National COVID Cohort Collaborative

Introduction to Analyzing Real-World Data Using the National COVID Cohort Collaborative (N3C)

Spring 2024 - N3C Education and Training Domain Team

Session 2, Assignment 2

Concept Set Creation and Basic Us

DUE: Feb. 1, 2024

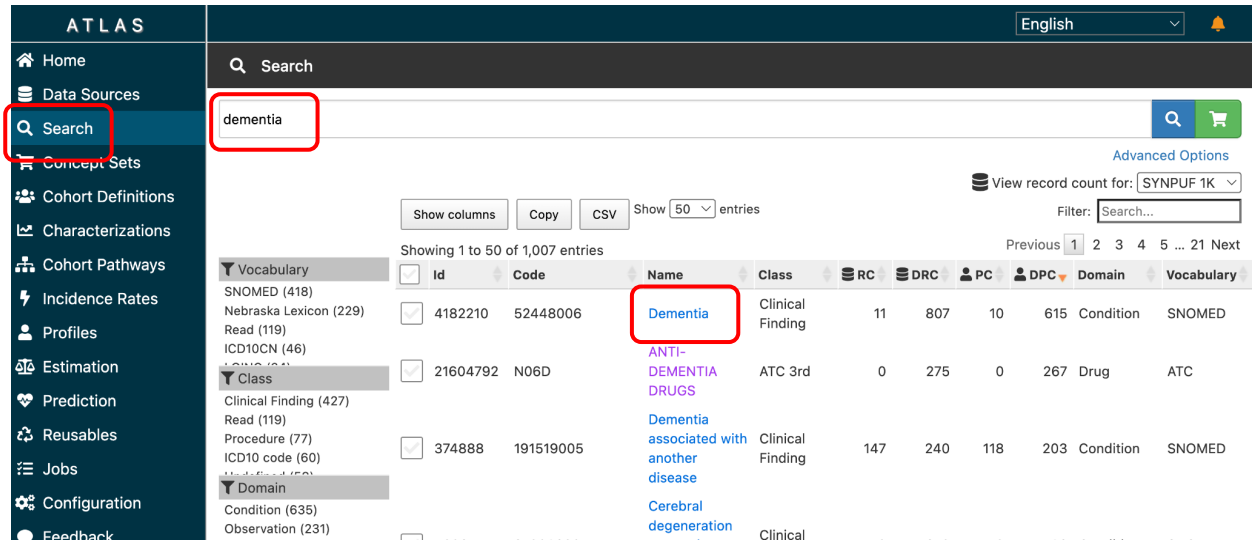
Pre-requisites

To complete this assignment, you will need to:

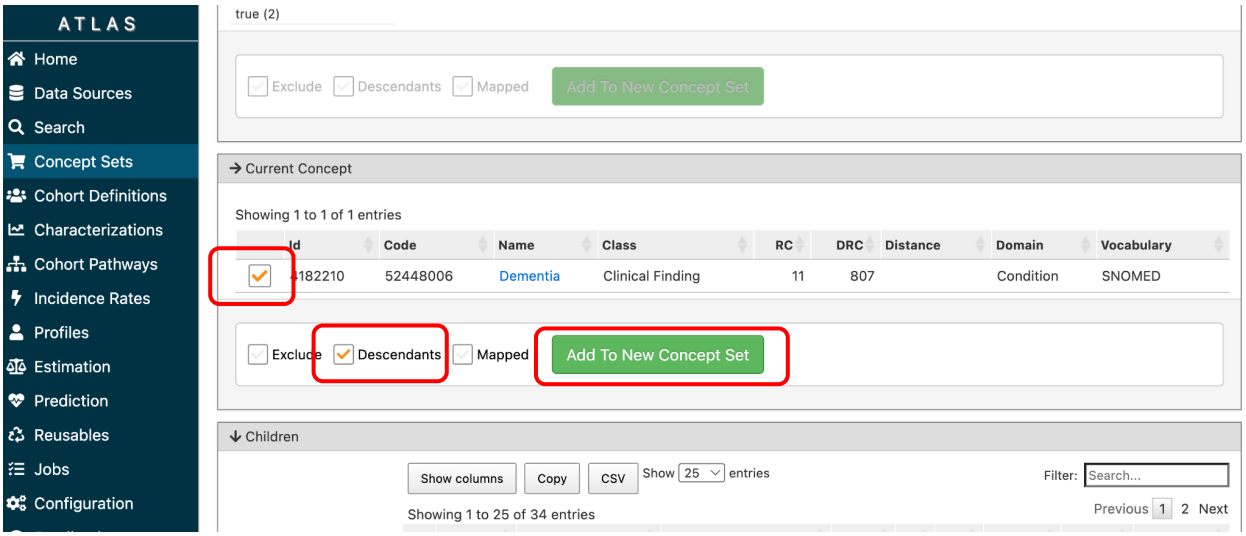
- Be able to log into the N3C Enclave at https://unite.nih.gov

Part 1 - Creating a Concept Set in ATLAS

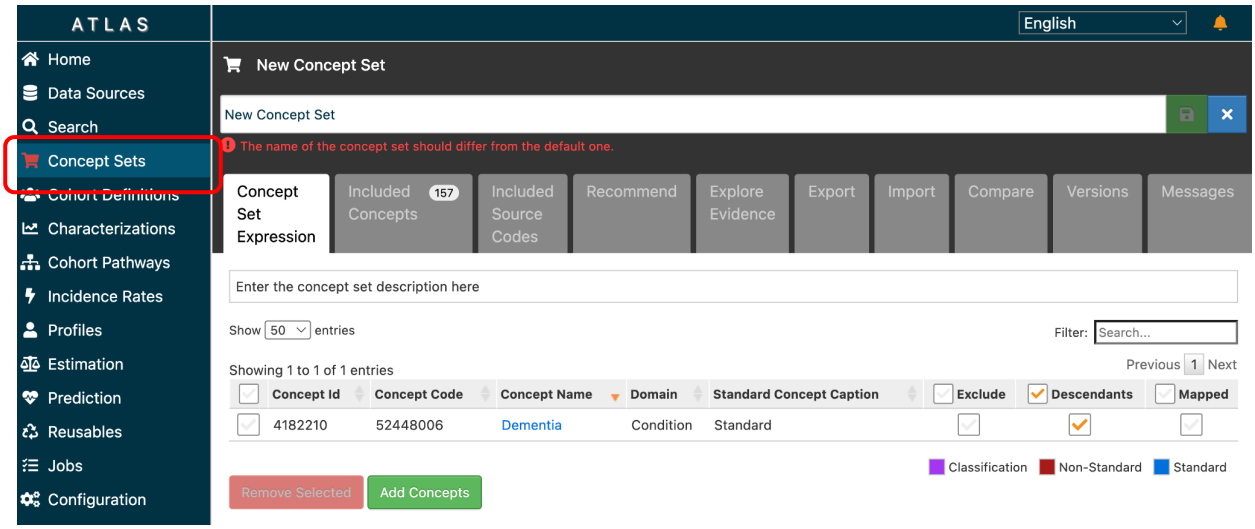
For this assignment we are going to create a simple concept set capturing various types of Dementia, with an exclusion for Huntington's chorea, which is under Dementia in the OMOP hierarchy. First, navigate to https://atlas-demo.ohdsi.org, and in the Search box, search for Dementia, and open the standard concept (OMOP concept_id 4182210).



In the resulting view, navigate to the **Hierarchy** tab, and scroll to find the Current Concept; click the checkbox next to it, and the **Descendants** checkbox, and then **Add To New Concept Set**.

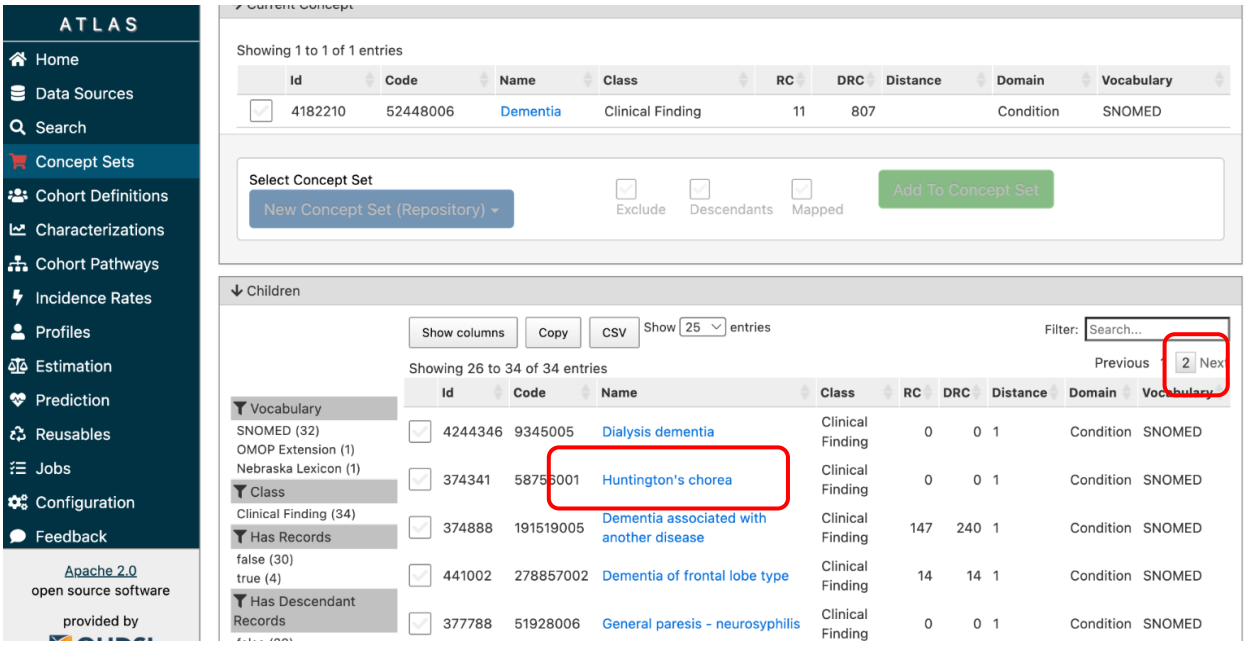


This will add the concept (and it's descendants) to a new concept set, which we can view by navigating to the Concept Sets menu item, which now has a red shopping cart icon (to indicate there's a new, unsaved concept set).

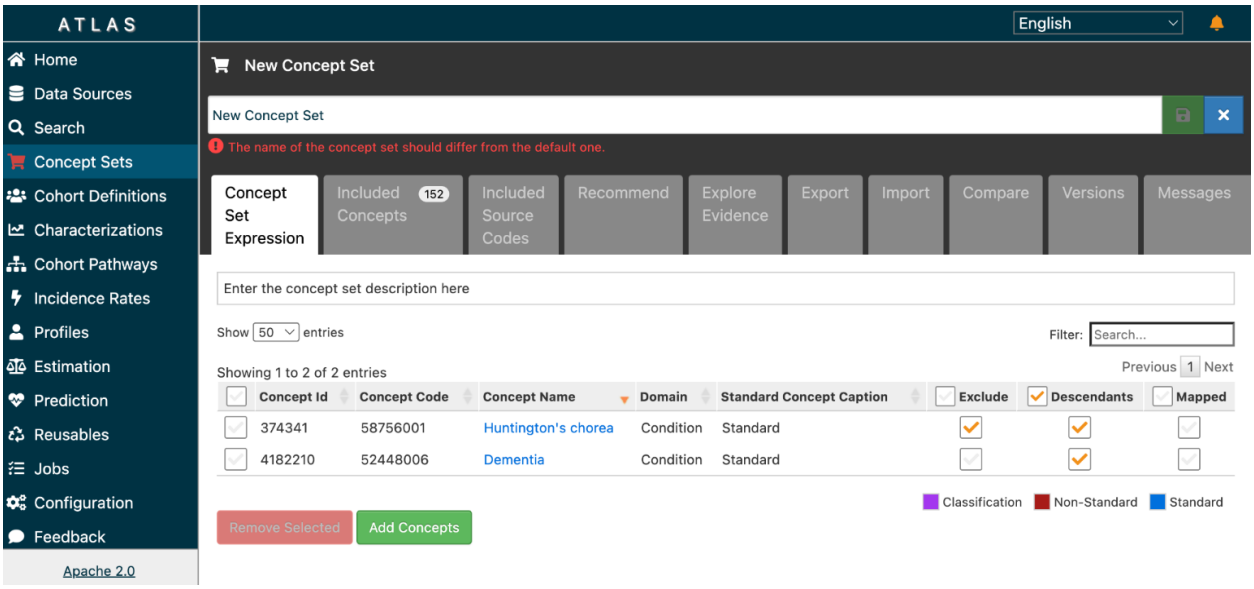


Self check: how many concepts are included in this concept set so far? What if you select the Mapped checkbox in this view? What does that number represent?

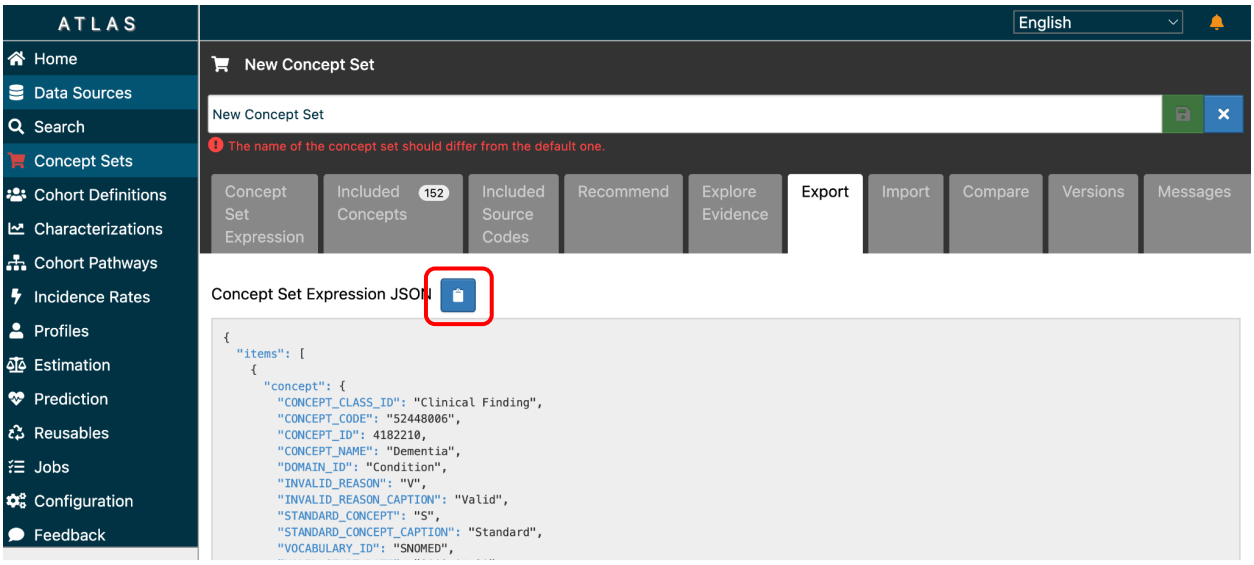
Next, re-open the Dementia concept, and navigate through its children in the hierarchy tab to find Huntington's chorea (it's on the second page of the children list).



Next, add this concept to the concept set, by selecting it and choosing “Descendants” and “Exclude” to exclude Huntington’s and its subtypes from the concept set. Once you’ve done this, navigate back to the Concept Set view to see the updated expressions in the set. It should look like so:

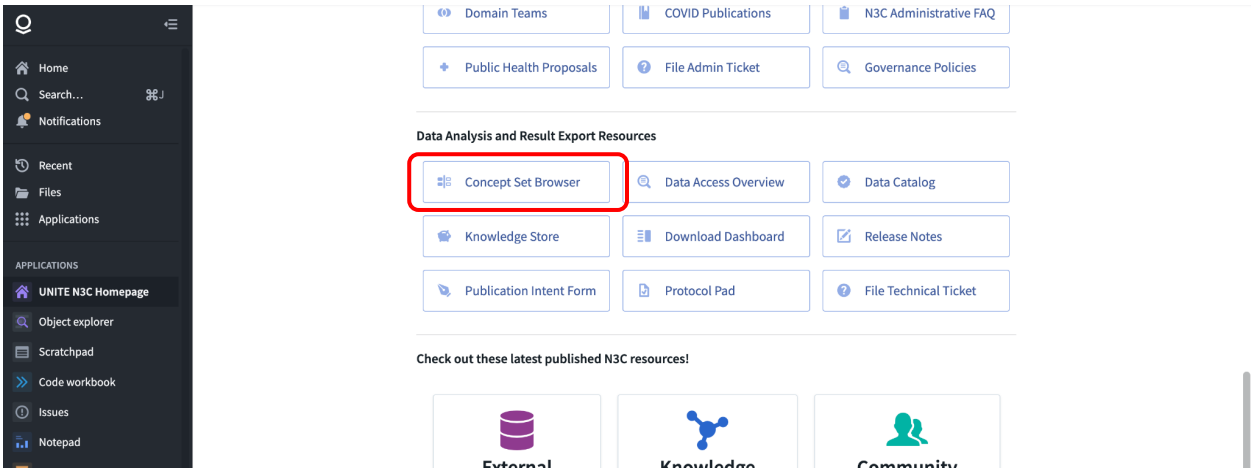


Finally, open the **Export** tab of this concept set; we aren’t going to save it in ATLAS, rather, we are going to copy and paste the provided JSON code that defines the concept set into the N3C concept set editor in the next steps. You can use the **Clipboard** icon to copy the **Concept Set Expression JSON** to your clipboard.

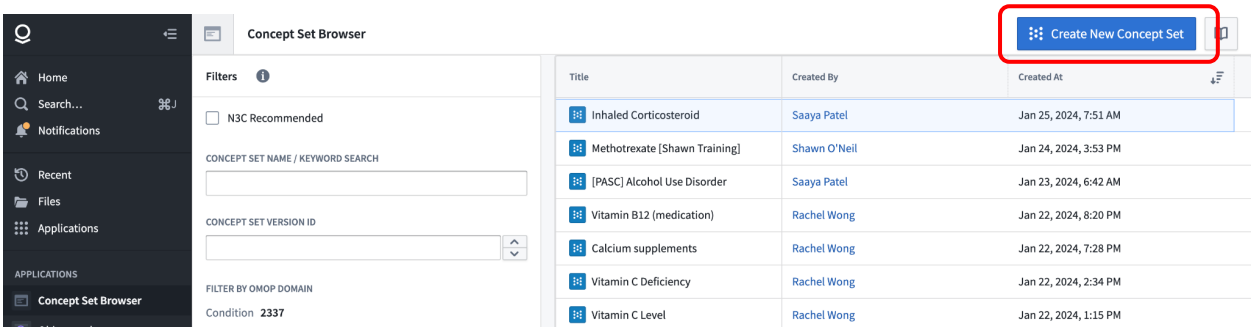


Part 2 - Creating a New Concept Set in N3C

Back in the N3C Enclave, open the **Concept Set Browser** from the homepage:

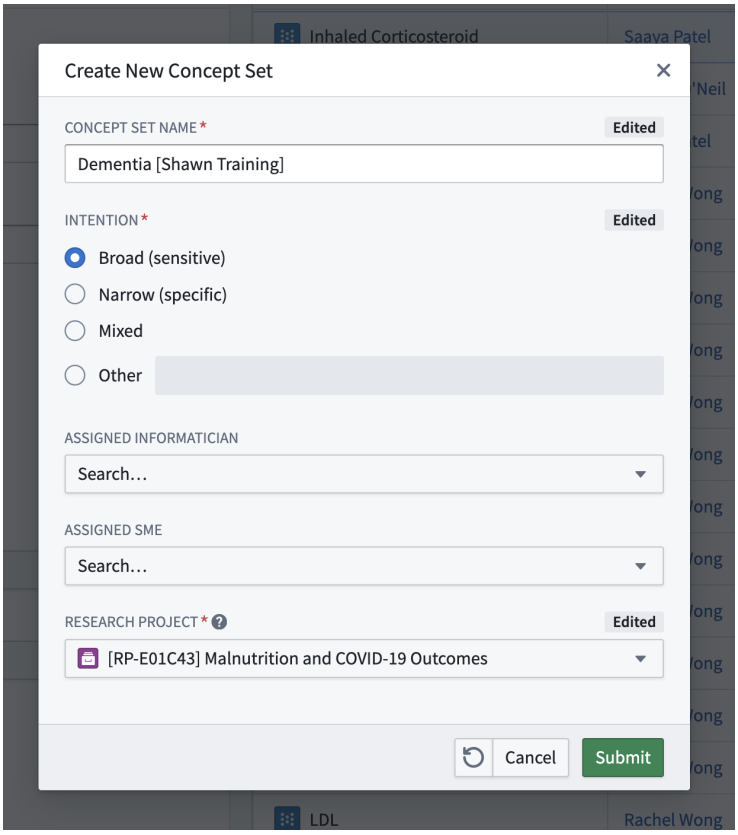


In class we saw how to browse and create new versions of existing concept sets. Let's create a brand new concept set, using the **Create New Concept Set** button in the browser.

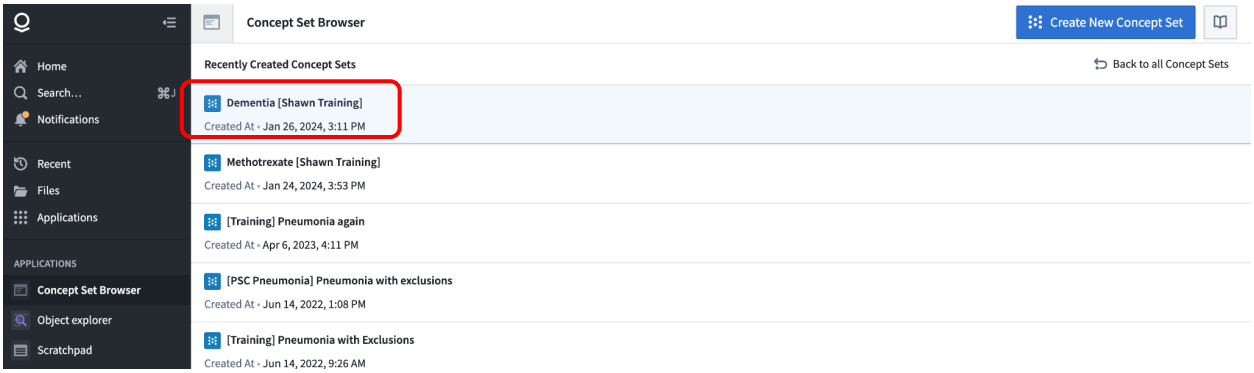


In the popup window, name your new concept set according to the pattern **Dementia [YourName Training]**. Concept set names are freeform; this will help others who might stumble across your concept set realize its purpose. You will also need to select an associated research project for the concept set. For lack of a better option, you can search for **E01C43** to pull up the Malnutrition project that we'll be seeing later in the course. When done, click **Submit**.

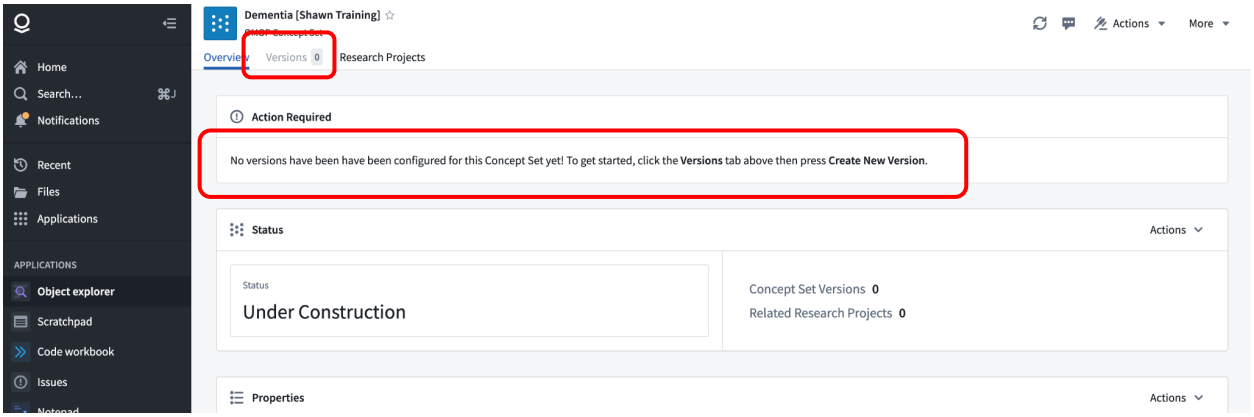
By the way, we are selecting Broad to add metadata indicating we intend this concept set to capture a broad range of concepts. We won't discuss the meaning of all these here - see the [Concept Set Browser Documentation] for details.



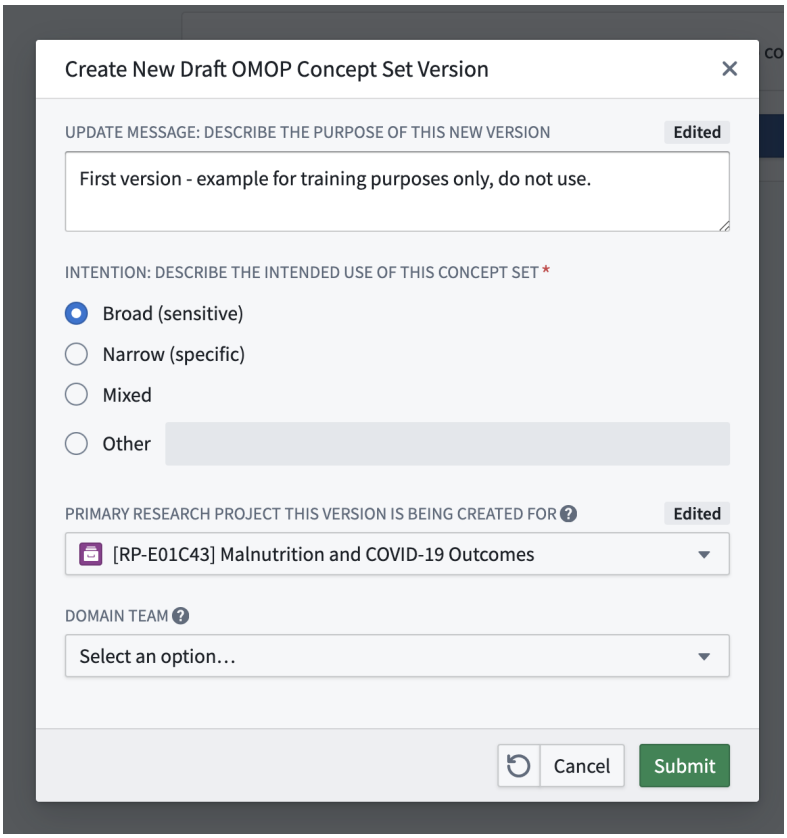
As we discussed in class, N3C concept sets consist of one or more *versions*. At this point, we have a new concept set, but it contains no versions! What we need to do is open the concept set and create a new version (much like we saw in class). To facilitate this, after creating a brand new concept set, the editor takes us to a page listing all recently created concept sets; the one you just created should be listed at the top - open it by clicking on it.



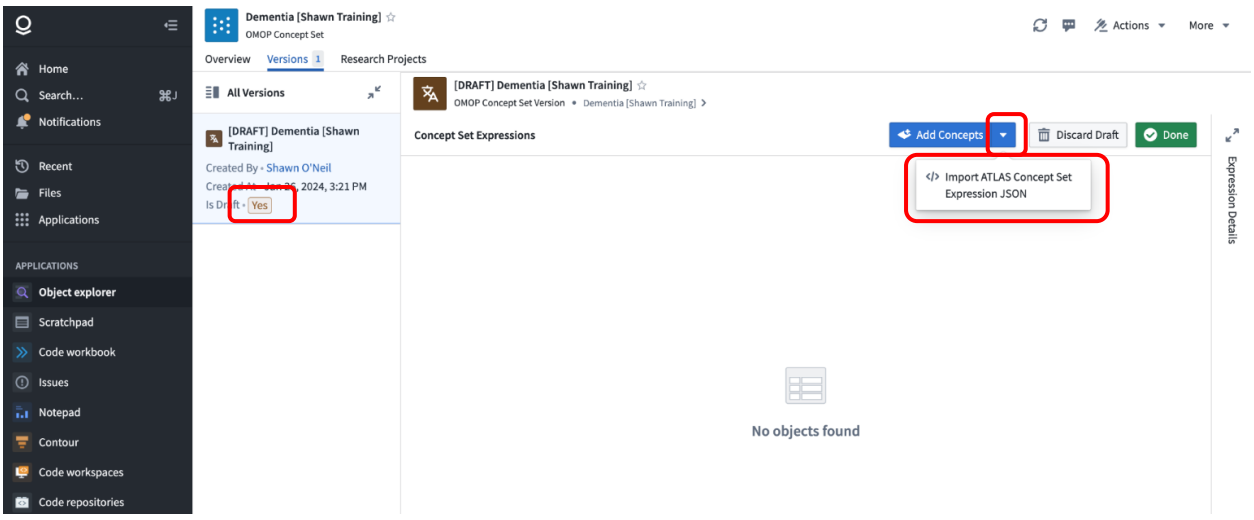
The resulting view of the concept set notifies us that it has no versions, and that we need to open the **Versions** tab to create one.



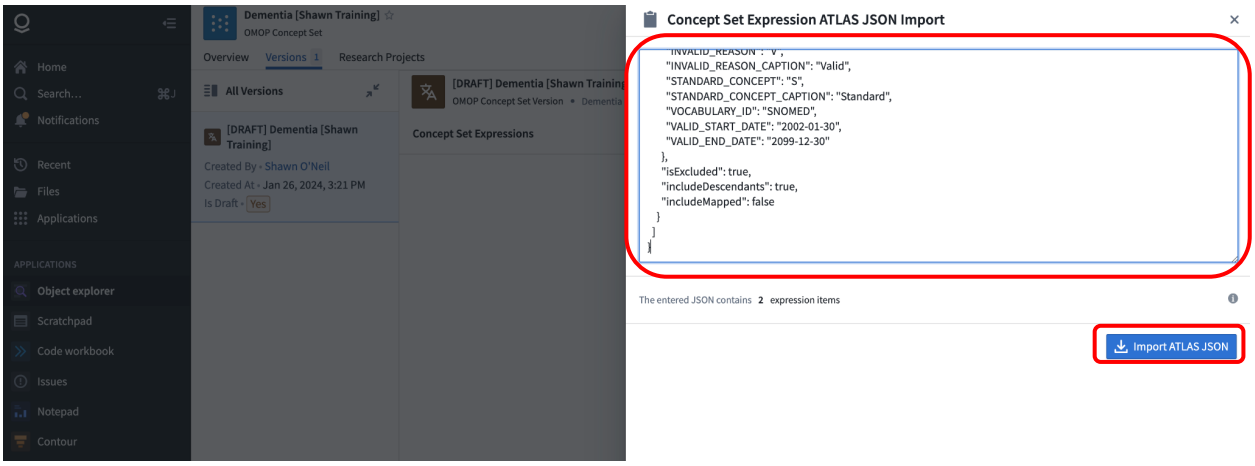
Open the Versions tab, and click the big “Create New Version” button to begin creating a new version of the concept set. This will pop up another dialog, where we are again prompted to enter descriptive information for the version and link to either an N3C project or Domain Team. Enter some text indicating that this is an exercise,



Now we are presented with the interface to add concepts to this new version of the concept set, which is labeled as a draft. Rather than adding them individually, we can import them from ATLAS by selecting the dropdown next to **Add Concepts**, and choosing the ATLAS JSON import option.



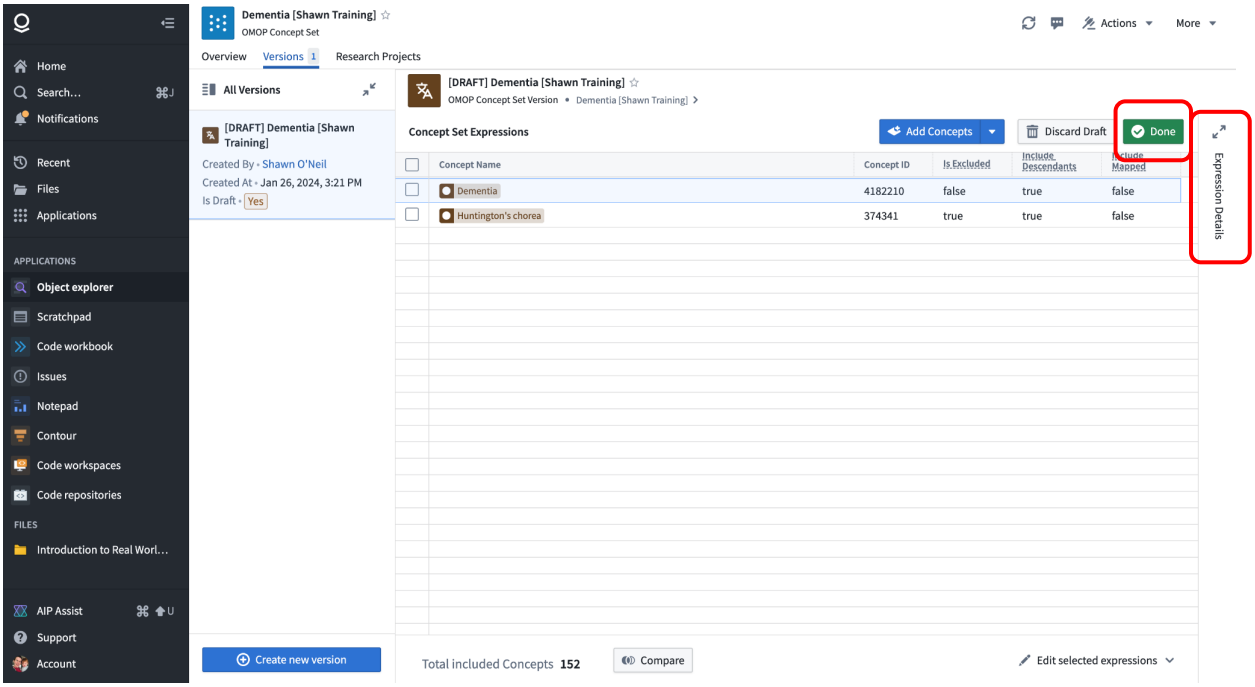
In the resulting dialog, paste the export JSON from ATLAS, and click **Import ATLAS JSON**.



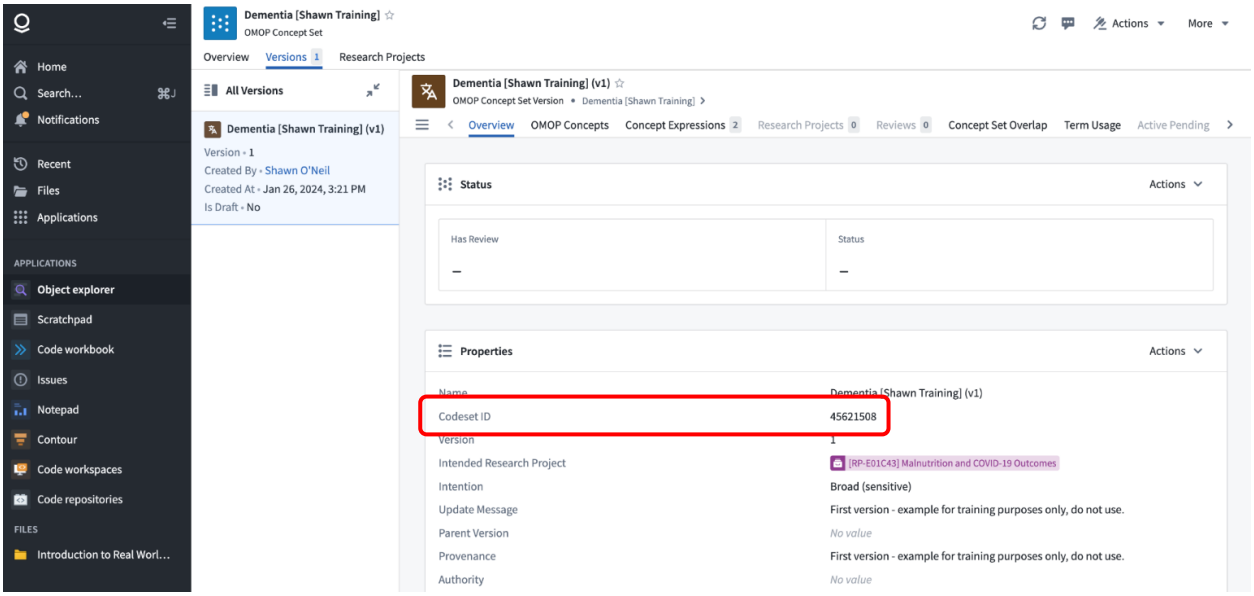
After you do this, the JSON import field will *stay open*. Just close it to get back to a view reminiscence of the expression view in ATLAS. With the Dementia item selected, click the **Expression Details** box to see information about this concept. How many records are covered by the selected concept and it's descendants? (Answer: millions). How many patients does this represent? (Answer: you can't tell, because a single patient might have multiple dementia records) How do these numbers compare to the entry for Huntington's?

While you are here, click the Add Concepts button to search and browse other concepts, noting the record and patient counts reported. Don't add any to the concept set - this is just for exploration.

Finally, to save this version of this concept set, click the **Done** button.



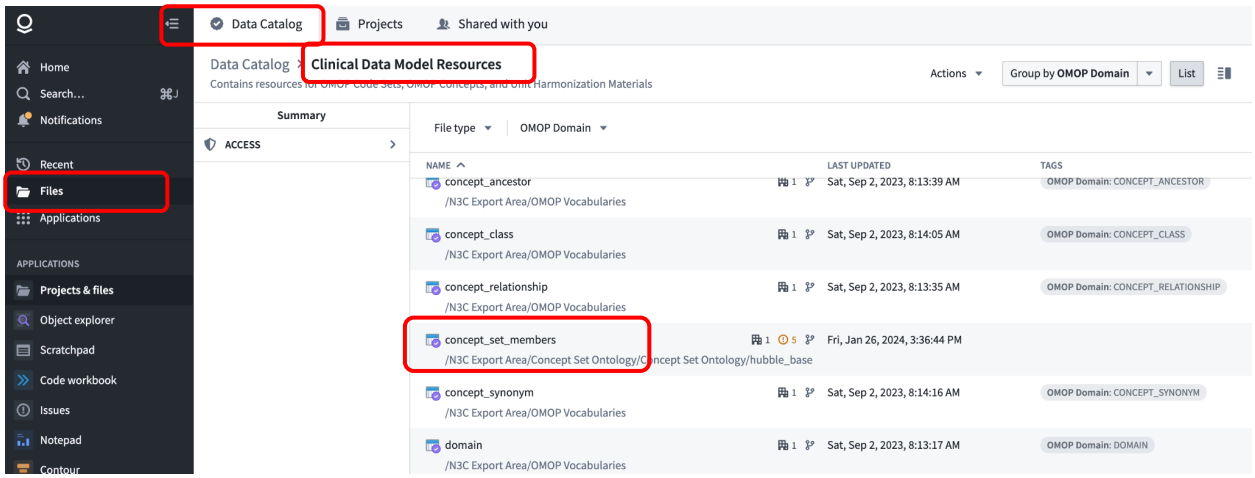
Once done, your new concept set version will be given a **Codeset ID**. Write it down for future use!



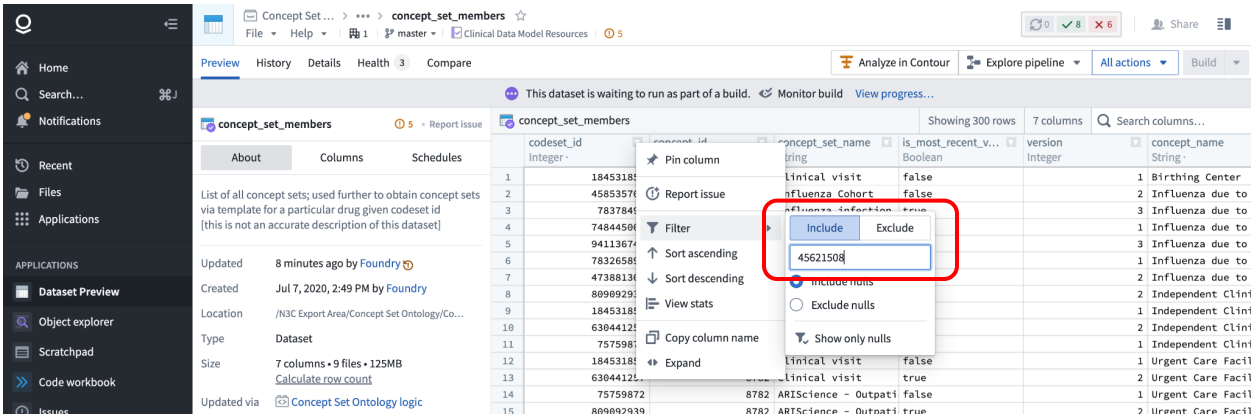
Part 3 - Viewing and Using the Concept Set

NOTE: After you create a new concept set, it can take up to 60 minutes for the set to be processed and made available for this next part.

Now that we've created a new concept set version, and it's been given a codeset ID, we can go and find it in the `concept_set_members` table, which is the *table* that concept set information is stored in. To find it, navigate to the enclave's **Projects and Files**, then the **Data Catalog**, and finally the **Clinical Data Model Resources** collection.



If we open the table, we can then use the dropdown arrow next to the `codeset_id` column to filter to show only rows matching the Codeset ID we wrote down earlier. (Again, this step won't work until the concept set has been fully created in this table, which can take up to 60 minutes.)



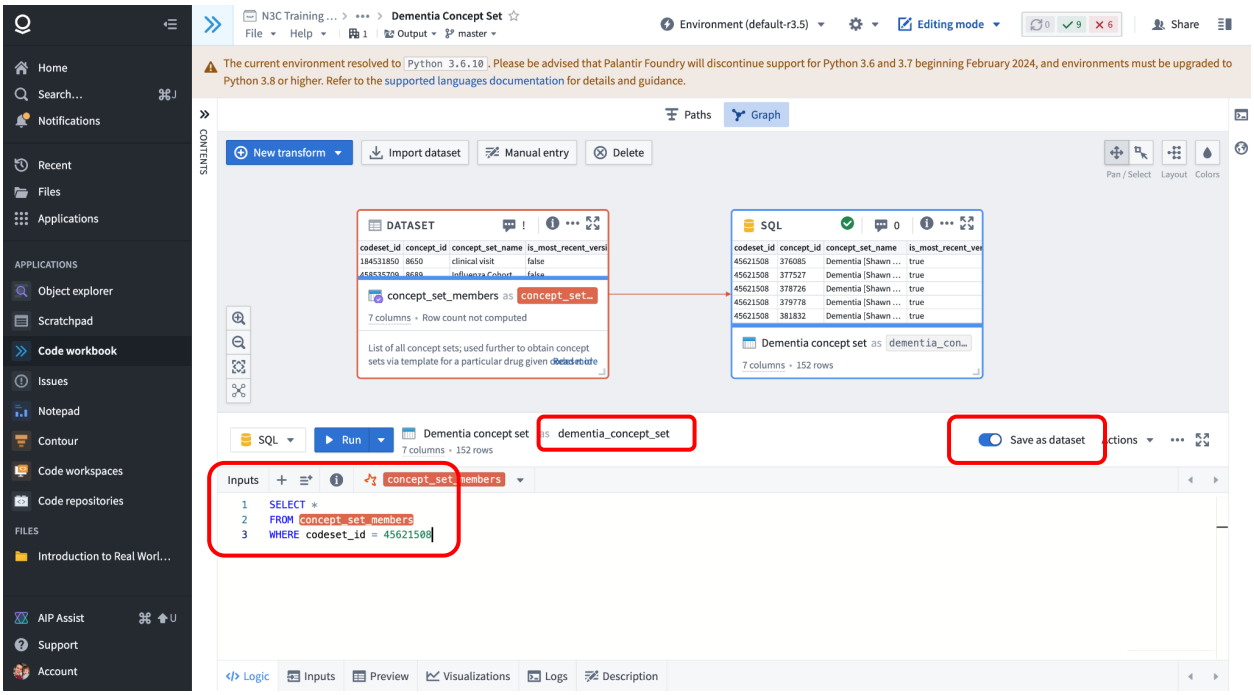
The resulting matching rows show the concepts that are part of the concept set, in tabular form and ready for analysis!

Part 4 - Filtering with a Concept Set in Code Workbook

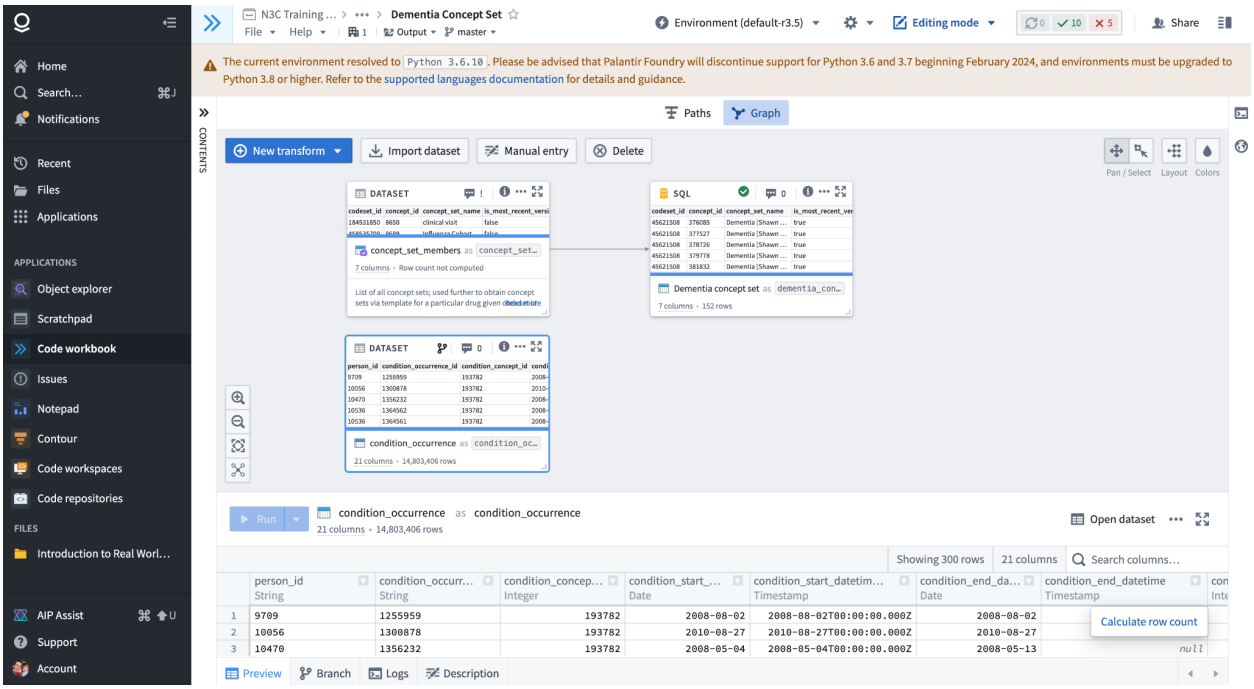
To get closer to an analysis using this concept set, follow the procedures in Assignment 1 to import the `concept_set_members` table into a new code workbook in your training folder; call the workbook Assignment 2 Workbook, and create a transform with the code

```
SELECT *
FROM concept_set_members
WHERE codeset_id = <your_codeset_id>
```

You should also save the resulting dataset via the **Save as dataset** option, and give it a name of **dementia_concept_set** in the workbook. After you **Run** the code for that node, the result should look like this:



Let's take it a step further, and match this concept set against some data. Import the condition_occurrence table from the SynPuf notional (fake) data by using **Import Dataset**, and choosing it via **All → Data Catalog → Patient Data → SynPuf Data (Open Source Synthetic) → condition_occurrence**. Move the imported data node so that it's underneath the imported concept_set_members table:



Now, we need to create a new SQL transform node, from the dementia_concept_set node, that *also* takes the condition_occurrence table as input.

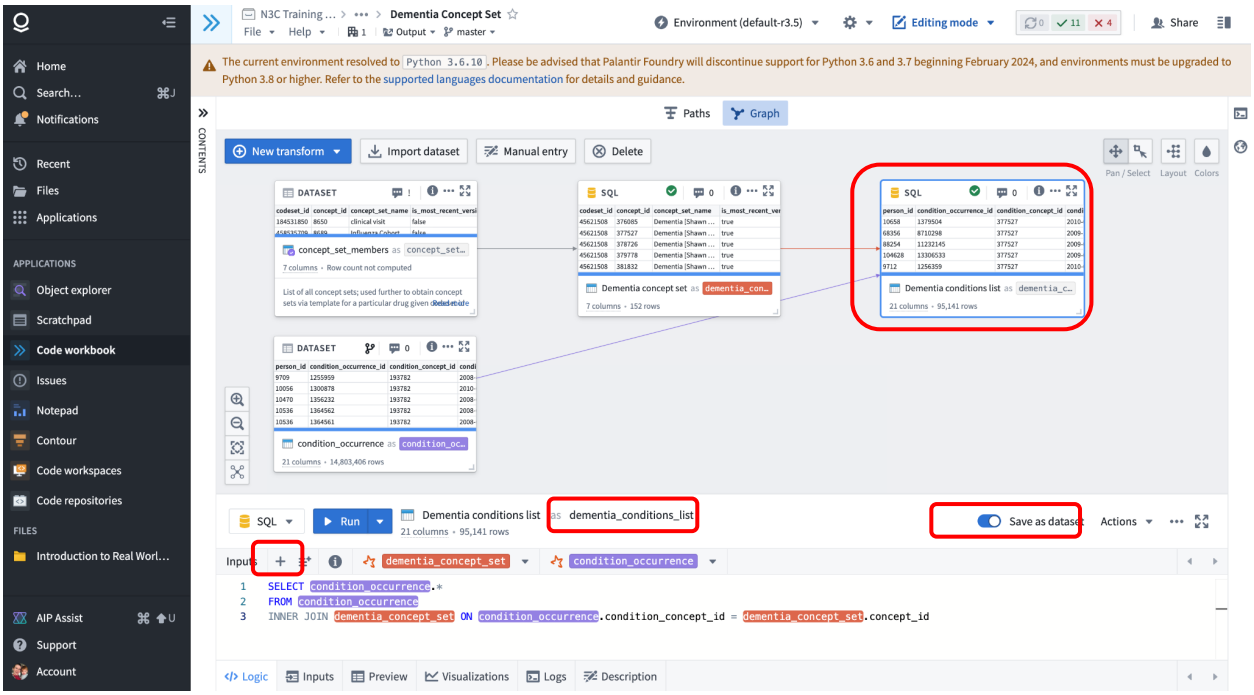
1. Start by creating a new SQL transform off of the dementia_concept_set node.
2. In the resulting new node, select the **+** button near the inputs section to select the `condition_occurrence` table as an additional input.
3. Add the following code to the new node:

```
SELECT condition_occurrence.*
FROM condition_occurrence
INNER JOIN dementia_concept_set ON condition_occurrence.condition_concept_id =
dementia_concept_set.concept_id
```

   This bit of SQL selects all of the columns from the select the condition occurrence table (`condition_occurrence`.*), using an *inner* join against the `dementia_concept_set` table, which results in only rows having entries in both tables in the output. For matching, we are using the `condition_concept_id` column from the `condition_occurrence` table (which contains standard `concept_ids` for the data) and the `concept_id` column of the `dementia_concept_set` table (which also contains standard concept IDs).

4. Finally, save the dataset as `dementia_conditions_list`, and run the transform to see the resulting 95K dementia records in the SynPuf data!

The screenshot below shows the final result, highlighting the UI elements used:



One handy feature of code workbooks is the **Preview** tab at the bottom of the screen which shows the dataset viewer. Try using the drop-downs next to the columns preview to View Stats and answer the following questions on your own:

- What is the most common entry in the `condition_concept_name` column? (This stores the human-readable name corresponding to the concept ID in the `condition_concept_id` column; this information can also be found in the OMOP `concept` table.)

- How many different patients have that most common entry? Hint: try using the dataset preview pane to first filter the `condition_concept_name` column to only include the most common entry, and then use the View Stats feature on the `person_id` column to count the number of "distinct" values, which is shown in the summary stats on the left. I got the answer 19,570.