Concept Set Curation

v1.0

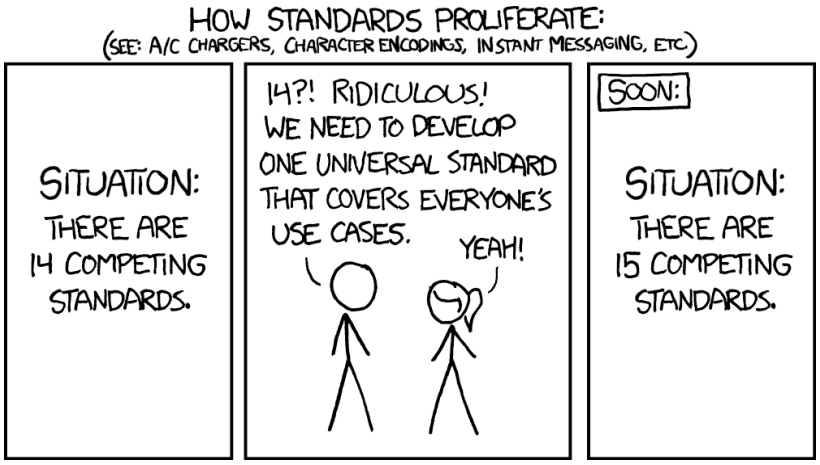
We've already seen some OMOP "concept sets" via the **concept_set_members** table and the Concept Set Browser, but we haven't gone into much detail on what a concept set *is*. (For those who may be familiar, they are often called "value sets" or "code sets" in non-OMOP contexts.)

A concept set is a list of medical codes–usually these pertain to some particular medical idea and are used to identify cohorts (sets of patients) for a research question. Consider pneumonia: there are many different kinds of pneumonia, such as pneumonia due to gram-negative bacteria, measles pneumonia, and adenoviral pneumonia, any of these may be found in a patient record. It may also be that the specific kind of pneumonia isn't known or recorded; gram-negative bacterial pneumonia is a kind of bacterial pneumonia, which is a kind of infective pneumonia, which is a kind of pneumonia. Any of these may be found in a patient record as well. A concept set meant to capture "all kinds of pneumonia" would need to include all of these for matching against patient records.

Concept sets can describe more than just diseases. Drugs are another example, where we might want to collect all of the drug administrations containing a fluoroquinolone antibiotic, of which there are many kinds, formulations, and doses. Lab tests may also be captured by concept sets. One of the most commonly used concept sets in N3C contains codes for all the existing COVID-19 PCR or Antigen tests. Even the *results* of lab tests may be captured by a concept set–it turns out there are various ways of encoding a positive result in a lab test ("positive", "detected", "reactive", and so on).

Mapping and Standards

All of these medical concepts are represented by medical codes (or concepts, in OMOP lingo), as a term in a medical *vocabulary*. For example, the Anatomical Therapeutic Chemical (ATC) classification system knows the drug ingredient "ciprofloxacin" as S02AA15. Another drug classification system known as RxNorm knows ciprofloxacin as ID 2551. The MeSH term (Medical Subject Headings) for ciprofloxacin is D002939. Meanwhile, there are codes and IDs for all the various drugs that contain cipro as an ingredient. In case it's not clear from these examples there are many competing standards for medical vocabularies, across drugs, diseases, procedures, lab tests, and more, and we can't guarantee that N3C data (which comes from dozens of contributing medical centers) will use any particular one.



How can we organize all of this information in a way that is useful to researchers? This is precisely what the OMOP Common Data Model (CDM) is meant to do. OMOP, which is short for Observational Medical Outcomes Partnership, was a US-funded effort to prove the usefulness of observational medical data and solve issues around EHR interoperability. Today, OMOP is stewarded by the Observational Health Data Sciences and Informatics (OHDSI, pronounced “odyssey”) project, a multinational, multidisciplinary community of researchers dedicated to enabling observational medical record research.

Important: *Before we get too far, know that there is a lot to know about OMOP, and we are only going to be able to hit some high points. OHDSI has published an important resource for understanding OMOP in the Book of OHDSI, and the Ehden Academy hosts video-based training courses in OMOP and OHDSI tools.*

While OMOP, being a Common Data Model, is yet another way to organize medical data (alongside the likes of the PCORNet, i2b2, and ACT CDMs), it does not define new medical terms. Instead, it leverages existing terms, like those found in ICD10, SNOMED, RxNorm, and others. Note that OMOP cannot guarantee that a code in one vocabulary–like 2551 in RxNorm–won’t be used by another vocabulary for something else. As a result, OMOP assigns to each term a unique, OMOP-specific ID, known as a **concept id**. When data are ingested into OMOP, the source vocabulary IDs are *mapped* to their corresponding OMOP concept ID, and potentially mapped *again* to a “standard” concept ID for use in analysis. While some vocabularies like ICD10 mix numbers and letters in their IDs, OMOP concept IDs are always just numbers.

To begin to make sense of this, let's consider a few columns of a hypothetical row in the OMOP **drug_exposure** table.

drug_exposure (data)

drug_concept_id	drug_concept_name	…	drug_source_concept_id	drug_source_concept_name	drug_source_value
1797513	ciprofloxacin	…	3193927	Ciprofloxacin Hydrochloride	20450009

When data are ingested into N3C's database from a contributing source, the data we get is primarily the **drug_source_value**, which is the code used by the originating system. In this case 20450009 is the code for "Ciprofloxacin Hydrochloride" in the Nebraska Lexicon vocabulary, which is recorded in the **drug_source_concept_name** column. Again, because OMOP cannot ensure that 20450009 isn't used for something else in some other vocabulary, a unique OMOP-specific concept ID is given, 3193927, as recorded in the **drug_source_concept_id** column. To summarize, OMOP concept ID 3193927 represents code 20450009 (a.k.a. "Ciprofloxacin Hydrochloride") in the Nebraska Lexicon.

But what about the first two columns, **drug_concept_id** and **drug_concept_name**? If we were to look up the OMOP concept ID 1797513, we'd find that it represents code 2551 (a.k.a. "ciprofloxacin") in the RxNorm vocabulary. We'd also discover that this concept ID is special, because it is considered the **standard** concept ID to represent ciprofloxacin. Consider, with all of the various ways to refer to cipro across various vocabularies, it would be most helpful for researchers if the world just standardized on one of them, which is what OMOP does, using **relationships** between concepts.

As a consequence, **you should only find standard concept IDs in the drug_concept_id column** (and their corresponding human-readable names in the **drug_concept_name** column).

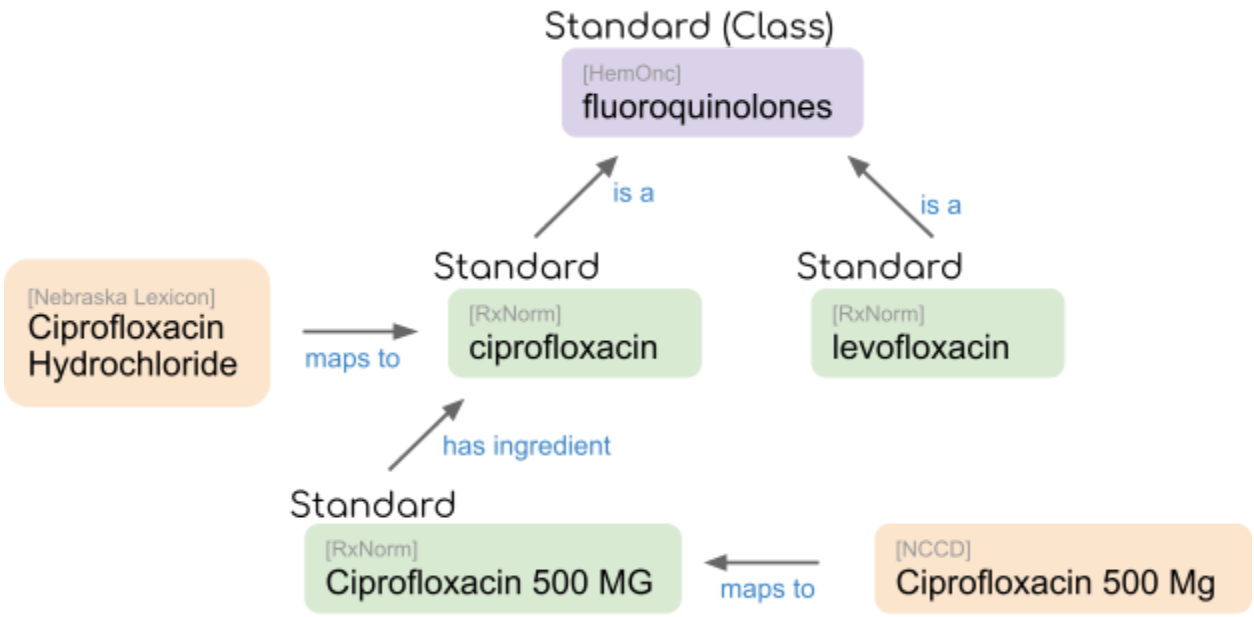
By the way, there's a naming convention in play here. For example **procedure_source_value**, **procedure_source_concept_name**, **procedure_concept_name**, and **procedure_concept_id** have analogous meanings in the OMOP procedure table. In general columns matching the pattern *_source_* refer to how information was represented at the data source, while

***_concept_id** and ***_concept_name** (without any mention of source) refer to the mapped standard concepts.

The OMOP Hierarchy

Standards certainly reduce the number of things a researcher will need to look for in the data, but they don't solve the problem of different levels of specificity that might be present in a patient record. As we noted above, for a patient with gram-negative bacterial pneumonia, we might find in a record indicating that exactly, or "bacterial pneumonia", "infective pneumonia", or just "pneumonia", each of which has a standard OMOP concept ID. In the case of drugs, the record might indicate a specific formulation or dose. Hopefully the record doesn't just say "ciprofloxacin," but EHR is often messy, underspecified, or missing altogether.

To address this, OMOP provides a "hierarchy" of relationships, *but only between standard concepts*. In the figure below, the standard concept ID for Cipro 500 MG is noted as a child (via a "has ingredient" relationship) of the concept for "ciprofloxacin", which is a child of the standard term for the class of "fluoroquinolones." Class concepts are a bit different–they are considered standards, but they wouldn't be found in the data directly, they are just used to help organize other concepts.



By the way, although we've drawn this portion of the hierarchy as a tree (where each child has only one parent), many concepts have multiple parents. Pneumonia, for example, is a child of

pneumonitis (lung inflammation) and lung consolidation (substance in lung), both of which happen to be children of “disorder of lung”.

Building Concept Sets with ATLAS

It is this hierarchical set of relationships that allow us to quickly generate a list of all concepts we might find in the data for a research question of interest. To find all prescriptions of drugs containing ciprofloxacin, we can gather the concept ID for ciprofloxacin and all of its descendent concept IDs. To find all kinds of pneumonia we can do the same. We can even exclude parts of the hierarchy. Suppose for example we want to identify all cases of pneumonia, except those that are known to be caused by something other than an infection (like rheumatic pneumonia). We can gather the standard concept for pneumonia and all its descendents, and then remove from that set the concept for “non-infective pneumonia” and all of its descendents.

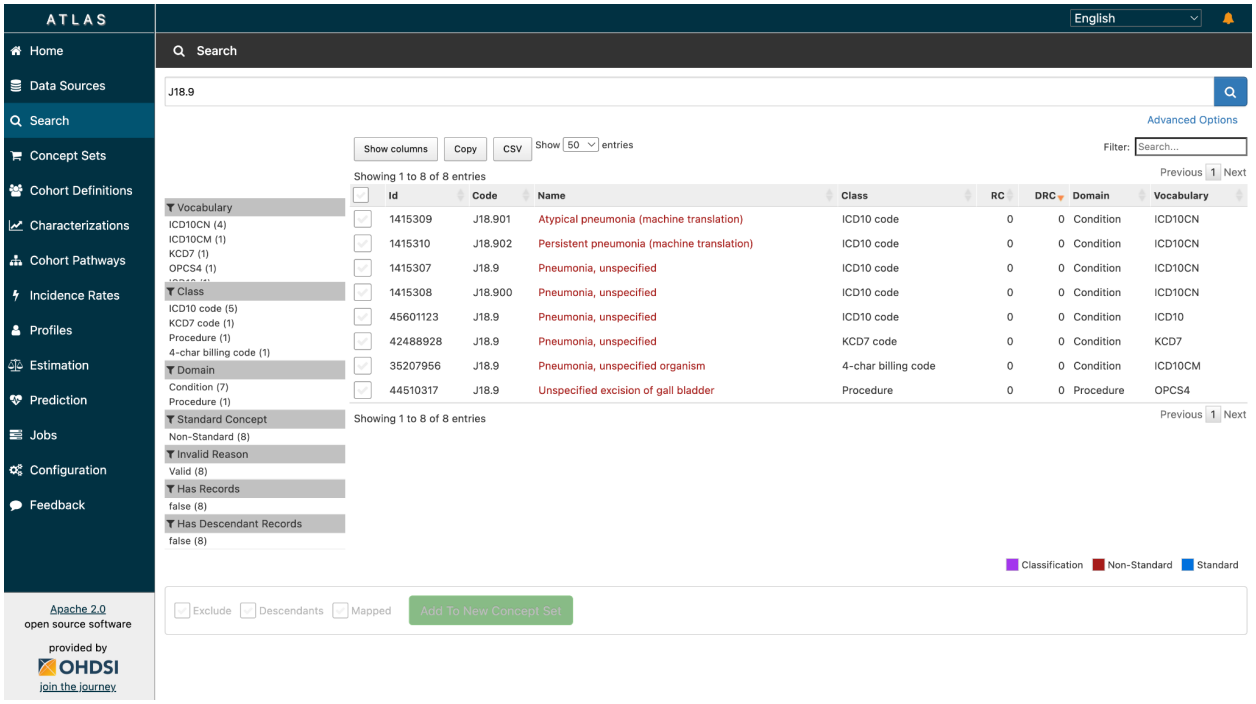
Important: *Although the OMOP hierarchy makes it quick and seemingly easy to develop a concept set, in real research this frequently is not the case. Identifying concepts associated with pregnancy is a good example, as there are many disparate conditions, drugs, procedures, and other kinds of terms to consider. Further, using the hierarchy to build concept sets makes it easy to accidentally include concepts that should be excluded, like non-infective pneumonias when the intent is to study an infectious disease like COVID-19. These errors are difficult to identify without training in medical vocabularies, so review by clinicians with relevant expertise is an important part of team science in N3C.*

Concepts and relationships between them are managed by tables; there’s a table called **concept** that lists all the vocabulary terms used by a given OMOP database and their information. There’s a table called **concept_relationships** that describes relationships between them. If you are curious, these and other OMOP meta-tables can be found in the Data Catalog under the “OMOP Concepts” collection.

We could, in theory, use something like SQL to use the hierarchy to find concepts of interest. However, both OHDSI and N3C have developed tools to help you navigate the hierarchy and generate concept sets. We’ll only have time to cover a few features of these.

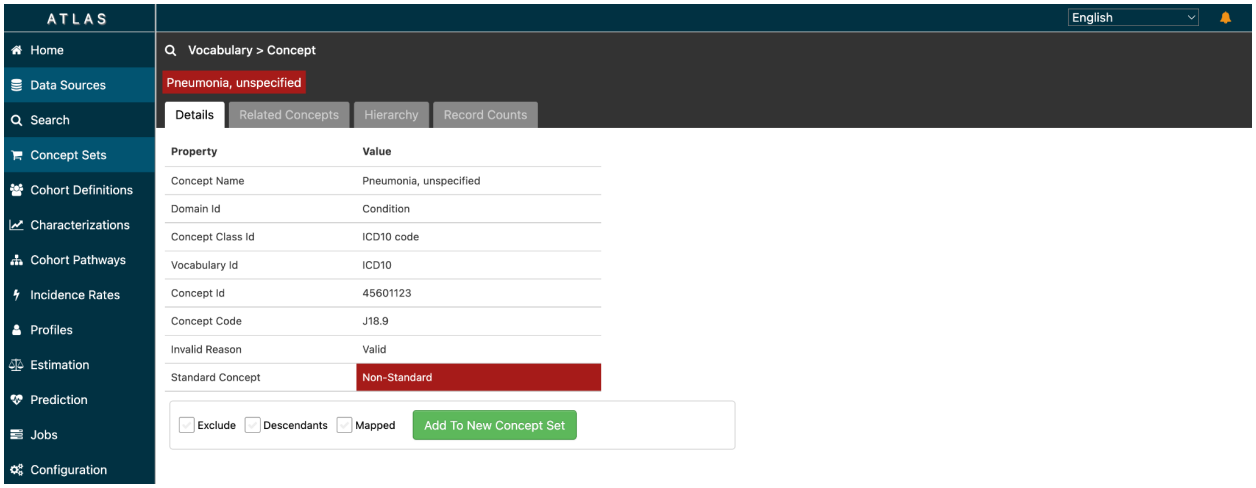
The first tool of interest is known as ATLAS, a web-based tool developed by OHDSI for the purpose of working with the OMOP concept hierarchy and developing concept sets. A publicly-available installation is accessible at https://atlas-demo.ohdsi.org/

For our purposes, let’s immediately visit the Search tab on the left menu bar, and search for J18.9, the ICD10 code for Pneumonia.

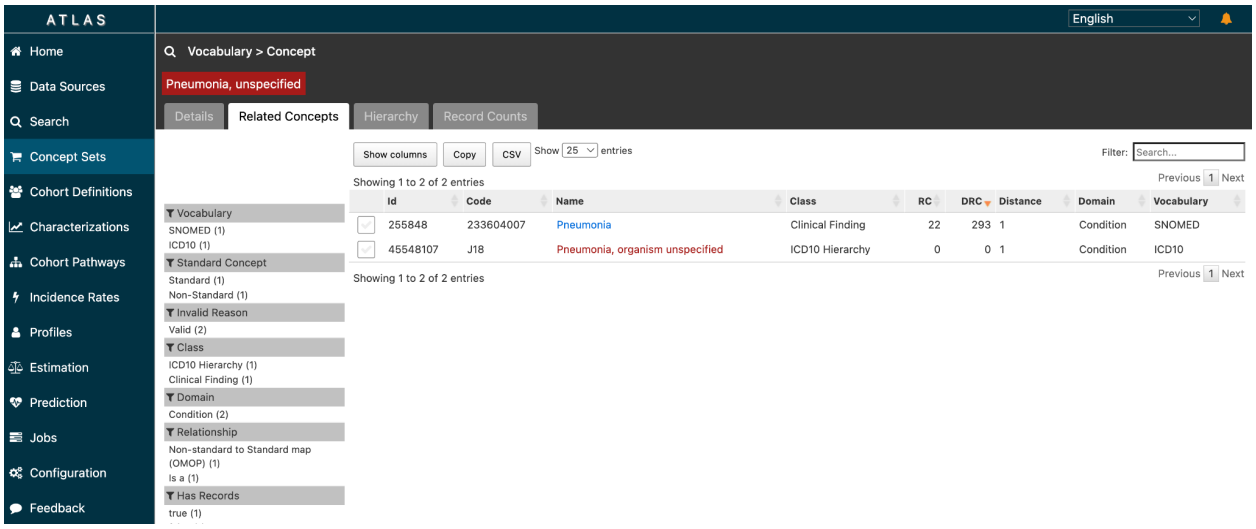


In the screenshot above, we find several matching variations from ICD10 and related vocabularies. According to the legend in the lower-left, they are colored red because these are not standard concepts. The **Id** column shows the OMOP-specific concept ID, while the **Code** column shows the source vocabulary codes. The **Domain** column gives a hint at what OMOP tables these concepts can be found in, in this case we'd look in the **condition_occurrence** table of EHR records for potential entries. The entries on the left let us filter the matches in various ways.

Because the matches aren't standards, we will click on one to see more information about it. The 5th entry seems to be the most generic, ICD10 J18.9 version. The resulting view highlights that the concept is not standard, and provides a little more information about it.



Because it's not standard, we can use the **Related Concepts** tab to see what standard concept this one maps to.



Here we see two mappings: one to another non-standard concept, and one to the standard concept colored in blue (which happens to come from the SNOMED vocabulary–many standard concepts do). Many concepts have many mappings, making it difficult to identify the relevant standard. The filters on the left help narrow down to the relationship of interest, especially the "non-standard to standard map" filter.

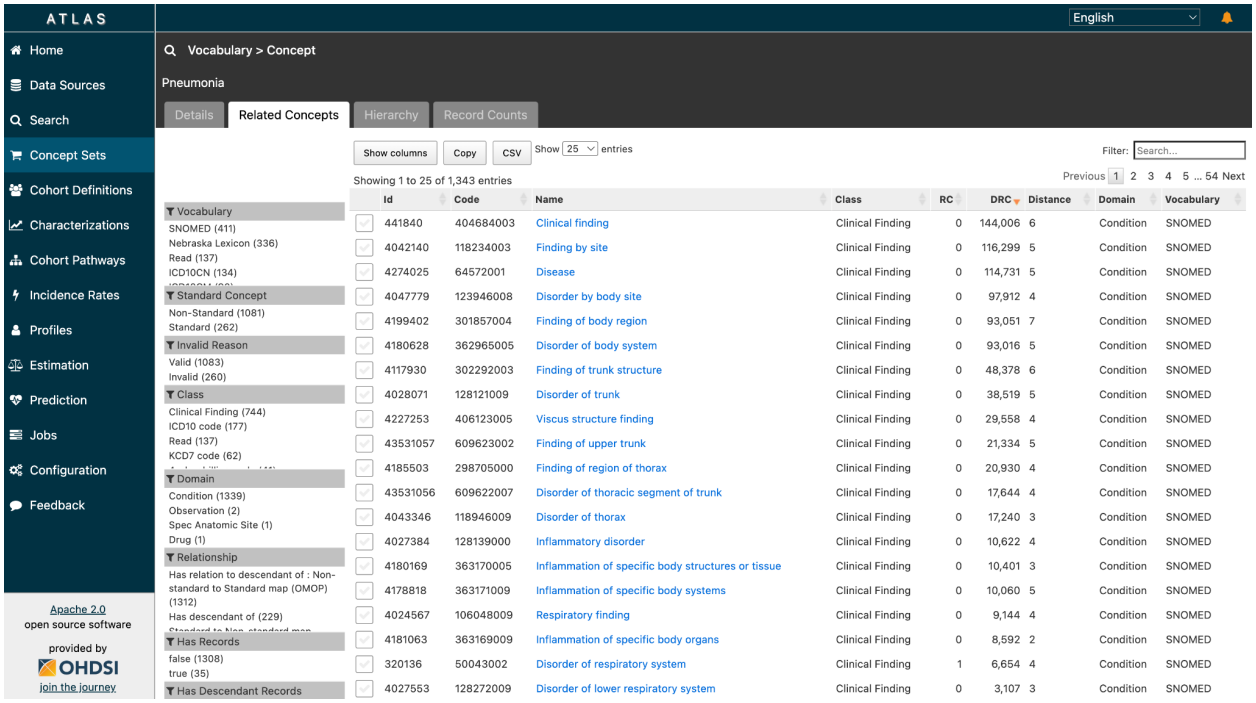
By the way, the RC and DRC columns of ATLAS shown indicated how many records use that code in the database connected to this installation of ATLAS. This is not the N3C database, just a small example OMOP dataset, so these record counts don't mean much.

Exercise

At https://atlas-demo.ohdsi.org/, search for a disease, and investigate a non-standard concept related to it. Find the corresponding standard concept. What is the concept ID of the non-standard concept you chose? What about the standard?

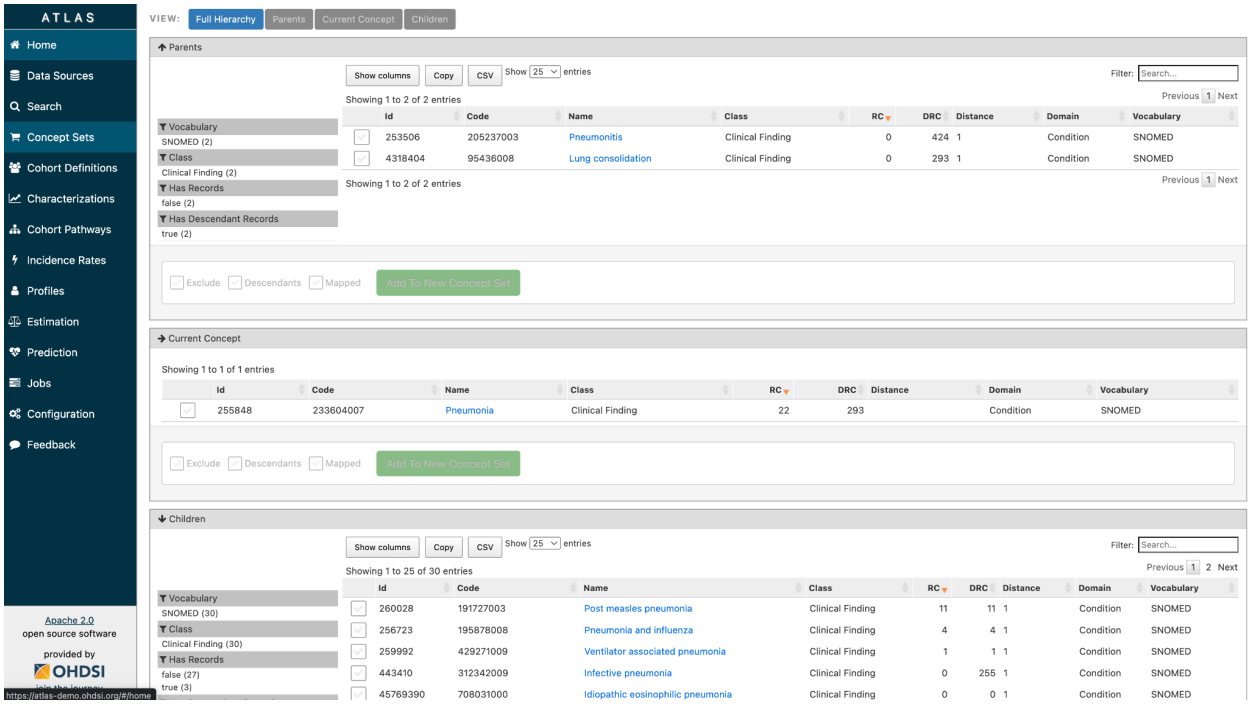
Selecting Concepts to Build the Set

Next, let's click on the Pneumonia standard concept, and similarly navigate to the Related Concepts tab to see its relationships to other concepts.



The result shows a table of 1,343 concepts! Many of these are also standard concepts, colored in blue, and below are many non-standards that also map to this standard concept for Pneumonia.

Now that we are working with a standard concept, we can open up the Hierarchy tab to see what other standard concepts are above and below it in the hierarchy.



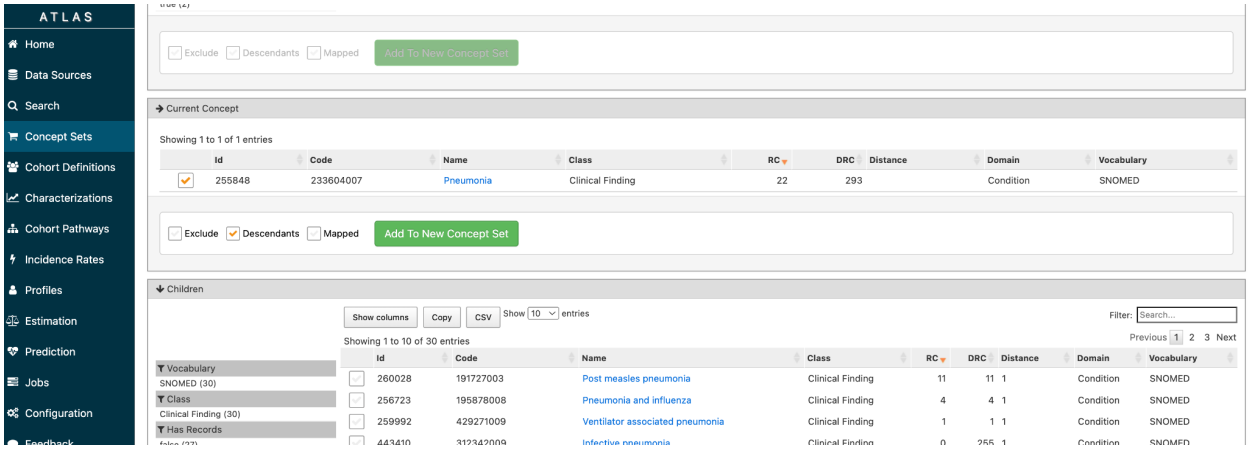
This hierarchy view is a little disorienting at first. It's broken up into three sections: in the middle sits Pneumonia, our current concept of interest. Below that are children of Pneumonia in the hierarchy, and above are parents.

Before continuing further, it's wise to investigate the parent concepts. Perhaps after consideration we aren't interested in types of pneumonia, but types of pneumonitis, one of the parent terms. Or maybe with further thought we want something even more general, in which case we'd open Pneumonitis up and see what *its* parents are in the hierarchy.

We should also review the child concepts. These may suggest narrower concepts better suited to our study, or specific concepts that should be excluded. In the case of Pneumonia, including cases of Ventilator Associated Pneumonia may not be appropriate for some studies. We should then visit Ventilator Associated Pneumonia and explore its parents, to see if there's a more general class like Hospital Associated Pneumonia we should exclude instead.

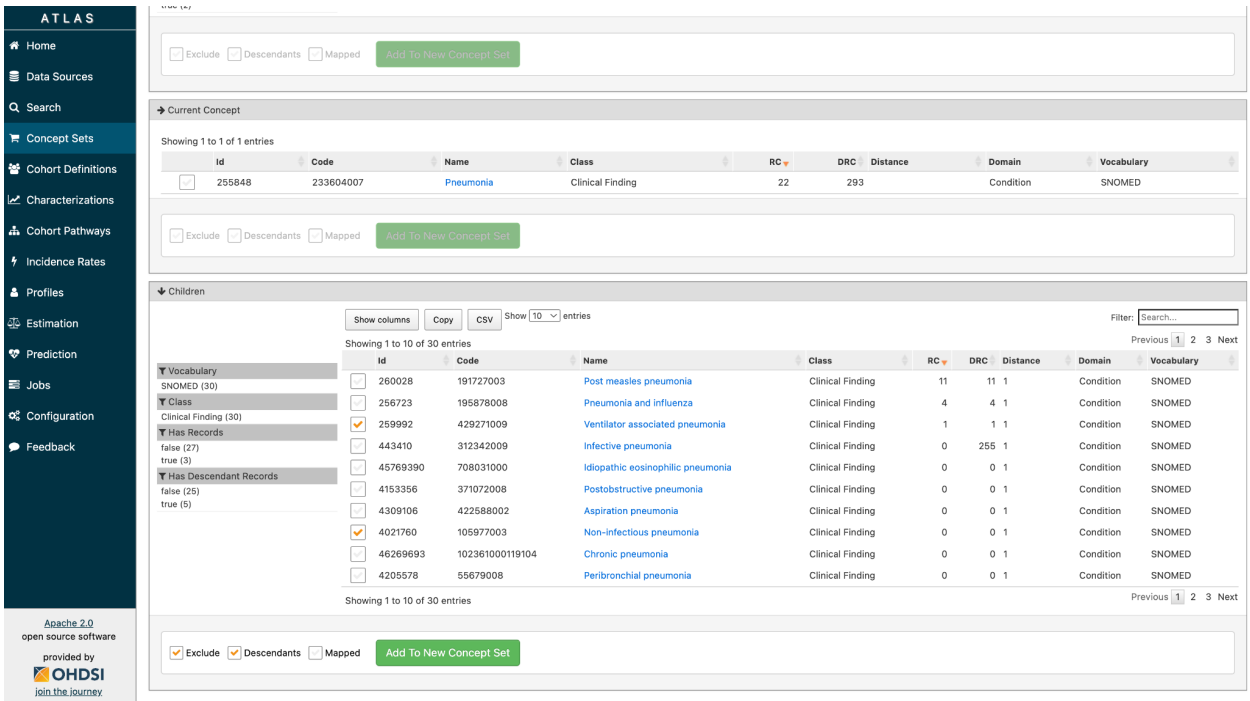
Exploring the "concept neighborhood" is one component building quality concept sets. Another is review of your choices by someone with relevant clinical knowledge. They may identify accidentally included concepts, or suggest clinically relevant concepts from a different part of the hierarchy altogether.

To keep it simple, let's say we want to include all kinds of pneumonia *except* Ventilator Associated Pneumonias and Non-infective Pneumonias. In ATLAS, this takes a couple of steps. First, we are going to include Pneumonia and all of its descendents by selecting the check box next to Pneumonia and then **Descendents** in the options. When we then click on Add to New Concept Set, the term and descendant concepts are included in a new draft concept set. (Not checking Descendants includes only the selected concept in the set.)

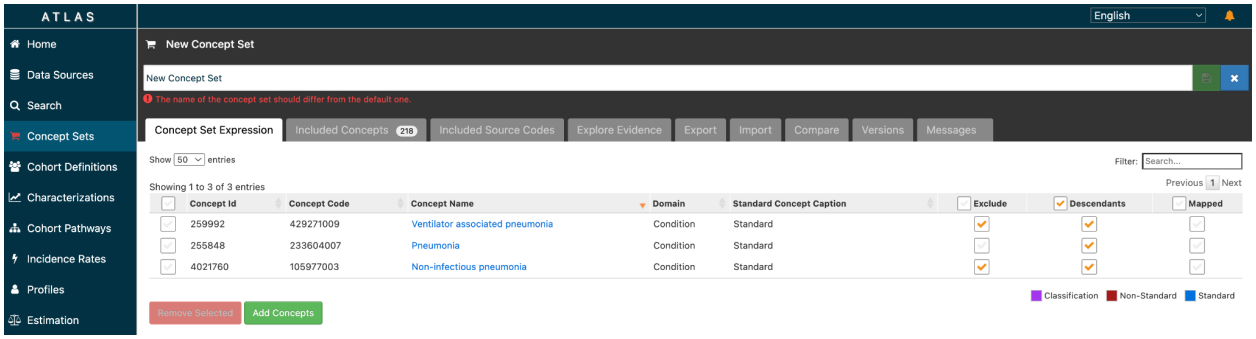


The **Mapped** option causes the concept set draft to also include all of the non-standard concepts that map to the standard concepts included. Since the columns we will be working with in the data (e.g. **drug_concept_id** in the **drug_exposure** table) should only contain standard concepts as a result of the data ingestion process, we won't select this.

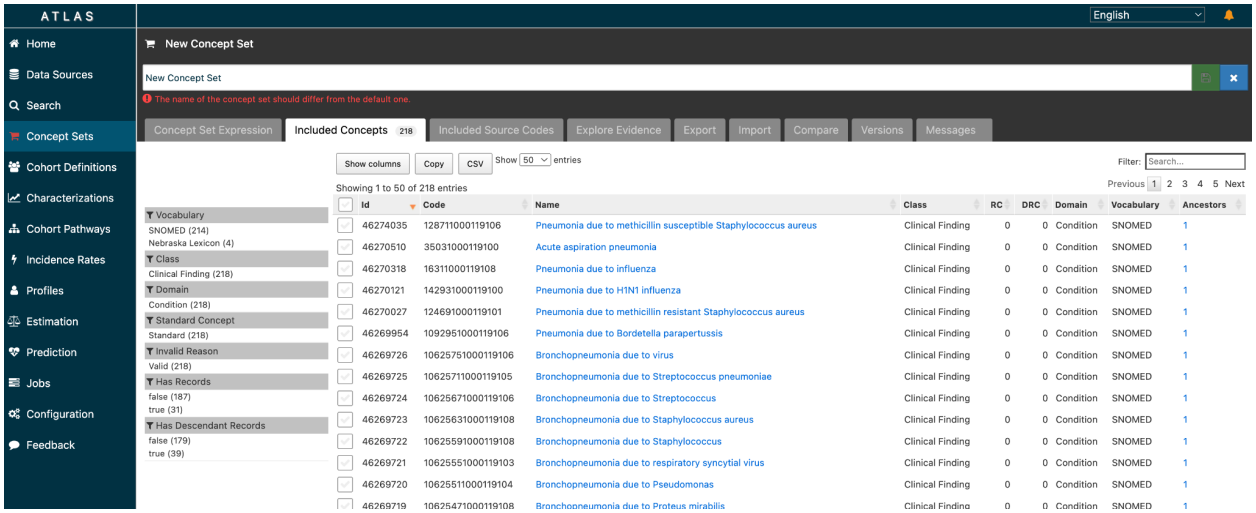
Next, to exclude Ventilator and Non-Infective Pneumonias, we select their check boxes and in the options choose Descendents *and* Exclude. This identifies these and all descendent concepts as those that should ultimately be excluded in the draft set, clicking "Add to Concept Set" notes them as exclusions.



With our draft concept set in progress, we can visit the **Concept Sets** menu item to see our progress.



The three additions we made are called “expressions.” Note that the order of the expressions doesn’t matter - the concept set generation first includes concepts from all the inclusion expressions, then removes those specified by the exclusion expressions. Clicking the **Included Concepts** tab shows the full list of concept IDs defined by the expressions. This is a good list for someone with relevant clinical expertise to review for unexpected concepts.

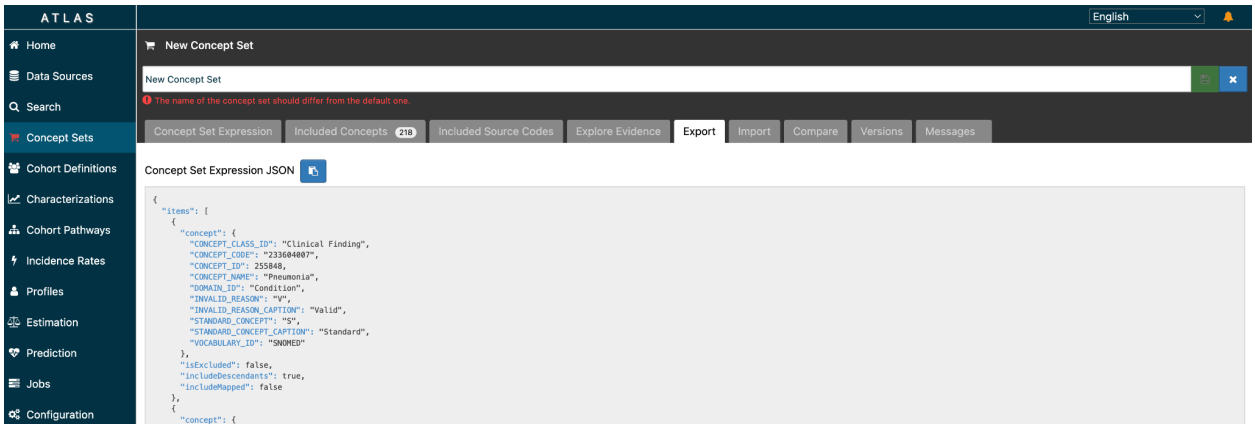


Exercise

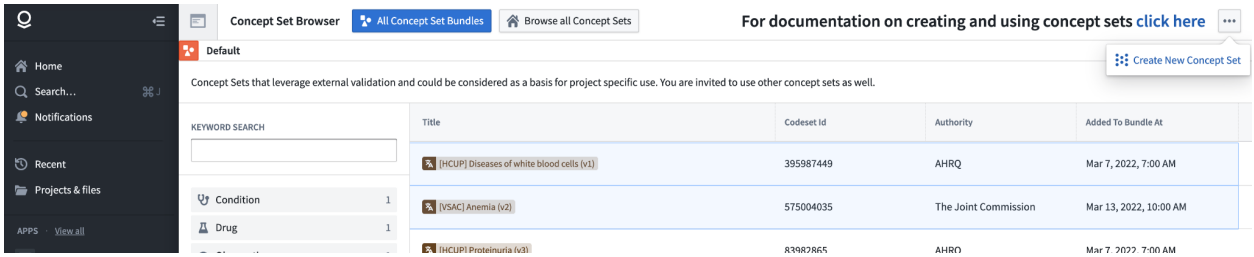
At https://atlas-demo.ohdsi.org/, build a draft concept set including at least one standard concept and its descendants, and excluding at least one child (or grand-child) of the concept and its descendants.

Concept Sets in the Enclave

The ATLAS tool is not (yet) available directly in the N3C Enclave. Even though a concept set is at its base a list of concept IDs, there are a number of tools inside the Enclave for creating and sharing concept sets. In order to get the concept set we've developed into the enclave, we'll need to export it from ATLAS. This can be done under the **Export** tab, where we an click the button next to **Concept Set Expression JSON** to copy the information we'll eventually need to paste into the Enclave.

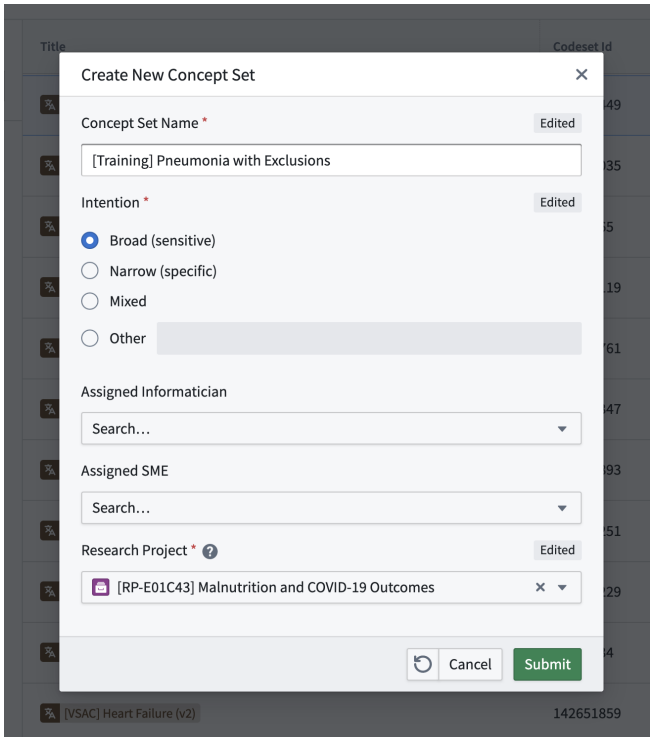


To create a new concept set from our ATLAS definition, we can first open the Concept Set Browser (linked from the Enclave homepage) and click **Create New Concept Set**, found under the … button in the upper-right.

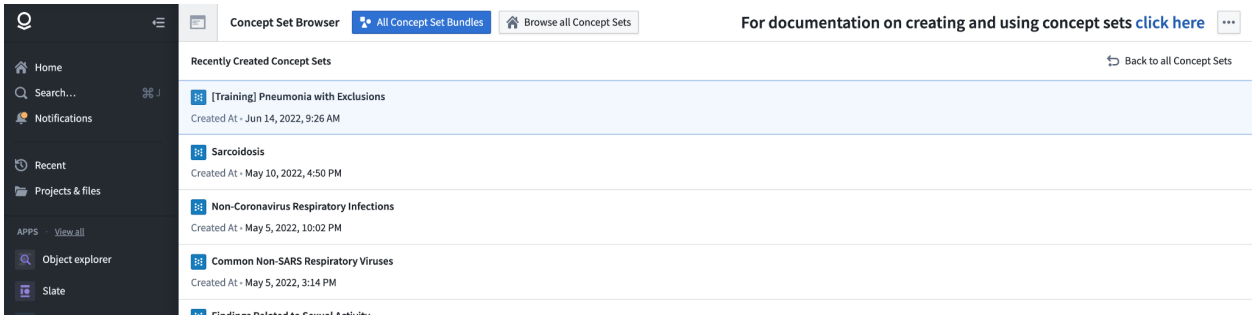


This will begin a step-by-step process to create a new concept set (as an object in the Enclave), that we can use and share with others via the concept_set_members table. The first set of questions asks for a name–it's common N3C practice to prefix names with some tag in brackets–intention (whether we intend for this concept to capture lots of data or highly specific concepts), and we must associate the concept set with a research project in the Enclave - here I've selected the Malnutrition project.

We can also optionally set an assigned informatician, the person who we think will be primarily using the concept set, and a Subject Matter Expert (SME), someone with relevant clinical experience to evaluate the concept set. These facilitate tracking of concept set provenance and enabling peer-review within the Enclave.

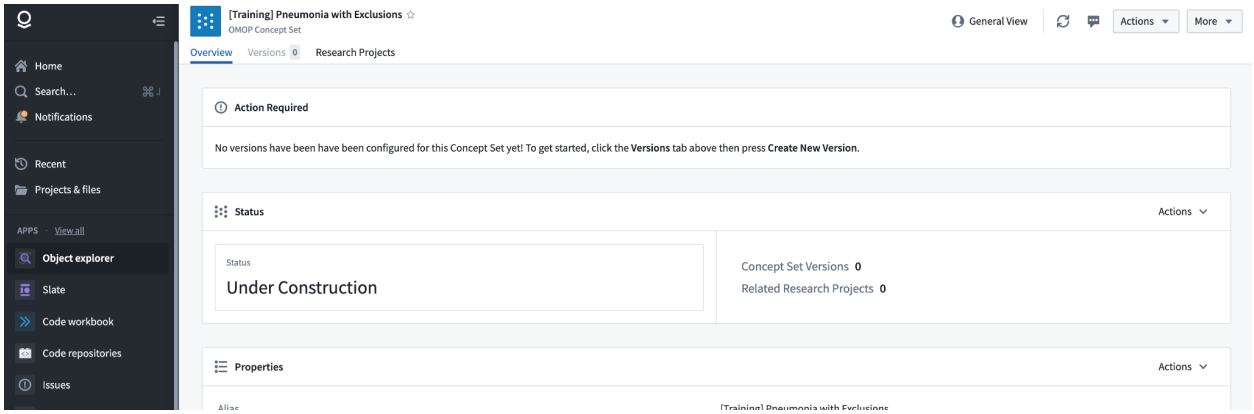


After we click Submit, we are taken to a page which lists all of the most recently created concept sets, including our own.

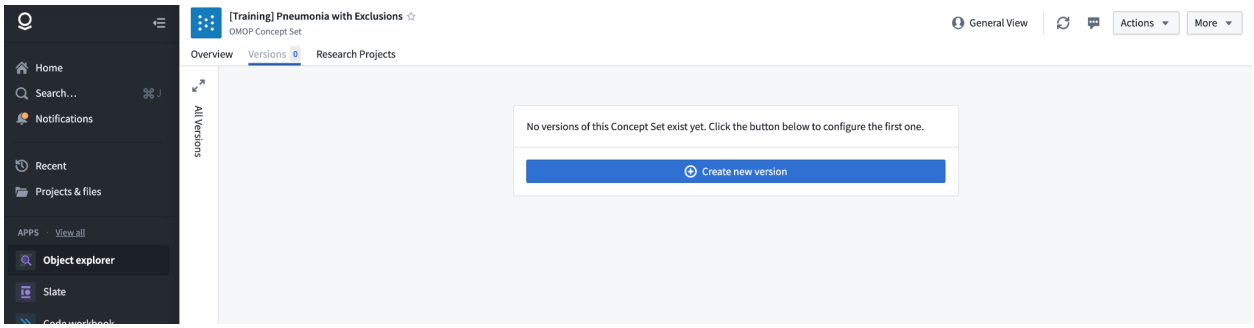


There's a technical limitation in this tool which prevents the system from opening our own new one automatically, so to continue we can click on the row for our new [Training] Pneumonia with Exclusions row.

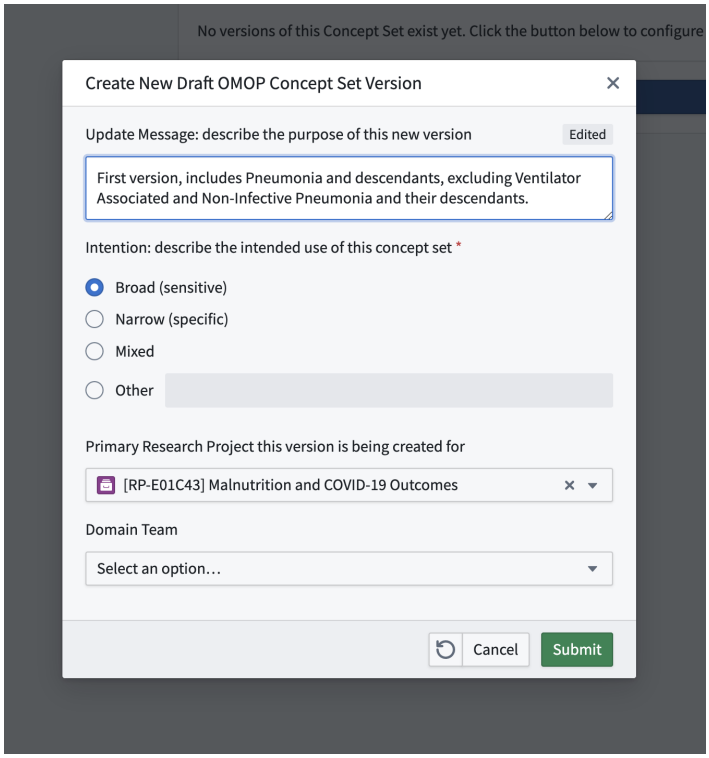
This opens the view for the concept set itself, including a warning that this set has no defined versions yet. Recall that because concept sets are shared with other researchers, it is not possible to edit a concept set "in place," which would affect others' work. Instead, concept sets are immutable (not editable), but can contain different versions.



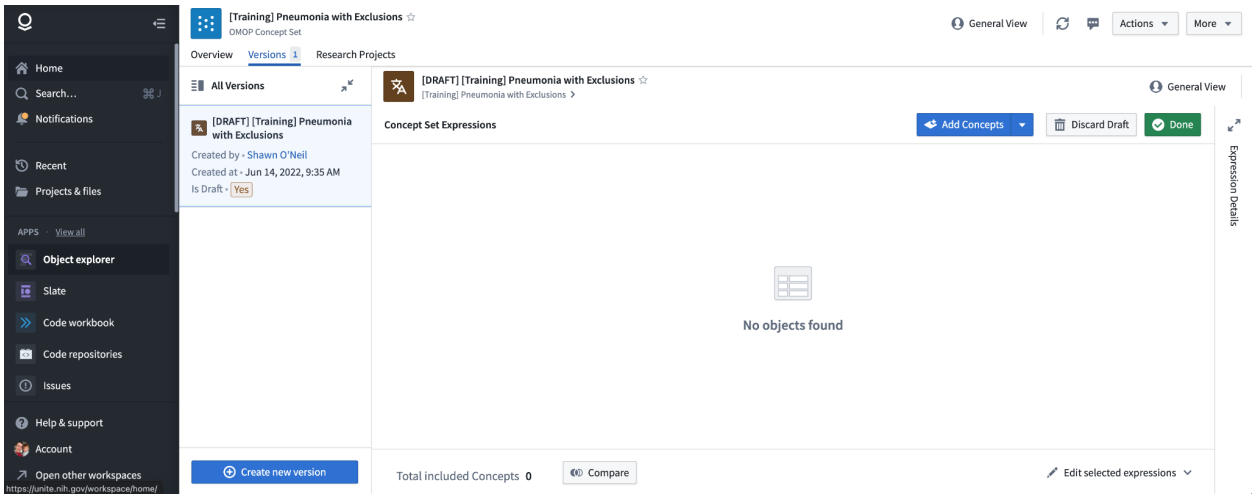
Clicking the **Versions** tab will prompt us to create a new version:



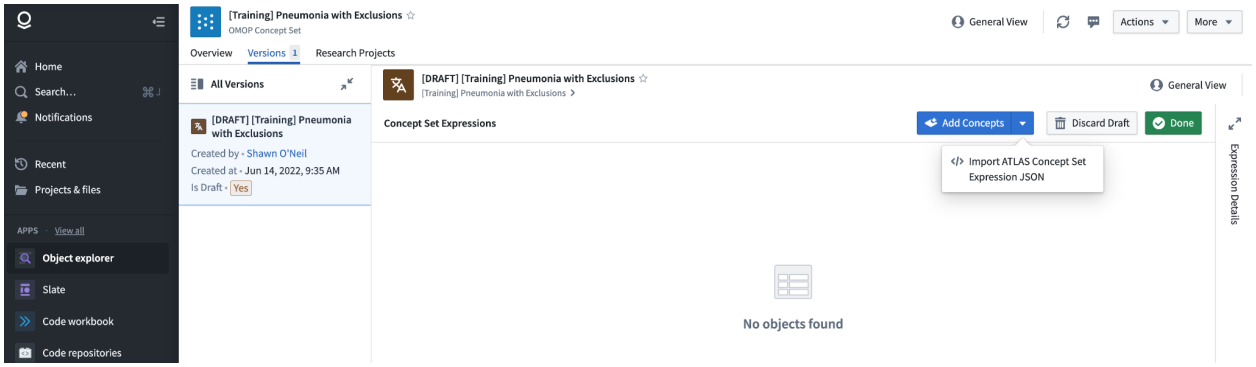
Which prompts another set of questions about the version specifically:



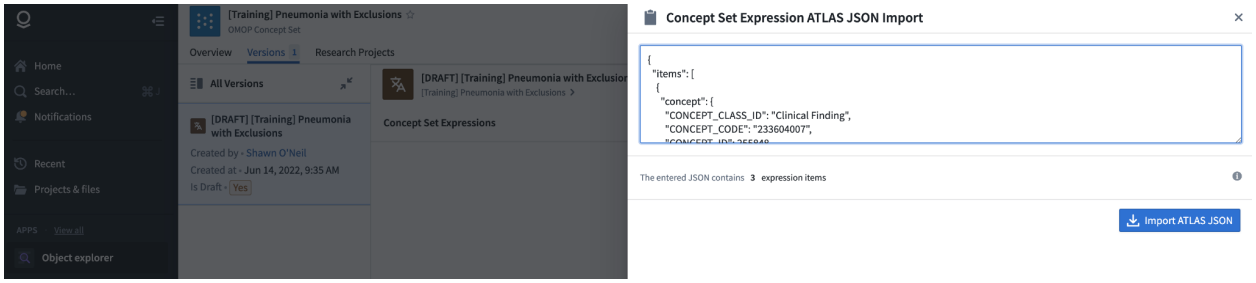
Now, finally, we have started the process of creating a draft concept set version inside the Enclave. Like all Enclave interfaces, there are many buttons and sidebars to pay attention to. Versions of the concept set are listed along the left, with our current first version marked as a draft. The main section allows us to **Add Concepts** to the set, **Discard** the draft altogether, or click **Done** to finalize the version and publish it along with all the other concept sets, after which this version is no longer editable or removable (though they can be “archived” under the **Actions** dropdown, which hides them in the Concept Set Browser).



To add concepts, we'll start by importing the exported JSON that we copied from ATLAS by clicking the arrow next to the Add Concepts button and selecting the ATLAS import option.

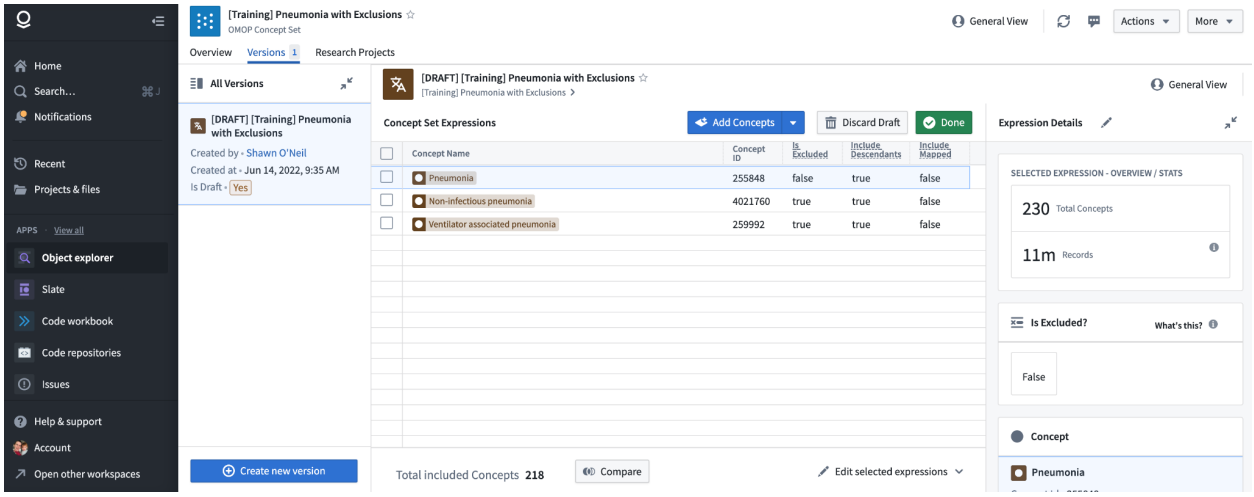


After we do this, a side window pops out where we can paste the previously copied export, and click **Import ATLAS JSON**.

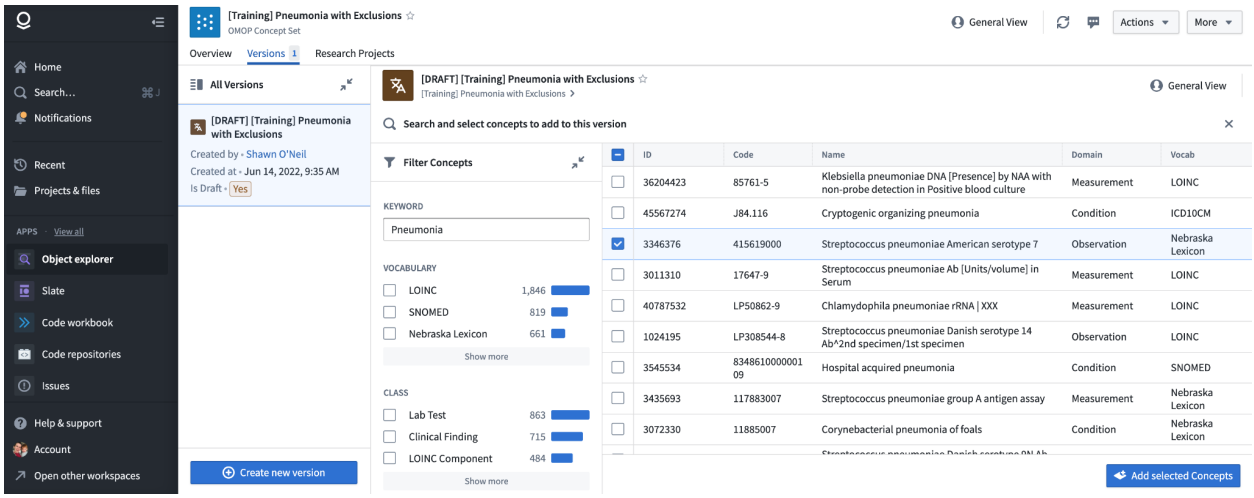


Now our editor shows a view very similar to the ATLAS interface: our expressions are there, and we can see which include descendents and which are excluded, and the total included concepts at the bottom (218). In the screenshot below we've also expanded the **Expression Details** sidebar, which allows us to see statistics about the currently selected expression. Here we see that Pneumonia with descendants includes 230 concepts on its own, with approximately 11 million matches in the N3C data.

This view allows us to edit expressions (to change inclusion/exclusion, descendants, and including mapped non-standards) via a button at the bottom, and we can also click the **Compare** button to see how our draft set compares to other similar sets already defined by others.

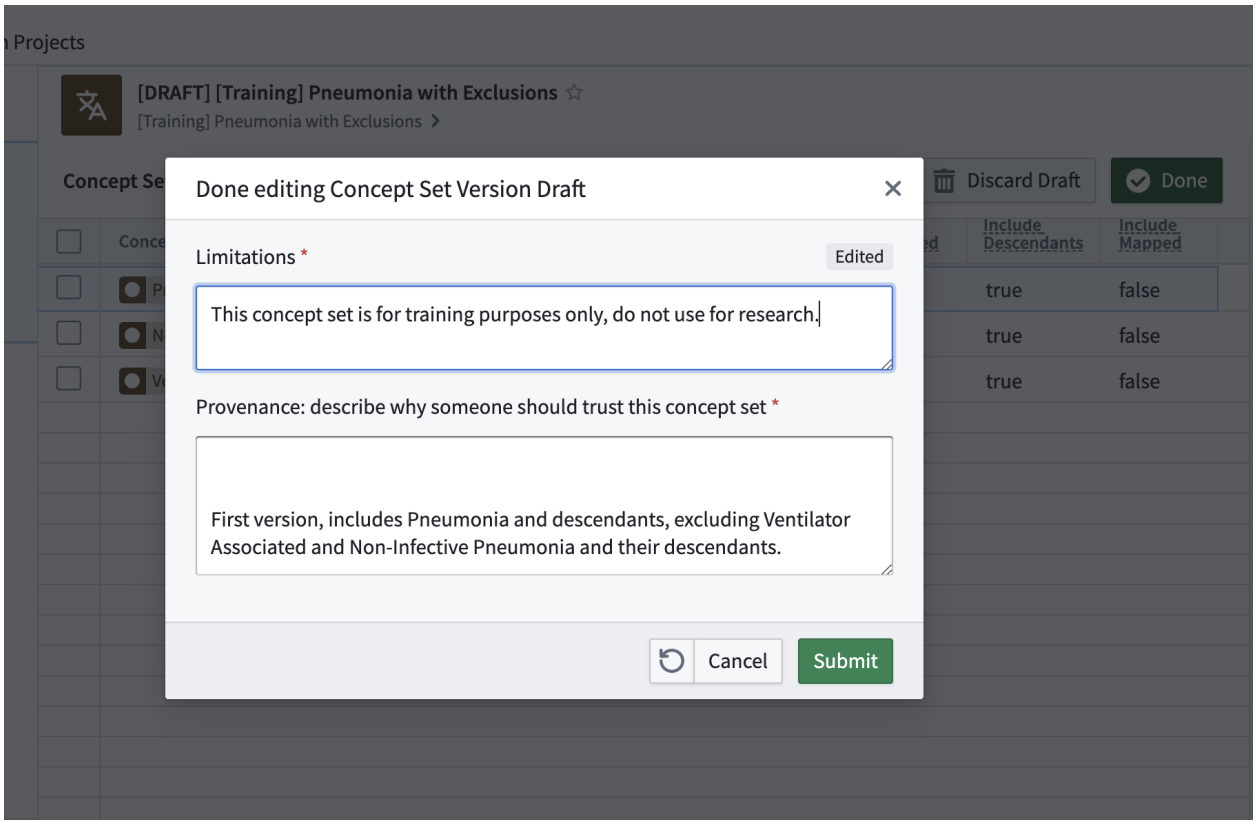


Although we've imported our concept set expressions from ATLAS, it is possible to search for and include concepts directly; the **Add Concepts** button opens a view where we can search for concepts, and then click **Add Selected Concepts** to be prompted as to whether we want to include descendants or not, define this as an exclusion, or include non-standard mapped concepts.

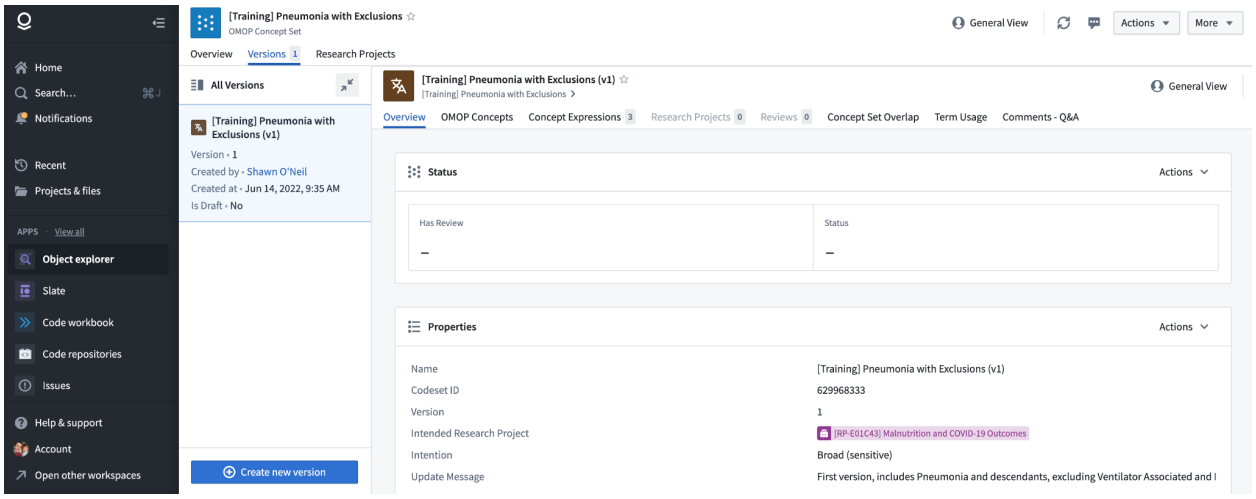


Overall, the in-Enclave concept set editor experience is designed to mimic the process and feel of ATLAS, though ATLAS provides the hierarchy exploration tools more readily in the process.

In any case, we can cancel the inclusion of this individual concept by clicking the small **X** in the upper right, and in the main view click the green **Done** button. This prompts us with more questions, specifically if there are any limitations to this concept set that we want others to be aware of before using it.



When we click **Submit**, our concept set will no longer be marked as a draft in the versions list; it is now version 1 of the [Training] Pneumonia with Exclusions concept set, no longer editable, and given a codeset ID for use in finding the concepts in the **concept_set_members** table!



Note: It takes time, up to 30 minutes, for a concept set to go from finalized in the editor to appearing in the concept_set_members table, and for the full list of concepts to be shown under the OMOP Concepts tab.

Exercise

Create a new concept set and initial version in the enclave for the example you built in ATLAS, and import the concept set. Name your set **[PSC Training] Your Name** (but use your actual name, of course; PSC being short for Pilot Short Course).