrics

Once the cohort definition for the study has been determined, the validity of the definition can be evaluated. A common approach to assess validity is by comparing some or all persons in a defined cohort to a reference 'gold standard' and expressing the results in a confusion matrix, a two-by-two contingency table that stratifies persons according to their gold standard classification and qualification within the cohort definition. Figure 17.1 shows the elements of the confusion matrix.

		Gold Standard	
		True	False
Cohort Definition	True	True Positive	False Positive
	False	False Negative	True Negative

Figure 17.1.: Confusion matrix.

The true and false results from the cohort definition are determined by applying the definition to a group of persons. Those included in the definition are considered positive for the health condition and are labeled "True." Those persons not included in the cohort definition are considered negative for the health condition and are labeled "False". While the absolute truth of a person's heath state considered in the cohort definition is very difficult to determine, there are multiple methods to establish a reference gold standard, two of which will be described later in the chapter. Regardless of the method used, the labeling of these persons is the same as described for the cohort definition.

In addition to errors in the binary indication of phenotype designation, the timing of the health condition may also be incorrect. For example, while the cohort definition may correctly label a person as belonging to a phenotype, the definition may incorrectly specify the date and time when a person without the condition became a person with the condition. This error would add

bias to studies using survival analysis results, e.g., hazard ratios, as an effect measure.

The next step in the process is to assess the concordance of the gold standard with the cohort definition. Those persons that are labeled by both the gold standard method and the cohort definition as "True" are called "True Positives." Those persons that are labeled by the gold standard method as "False" and by the cohort definition as "True" are called "False Positives," i.e., the cohort definition misclassified these persons as having the condition when they do not. Those persons that are labeled by both the gold standard method and the cohort definition as "False" are called "True Negatives." Those persons that are labeled by the gold standard method as "True" and by the cohort definition as "False" are called "False Negatives," i.e., the cohort definition incorrectly classified these persons as not having the condition, when it fact they do belong to the phenotype. Using the counts from the four cells in the confusion matrix, we can quantify the accuracy of the cohort definition in classifying phenotype status in a group of persons. There are standard performance metrics for measuring cohort definition performance:

1. **Sensitivity of the cohort definition** – what proportion of the persons who truly belong to the phenotype in the population were correctly identified to have the health outcome based on the cohort definition? This is determined by the following formula:

- Sensitivity = True Positives / (True Positives + False Negatives)

1. **Specificity of the cohort definition** – what proportion of the persons who do not belong to the phenotype in the population were correctly identified to not have the health outcome based on the cohort definition? This is determined by the following formula:

- Specificity = True Negatives / (True Negatives + False Positives)

1. **Positive predictive value (PPV) of the cohort definition** – what proportion of the persons identified by the cohort definition to have the health condition actually belong to the phenotype? This is determined by the following formula:

- PPV = True Positives / (True Positives + False Positives)

1. **Negative predictive value (NPV) of the cohort definition** – what proportion of the persons identified by the cohort definition to not have the health condition actually did not belong to the phenotype? This is determined by the following formula:

- NPV = True Negatives / (True Negatives + False Negatives)

Perfect scores for these measures are 100%. Due to the nature of observational data, perfect scores are usually far from the norm. Rubbo et al. reviewed studies validating cohort definitions for myocardial infarction. (2) Of the 33 studies they examined, only one cohort definition in one dataset obtained a perfect score for PPV. Overall, 31 of the 33 studies reported PPVs 70%. They also found, however, that of the 33 studies only 11 reported sensitivity and 5 reported specificity. PPV is a function of sensitivity, specificity, and prevalence. Datasets with different values for prevalence will produce different values for PPV with sensitivity and specificity held constant. Without sensitivity and specificity, correcting for bias due to imperfect cohort definitions is not possible. Additionally, the misclassification of the health condition may be differential, meaning the cohort definition performs differently on one group of persons relative to the comparison group, or non-differentially, when the cohort definition performs similarly on both comparison groups. Prior cohort definition validation studies have not tested for potential differential misclassification, even though it can lead to strong bias in effect estimates.

Once the performance metrics have been established for the cohort definition, these may be used to adjust the results for studies using these definitions. In theory, adjusting study results for these measurement error estimates has been well established. In practice, though, because of the difficulty in obtaining the performance characteristics, these adjustments are rarely considered. The methods used to determine the gold standard are described in the remainder of this section.

17.3. Source Record Verification

A common method used to validate cohort definitions has been clinical adjudication through source record verification: a thorough examination of a person's records by one or more domain experts with sufficient knowledge to

competently classify the clinical condition or characteristic of interest. Chart review generally follows the following steps:

1. Obtain permission from local institutional review board (IRB) and/or persons as needed to conduct study including chart review.
2. Generate cohort using cohort definition to be evaluated. Sample a subset of the persons to manually review if there are insufficient resources to adjudicate the entire cohort.
3. Identify one or more persons with sufficient clinical expertise to review person records.
4. Determine guidelines for adjudicating whether a person is positive or negative for the desired clinical condition or characteristic.
5. Clinical experts review and adjudicate all available data for the persons within the sample to classify each person as to whether they belong to the phenotype or not.
6. Tabulate persons according to the cohort definition classification and clinical adjudication classification into a confusion matrix, and calculate the performance characteristics possible from the data collected.

Results from a chart review are typically limited to the evaluation of one performance characteristic, positive predictive value (PPV). This is because the cohort definition under evaluation only generates persons that are believed to have the desired condition or characteristics. Therefore, each person in the sample of the cohort is classified as either a true positive or false positive based on the clinical adjudication. Without knowledge of all persons in the phenotype in the entire population (including those not identified by the cohort definition), it is not possible to identify the false negatives, and thereby fill in the remainder of the confusion matrix to generate the remaining performance characteristics. Potential methods of identifying all persons in the phenotype across the population include chart review of the entire database, which is generally not feasible unless the overall population is small, or the utilization of comprehensive clinical registries in which all true cases have already been flagged and adjudicated, such as tumor registries (see example below). Alternatively, one can sample persons who do not qualify for the cohort definition to produce a subset of predicted negatives, and then repeating steps 3-6 of the chart review above to check whether these patients are truly lacking the clinical condition or characteristic of interest can identify true negatives or

false negatives. This would allow the estimation of negative predictive value (NPV), and if an appropriate estimate of the phenotype prevalence is available, then sensitivity and specificity can be estimated.

There are a number of limitations to clinical adjudication through source record verification. As alluded to earlier, chart review can be a very time-consuming and resource-intensive process, even just for the evaluation of a single metric such as PPV. This limitation significantly impedes the practicality of evaluating an entire population to fill out a complete confusion matrix. In addition, multiple steps in the above process have the potential to bias the results of the study. For example, if records are not equally accessible in the EHR, if there is no EHR, or if individual patient consent is required, then the subset under evaluation may not be truly random and could introduce sampling or selection bias. In addition, manual adjudication is susceptible to human error or misclassification and thereby may not represent a perfectly accurate metric. There can often be disagreement between clinical adjudicators due to the data in the person's record being vague, subjective, or of low quality. In many studies, the process involves a majority-rules decision for consensus which yields a binary classification for persons that does not reflect the inter-rater discordance.

17.3.1. Example of Source Record Verification

An example of the process to conduct a cohort definition validation using chart review is provided from a study by the Columbia University Irving Medical Center (CUIMC), which validated a cohort definition for multiple cancers as part of a feasibility study for the National Cancer Institute (NCI). The steps used to conduct the validation for the example of one of these cancers—prostate cancer—are as follows:

1. Submitted proposal and obtained IRB consent for OHDSI cancer phenotyping study.
2. Developed a cohort definition for prostate cancer: Using ATHENA and ATLAS to explore the vocabulary, we created a cohort definition to include all patients with a condition occurrence for Malignant Tumor of Prostate (concept ID 4163261), excluding Secondary Neoplasm of Prostate (concept ID 4314337) or Non-Hodgkin's Lymphoma of Prostate (concept ID 4048666).

3. Generated cohort using ATLAS and randomly selected 100 patients for manual review, mapping each PERSON_ID back to patient MRN using mapping tables. 100 patients were selected in order to achieve our desired level of statistical precision for the performance metric of PPV.
4. Manually reviewed records in the various EHRs—both inpatient and outpatient—in order to determine whether each person in the random subset was a true or false positive.
5. Manual review and clinical adjudication were performed by one physician (although ideally in future more rigorous validation studies would be done by a higher number of reviewers to assess for consensus and inter-rater reliability).
6. Determination of a reference standard was based on clinical documentation, pathology reports, labs, medications and procedures as documented in the entirety of the available electronic patient record.
7. Patients were labeled as 1) prostate cancer 2) no prostate cancer or 3) unable to determine.
8. A conservative estimate of PPV was calculated using the following: prostate cancer/ (no prostate cancer + unable to determine).
9. Then, using the tumor registry as an additional gold standard to identify a reference standard across the entire CUIMC population, we counted the number of persons in the tumor registry which were and were not accurately identified by the cohort definition, which allowed us to estimate sensitivity using these values as true positives and false negatives.
10. Using the estimated sensitivity, PPV, and prevalence, we could then estimate specificity for this cohort definition. As noted previously, this process was time-consuming and labor-intensive, as each cohort definition had to be individually evaluated through manual chart review as well as correlated with the CUIMC tumor registry in order to identify all performance metrics. The IRB approval process itself took weeks despite an expedited review while obtaining access to the tumor registry, and the process of manual chart review itself took a few weeks longer.

A review of validation efforts for myocardial infarction (MI) cohort definitions by Rubbo et al. found that there was significant heterogeneity in the cohort definitions used in the studies as well as in the validation methods and the results reported. (2) The authors concluded that for acute myocardial infarction there is no gold standard cohort definition available. They noted that the process was both costly and time-consuming. Due to that limitation, most studies had small sample sizes in their validation leading to wide variations

in the estimates for the performance characteristics. They also noted that in the 33 studies, while all the studies reported positive predictive value, only 11 studies reported sensitivity and only five studies reported specificity. As mentioned previously, without estimates of sensitivity and specificity, statistical correction for misclassification bias cannot be performed.

17.4. PheValuator

The OHDSI community has developed a different approach to constructing a gold standard by using diagnostic predictive models.(3,4) The general idea is to emulate the ascertainment of the health outcome similar to the way clinicians would in a source record validation, but in an automated way that can be applied at scale. The tool has been developed as an open-source R package called PheValuator.[1] PheValuator uses functions from the Patient Level Prediction package.

The process is as follows:

1. Create an extremely specific ("**xSpec**") cohort: Determine a set of persons with a very high likelihood of having the outcome of interest to be used as noisy positive labels when training a diagnostic predictive model.
2. Create an extremely sensitive ("**xSens**") cohort: Determine a set of persons that should include anyone who could possibly have the outcome. This cohort will be used to identify its inverse: the set of people we are confident do not have the outcome, to be used as noisy negative labels when training a diagnostic predictive model.
3. Fit a predictive model using the xSpec and xSens cohort: As described in Chapter 13, we fit a model using a wide array of patient features as predictors and aim to predict whether a person belongs to the xSpec cohort (those we believe have the outcome) or the inverse of the xSens cohort (those we believe do not have the outcome).
4. Apply the fitted model to estimate the probability of the outcome for a hold-out set of persons who will be used to evaluate cohort definition performance: The set of predictors from the model can be applied to a person's data to estimate the predicted probability that the person belongs to the phenotype. We use these predictions as a **probabilistic gold standard**.

5. Evaluate the performance characteristics of the cohort definitions: We compare the predicted probability to the binary classification of a cohort definition (the test conditions for the confusion matrix). Using the test conditions and the estimates for the true conditions, we can fully populate the confusion matrix and estimate the entire set of performance characteristics, i.e., sensitivity, specificity, and predictive values.

The primary limitation to using this approach is that the estimation of the probability of a person having the health outcome is limited by the data in the database. Depending on the database, important information, such as clinician notes, may not be available.

In diagnostic predictive modeling we create a model that discriminates between those with the disease and those without the disease. As described in the Patient-Level Prediction chapter (Chapter 13), prediction models are developed using a *target cohort* and an *outcome cohort*. The target cohort includes persons with and without the health outcome; the outcome cohort identifies those persons in the target cohort with the health outcome. For the PheValuator process, we use an extremely specific cohort definition, the "xSpec" cohort, to determine the outcome cohort for the prediction model. The xSpec cohort uses a definition to find those with a very high probability of having the disease of interest. The xSpec cohort may be defined as those persons who have multiple condition occurrence records for the health outcome of interest in a specified period of time. For example, for a chronic disease such as atrial fibrillation, we may have persons who have two or more records with the atrial fibrillation diagnosis code in a 14-day period. For MI, an acute outcome, we may use two or more occurrences of MI during a single day and include the requirement of having at least two occurrences from an inpatient setting. The target cohort for the predictive model is constructed from the union of persons with a low likelihood of having the health outcome of interest and those persons in the xSpec cohort. To determine those persons with a low likelihood of having the health outcome of interest, we sample from the entire database and exclude persons who have some evidence suggestive of belonging to the phenotype, typically by removing persons with any records containing the concepts used to define the xSpec cohort. There are limitations to this method. It is possible that these xSpec cohort persons may have different characteristics than others with the disease. We use LASSO logistic regression to create the prediction model used to generate the probabilistic gold standard.(5) This algorithm produces a parsimonious model and typically removes

many of the collinear covariates which may be present across the dataset. In the current version of the PheValuator software, outcome status (yes/no) is evaluated based on parameters set in the analysis specification, For example, for chronic conditions, we may set the time to determine the characteristics within 365 days of the start of the condition. To add greater detail to the model, we may break the 365 day observation time into three separate time windows, such as 0-30 days, 31-90 days, and 91 to 365 days. For acute conditions, we may limit the time to 30 days after the diagnosis. PheValuator does not evaluate the accuracy of the cohort start date.

17.4.1. Example Validation By PheValuator

We may use PheValuator to assess the complete performance characteristics for a cohort definition to be used in a study where it is necessary to determine those persons who have had an acute myocardial infarction (MI).

The following are the steps for testing cohort definitions for MI using PheValuator:

17.4.1.1. Step 1: Define the xSpec Cohort

Determine those with MI with a high probability: For the cohort entry event, we required a condition occurrence record with a concept for myocardial infarction or any of its descendants (Figure 17.2). We required that each subject in the xSpec cohort have at least 30 days observation time after the cohort entry event to ensure that data for the model will be available through the complete observed prediction window.

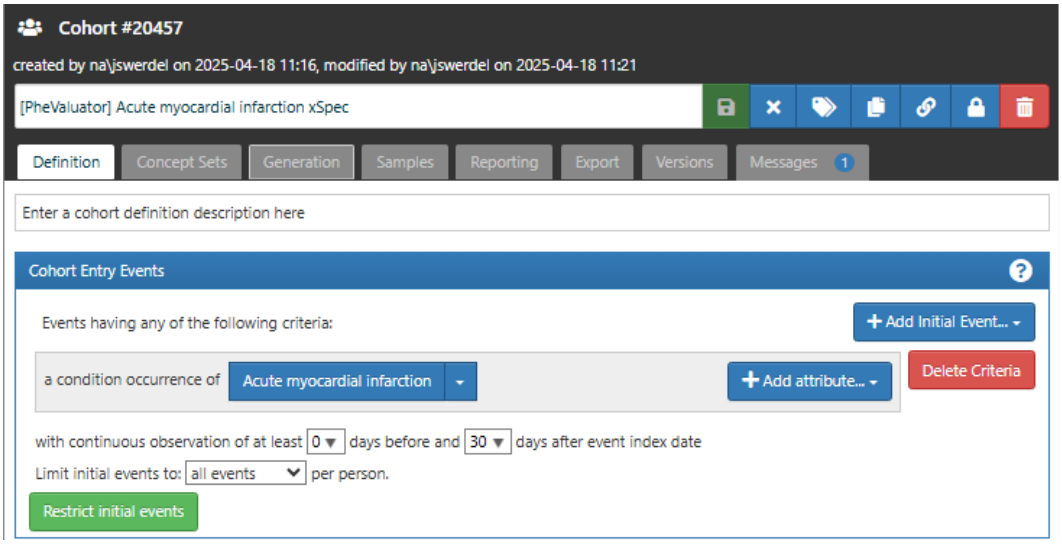


Figure 17.2.: Cohort entry event in ATLAS for an extremely specific cohort definition (xSpec) for myocardial infarction.

We next added an inclusion criteria requiring either a drug exposure of antithrombotic agent or a second diagnosis code for myocardial infarction on the same day as the cohort entry event (Figure 17.3). These inclusion criteria increase the specificity of the cohort, i.e., increasing the likelihood that the subjects selected by the definition is a case of MI.

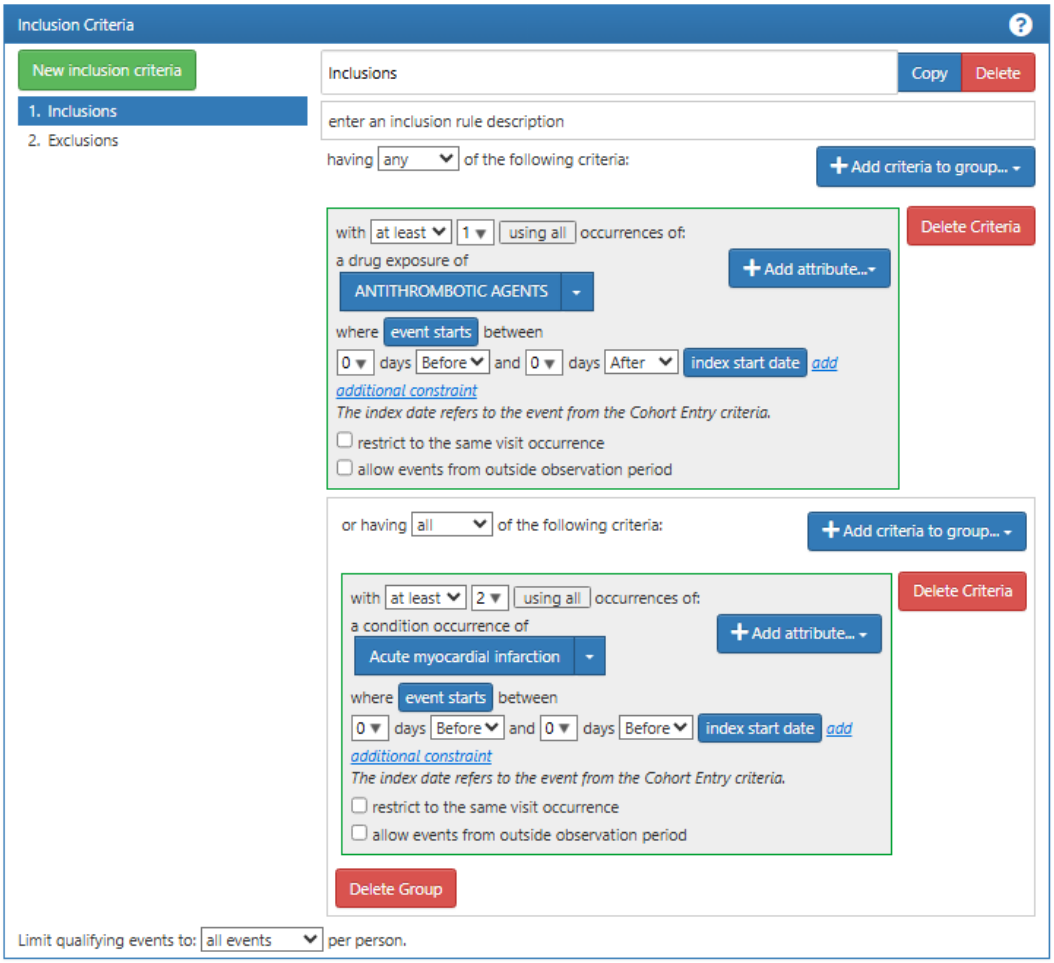


Figure 17.3.: Cohort inclusion event in ATLAS for an extremely specific cohort definition (xSpec) for myocardial infarction.

Finally, we add an exclusion event where we exclude a set of differential diagnoses in the period 7 days before and 14 days after the cohort entry event (Figure 17.4). These include conditions such as myocarditis and anxiety disorder. These exclusion events help to increase the validity of the diagnosis of myocardial infarction, further ensuring that the subjects in this cohort have a high probability of having the condition of interest.

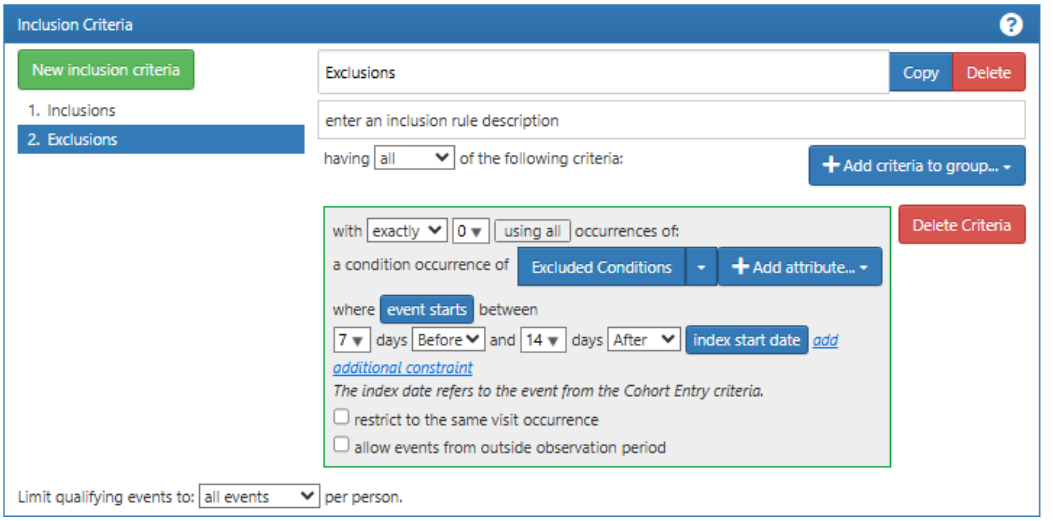


Figure 17.4.: Cohort exclusion event in ATLAS for an extremely specific cohort definition (xSpec) for myocardial infarction.

17.4.1.2. Step 2: Define the xSens Cohort

Next, we develop an extremely sensitive cohort (xSens). This cohort may be defined for MI as those persons with at least one condition occurrence record containing a myocardial infarction concept at any time in their medical history. Figure 17.5 illustrates the xSens cohort definition for MI in ATLAS.

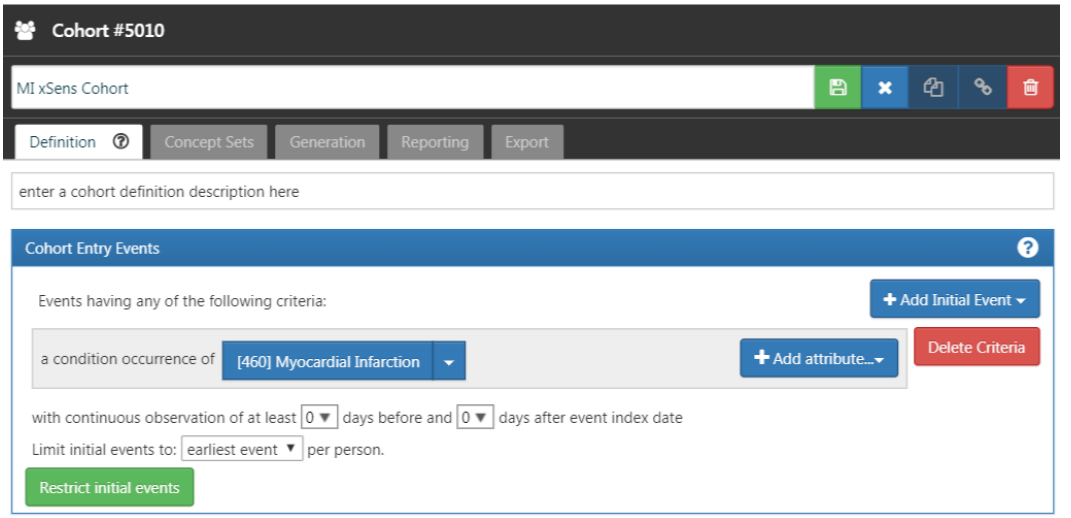


Figure 17.5.: An extremely sensitive cohort definition (xSens) for myocardial infarction.

The xSens cohort will be used to find all subjects with even a low probability of having the outcome of interest. When we select a large random set of subjects and remove those in the xSens cohort, the subjects remaining will likely all have a very low probability of having the outcome of interest.

17.4.1.3. Step 3: Running the PheValuator process

The PheValuator process involves three parts:

1. Developing the diagnostic predictive model
2. Applying the model to a large, random set of subjects in the database, the "evaluation cohort", to produce a "probabilistic gold standard"
3. Using the "probabilistic gold standard" to calculate the performance characteristics for phenotypes to be used in research studies

These steps can be performed using the *runPheValuatorAnalyses* function as described in the PheValuator vignette stored in the PheValuator repository in GitHub (https://github.com/OHDSI/PheValuator/tree/main).

At the end of the process, PheValuator will produce an output file with the performance characteristics for the tested phenotypes. The desired performance characteristics for each may depend on the intended use of the cohort to address the research question of interest. For certain questions, a very sensitive algorithm may be required; others may require a more specific algorithm. The process for determining the performance characteristics for a cohort definition using PheValuator is shown in Figure 17.6.

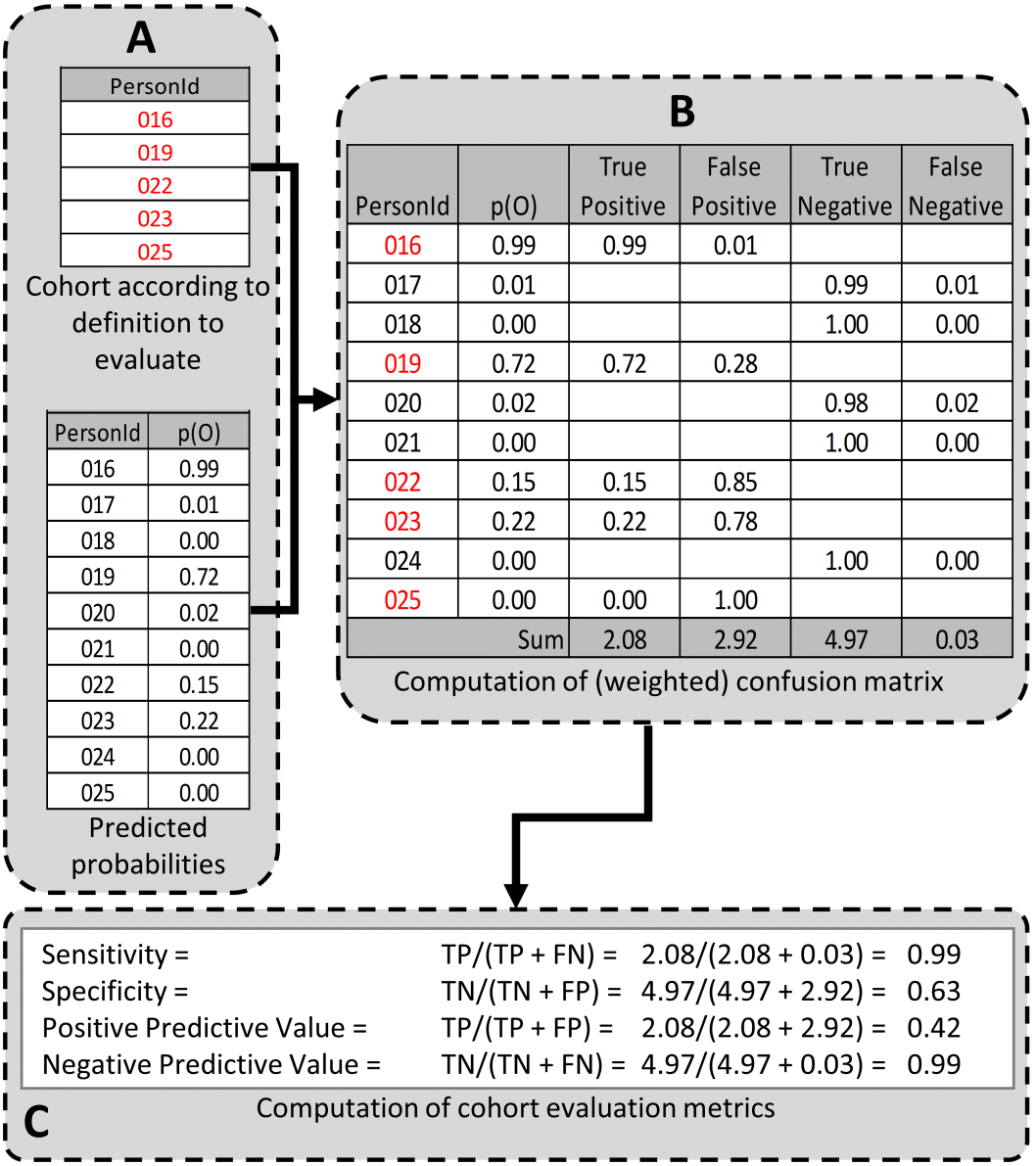


Figure 17.6.: Determining the Performance Characteristics of a cohort definition using PheValuator. p(O) = Probability of outcome; TP = True Positive; FN = False Negative; TN = True Negative; FP = False Positive.

In part A of Figure 17.6, we examined the persons from the cohort definition

to be tested and found those persons from the evaluation cohort (created in the previous step) who were included in the cohort definition (Person IDs 016, 019, 022, 023, and 025) and those from the evaluation cohort who were excluded from the cohort definition (Person Ids 017, 018, 020, 021, and 024). For each of these included/excluded persons, we had previously determined the probability of the health outcome using the predictive model (p(O)).

We estimated the values for True Positives, True Negatives, False Positives, and False Negatives as follows (Part B of Figure 17.6):

1. If the cohort definition included a person from the evaluation cohort, i.e., the cohort definition considered the person a "positive." The predicted probability for the health outcome indicated the expected value of the number of counts contributed by that person to the True Positives, and one minus the probability indicated the expected value of the number of counts contributed by that person to the False Positives for that person. We added all the expected values of counts across persons to get the total expected value. For example, PersonId 016 had a predicted probability of 99% for the presence of the health outcome, 0.99 was added to the True Positives (expected value of counts added 0.99) and 1.00–0.99 = 0.01 was added to the False Positives (0.01 expected value). Another way to think of this is that the cohort definition that selected this person got it 99% right and 1% wrong. This was repeated for all the persons from the evaluation cohort included in the cohort definition (i.e., PersonIds 019, 022, 023, and 025).

2. Similarly, if the cohort definition did not include a person from the evaluation cohort, i.e. the cohort definition considered the person a "negative," one minus the predicted probability for the phenotype for that person was the expected value of counts contributed to True Negatives and was added to it, and, in parallel, the predicted probability for the phenotype was the expected value of counts contributed to the False Negatives and was added to it. For example, PersonId 017 had a predicted probability of 1% for the presence of the health outcome (and, correspondingly, 99% for the absence of the health outcome) and 1.00 – 0.01 = 0.99 was added to the True Negatives and 0.01 was added to the False Negatives. This was repeated for all the persons from the evaluation cohort not included in the cohort definition (i.e., PersonIds 018, 020, 021, and 024).

After adding these values over the full set of persons in the evaluation cohort,

we filled the four cells of the confusion matrix with the expected values of counts for each cell, and we were able to create point estimates for the tested cohort's performance characteristics, i.e., sensitivity, specificity, and positive and negative (NPV) predictive value (Figure 1C). In the example, the sensitivity, specificity, PPV, and NPV were 0.99, 0.63, 0.42, and 0.99, respectively. PheValuator also calculated the confidence intervals from these estimates.

The desired performance characteristics may depend on the intended use of the cohort to address the research question of interest. For certain questions, a very sensitive algorithm may be required; others may require a more specific algorithm. An example of the output from an analysis for acute myocardial infarction is shown in Figure 17.7.

databaseId	description	cohortId	sensitivity95Ci	ppv95Ci	specificity95Ci	npv95Ci	f1Score
Optum DOD	xSpec	11081	0.454 (0.451 - 0.457)	0.960 (0.959 - 0.962)	0.999 (0.999 - 0.999)	0.967 (0.967 - 0.968)	0.617
Optum EHR	xSpec	11081	0.392 (0.389 - 0.395)	0.980 (0.978 - 0.981)	1.000 (1.000 - 1.000)	0.966 (0.966 - 0.967)	0.56
Optum DOD	[PL] All events of Acute Myocardial Infarction, inpatient setting with washout period of 365 days	2072	0.777 (0.775 - 0.780)	0.887 (0.885 - 0.889)	0.994 (0.993 - 0.994)	0.986 (0.986 - 0.986)	0.829
Optum EHR	[PL] All events of Acute Myocardial Infarction, inpatient setting with washout period of 365 days	2072	0.567 (0.564 - 0.570)	0.937 (0.935 - 0.938)	0.998 (0.998 - 0.998)	0.975 (0.975 - 0.975)	0.707

Figure 17.7.: Example output from a PheValuator analysis for acute myocardial infarction.

In this example, we included the results from the xSpec cohort (cohort ID 11081) as well as the cohort of interest (cohort ID 2072). The performance characteristics of the xSpec cohort showed high PPV and low sensitivity compared to the test cohort, "[PL] All events of Acute Myocardial Infarction, inpatient setting with washout period of 365 days". This is expected as the criteria for the xSpec cohort was very specific for acute myocardial infarction leading to a high PPV. PheValuator calculates the F1 score which is the harmonic mean of the sensitivity and the PPV.

17.5. Generalizability of the Evidence

While a cohort can be well-defined and fully evaluated within the context of a given observational database, the clinical validity is limited by the extent

to which the results are considered generalizable to the target population of interest. Multiple observational studies on the same topic can yield different results, which can be caused by not only by their designs and analytic methods, but also bt their choice of data source. Madigan et al. (2013) demonstrated that choice of database affects the result of observational study. They systematically investigated heterogeneity in the results for 53 drug-outcome pairs and two study designs (cohort studies and self-controlled case series) across the 10 observational databases. Even though they held study design constant, substantial heterogeneity in effect estimates was observed.

Across the OHDSI network, observational databases vary considerably in the populations they represent (e.g. pediatric vs. elderly, privately-insured employees vs. publicly-insured unemployed), the care settings where data are captured (e.g. inpatient vs. outpatient, primary vs. secondary/specialty care), the data capture processes (e.g. administrative claims, EHRs, clinical registries), and the national and regional health system from which care is based. These differences can manifest as heterogeneity observed when studying disease and the effects of medical interventions and can also influence the confidence we have in the quality of each data source that may contribute evidence within a network study. While all databases within the OHDSI network are standardized to the CDM, it is important to reinforce that standardization does not reduce the true inherent heterogeneity that is present across populations, but simply provides a consistent framework to investigate and better understand the heterogeneity across the network. The OHDSI research network provides the environment to apply the same analytic process on various databases across the world, so that researchers can interpret results across multiple data sources while holding other methodological aspects constant. OHDSI's collaborative approach to open science in network research, where researchers across participating data partners work together alongside those with clinical domain knowledge and methodologists with analytical expertise, is one way of reaching a collective level of understanding of the clinical validity of data across a network that should serve as a foundation for building confidence in the evidence generated using these data.

17.6. Summary

- Clinical validity can be established by understanding the characteristics of the underlying data source, evaluating the performance characteristics of the cohorts within an analysis, and assessing the generalizability of the study to the target population of interest.
- A cohort definition can be evaluated on the extent to which persons identified in the cohort based on the cohort definition and the available observational data accurately reflect the persons who truly belong to the phenotype.
- Cohort definition validation requires estimating multiple performance characteristics, including sensitivity, specificity, and positive predictive value, to fully summarize and enable adjustment for measurement error.
- Clinical adjudication through source record verification and PheValuator represent two alternative approaches to estimating cohort definition validation.
- OHDSI network studies provide a mechanism to examine data source heterogeneity and expand the generalizability of findings to improve clinical validity of real-world evidence.

17.7. References

1. Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. J Am Med Inform Assoc. 2018 Mar 1;25(3):289–94.

2. Rubbo B, Fitzpatrick NK, Denaxas S, Daskalopoulou M, Yu N, Patel RS, et al. Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. Int J Cardiol. 2015 May;187:705–11.

3. Swerdel JN, Hripcsak G, Ryan PB. PheValuator: Development and evaluation of a phenotype algorithm evaluator. J Biomed Inform. 2019 Sep;97:103258.

4. Swerdel JN, Schuemie M, Murray G, Ryan PB. PheValuator 2.0: Methodological improvements for the PheValuator approach to semi-automated phenotype algorithm evaluation. J Biomed Inform. 2022 Nov;135:104177.

5. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive Parallelization of Serial Inference Algorithms for a Complex Generalized Linear Model. ACM Trans Model Comput Simul. 2013 Jan;23(1):1–17.

[1] https://github.com/OHDSI/PheValuator

18. Software Validity

TODO- write abstract

Chapter Leads: Martijn Schuemie, Anthony Sena

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

19. Diagnostics

TODO- write abstract

Chapter Leads: Martijn Schuemie, Mitch Conover

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

Part V.

OHDSI Research

20. Study Steps

TODO- write abstract

Chapter Leads: Anthony Sena

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

21. OHDSI Network Research

TODO- write abstract

Chapter Leads: Kristin Kostka

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

22. Engagement with Networks

TODO- write abstract

Chapter Leads: Clair Blacketer

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

Part VI.

OHDSI in Action

23. Study Steps

TODO- write abstract

Chapter Leads: Julie Green

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

24. Generative AI

TODO- write abstract

Chapter Leads: tbc

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

Part VII.

Back Matter

Glossary

TODO- write abstract

Chapter Leads: tbc

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory ?
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page

Protocol Template

TODO- write abstract

Chapter Leads: tbc

> **Note**
>
> This is page is currently a stub. The chapter is being written in the OHDSI Teams directory. When the draft is complete, it will be converted to markdown and moved to this file.
> Author Resources (requires an OHDSI Teams account):
>
> - Chapter Directory ?
> - Book Layout
> - Education Working Group SharePoint Drive
>
> Public Resources:
>
> - Book of OHDSI, Edition 1
> - Source Code for Book of OHDSI, Edition 1
> - OHDSI Home Page