SCILHS i2b2 PCORnet Common Data Model (CDM) mapping documentation

April 21, 2016

Jeff Klann, PhD

Please refer to our GitHub repository for documentation on installation and use of the ontology. http://www.github.com/SCILHS/scilhs-ontology This document describes the mapping process.

Adding local mappings to the ontology

Some queries will work unchanged (e.g., age if you populate birth_date in the patient dimension), but a majority will require changes to the ontology to represent your local data. We do not anticipate you needing to modify your fact tables (except to ETL data that is not presently loaded - e.g., vitals). Please edit the ontology tables to reflect your local data. Be aware that there will be revisions to the ontology that you will need to incorporate later. We will attempt to minimize changes to each table before you engage in mapping, but this process will require some iterative development.

Important mapping rules

There are a variety of ways to map, and as long as you follow this rule, you will remain compatible with all of our upcoming milestones.

- ***Do not ever change the c_fullname or the m_applied_path of any element in the supplied ontology.***

Also, another important rule:

- **You do not need to map to multiple terminologies in diagnosis and procedure (e.g., ICD-9 and HCPCS). Map to the terminologies your data use.**

Mapping Demographics

Detailed examples for mapping demographics are provided in the attached spreadsheet ‘DemographicsMapping’. The spreadsheet has instructions for modifying the examples to reflect your local data, which can then be imported directly into your ontology table. Remember to not insert duplicate rows in your table…. either DELETE the existing rows first or perform a SQL update.

Our recommended approach is as follows:

1. **Modify the dimcode for dimension table queries.** For queries against the patient_dimension, replace the list of values in the dimcode with your local values. This has already been done to support the demodata in race, sex, and ethnicity. See these for examples. The Excel sheets ‘PCORIRaceEth’ and ‘PCORISex’ provide a template for updating your dimcodes. It is your responsibility to get the updated dimcodes back into your metadata table. Approaches include deleting the data in the demographics table and reinserting it from this file, or copying and pasting in a visual editor.

You can also use the approaches in the diagnosis section if you store some of your demographics outside of the patient_dimension table. For an example template of this more complex approach, see the Excel sheet ‘PCORIRaceEth2’, which shows a hypothetical example using race codes for both race and ethnicity. Note that only the dimcode approach is currently supported by the PopMedNet transform.

Mapping Diagnoses

Sites should map diagnoses using the following two approaches (in this order):

1. **Modify the the c_basecodes for standard terminologies to match your local codes** (preferred when there is a 1:1 mapping between standard codes and local codes). Specifically, it appears many sites have their diagnoses in ICD-9 format but with a different scheme or code format. If your ICD-9 codes are, for example, ‘PHSICD9:xxxx’ instead of ‘ICD9:xxx.x’, it is permissible to change the c_basecode in the ontology table to match your local code. (And note the concept_cd in the concept_dimension will also need to change.) The new column, PCORI_BASECODE, provides a reference for the standard code and should not be changed.
2. **Create local children of the terms in the ontology table.** For additional local diagnosis codes, sites will need to create child terms of nodes, either manually or using the mapping cell. The Mapping Cell for SCILHS, which automates this process, is

released here: https://community.i2b2.org/wiki/display/NCBO/PCORI+Mapping+Tools+version+1.0 The approach is this: If a particular ICD-9 code maps to five local codes, create five children underneath that ICD-9 code and queries will automatically gather all the local codes. In general, children should have a fullname and dimcode that corresponds to the PCORI ontology, but with a basecode corresponding to your local system. Run the included BASECODE fix script in Other Notes below to fix the PCORI_BASECODE column after the mapping process is complete.

Version 2.0.2 adds a Data Type modifier tree. Your diagnosis facts must have a modifier_cd (or a separate modifier row) that indicates whether the fact originated from a condition list (e.g., problem list) or a diagnosis (i.e. generated during an encounter). The transform will not function correctly otherwise! See the meds mapping section for more information on this. Note that it is permissible to use the 'Unknown' modifier codes (CONDITION_SOURCE:UN and DX_SOURCE:UN) if you do not know the specific type of condition or diagnosis source.

Mapping Vitals

Detailed examples for mapping vitals are provided in the attached spreadsheet 'Vitals_workbook'. The spreadsheet has instructions for modifying the examples to reflect your local data, which can then be imported directly into your ontology table. Remember to not insert duplicate rows in your table…. either DELETE the existing rows first or perform a SQL update. Note that the 2.0.2 update includes new tobacco usage codes that must be mapped.

Our recommended approach is as follows:

1. **Modify the the c_basecodes for standard terminologies to match your local codes** (preferred when there is a 1:1 mapping between standard codes and local codes). The Excel sheet provides a template for updating your c_basecodes. It is your responsibility to get the updated c_basecodes back into your metadata table. Approaches include deleting the data in the vitals table and reinserting it from this file, or copying and pasting in a visual editor. Note that you will also need to modify the basecodes for the two modifiers in vitals, if you have that data.
2. **Create local children of the terms in the ontology table.** For additional local vitals codes, sites will need to create child terms of nodes, as you did for diagnoses. We do not recommend using the mapping tool for this, because Vitals is a small table. There are examples in the spreadsheet of doing this approach manually.

In addition, if you have units in your fact table, make sure you enable unit conversions in i2b2 and make sure your unit codes match those in the metadata_xml in the vitals ontology.

If you do not have units in your fact table, make sure the default unit in the metadata_xml field matches the units your data are stored in. Please do not change the default unit - it is assumed by SHRINE. If your data are not stored in this format, you will either need to convert your data or turn unit conversions on.

See here for more information:
https://community.i2b2.org/wiki/display/DevForum/Metadata+XML+for+Medication+Modifiers

Mapping Procedures

Mapping procedures will hopefully be a straightforward process, similar to option 1 (renaming basecodes) in the diagnosis mapping above. Please note that the procedures table has many synonyms (duplicate terms with c_synonym_cd='Y'). The duplicates should not be put into your concept_dimension table. The script in step 6 of Installing the Ontology above should help.

Mapping Encounters

We anticipate sites will need to ETL additional encounter information to meet PCORI requirements. We expect this data, except for DRGs, will be stored as **additional columns in the visit_dimension table**. This is detailed in the SCILHS_Encounter_ETL_Guidance_v11.xslx spreadsheet. See especially the red columns (new columns) and the yellow columns (existing columns, but potentially new data required).

SQL code to add the columns (in MSSQL format) is below:

```
    ALTER TABLE [dbo].[VISIT_DIMENSION]
        ADD [DRG] varchar(50) NULL,
        [DISCHARGE_STATUS] varchar(25) NULL,
        [DISCHARGE_DISPOSITION] varchar(25) NULL,
        [LOCATION_ZIP] varchar(25) NULL,
        [ADMITTING_SOURCE] varchar(25) NULL,
       [FACILITYID] varchar(25) NULL,
       [PROVIDERID] varchar(25) NULL
```

Likely the data added to the visit_dimension can be ETLd using PCORI codes, but should you need to change the coding system, please see the Encounters-noDRGs.xslx spreadsheet. You will use the same mapping approach used for demographics. Note that the 2.0.2 update adds a new encounter type for encounters which begin in the emergency room and transition to inpatient. This is not a preferred encounter type; only use it if you have such data - do not generate it.

- **Modify the dimcode for dimension table queries.** For queries against the patient_dimension and encounter_dimension, replace the list of values in the dimcode with your local values. The Excel sheets provide a template for updating your dimcodes. It is your responsibility to get the updated dimcodes back into your metadata table. Approaches include deleting the data in the demographics table and reinserting it from this file, or copying and pasting in a visual editor.

DRGs should be ETLd into the fact table. If your site does not currently have DRGs, consider adding them using the same coding system in the ontology (MSGRG:xxx for MS-DRG and CMSDRG:xxx for CMS-DRG). You are only required to have one type of DRG - you do not need to populate both MS-DRG and CMS-DRG. If you plan to use a different coding system, please refer to the Encounters-DRGs.xslx spreadsheet. You will use the same mapping approach as diagnosis:

- **Modify the the c_basecodes for standard terminologies to match your local codes** (preferred when there is a 1:1 mapping between standard codes and local codes). Specifically, you might be using MS-DRG but with a different scheme. If your MS-DRG codes are, for example, 'PHSMSDRG:xxx' instead of 'MSDRG:xxx', it is permissible to change the c_basecode in the ontology table to match your local code. (Also note the concept_cd in the concept_dimension will also need to change.) The new column, PCORI_BASECODE, provides a reference for the standard code.
- **OR, Create local children of the terms in the ontology table.** For local DRG codes that are not MS-DRG or CMS-DRG, sites will need to create child terms of nodes, either manually or using the mapping cell. The approach is this: If a particular DRG code maps to five local codes, create five children underneath that DRG code and queries will automatically gather all the local codes. In general, children should have a fullname and dimcode that corresponds to the PCORI ontology, but with a basecode corresponding to your local system. We are not providing a configuration for the mapping cell for encounters, because we expect this approach will not be necessary. The spreadsheet documents this approach, however.

Mapping Labs (CDMv2)

The SCILHS Labs table is a tree of approximately 80 labs identified as having high importance for research (see the list here). These labs are mapped to about 350 LOINC codes.

First, you will need to map these 80 labs using LOINC codes, though they don't need to be the LOINC codes we have included, and most of the 700 included codes will not be used. Your data mapping process will vary depending on your source data:

1. **Local data is mapped to LOINC codes that are in the SCILHS lab tree:** Modify the the c_basecodes to match your local codes. Specifically, you might be using LOINC codes but with a different scheme. If your LOINC codes are, for example, 'LOINC2014AB:xxx' instead of LOINC:xxx', change the c_basecode in the ontology table to match your local code. (Also note the concept_cd in the concept_dimension will also need to change.) The new column, PCORI_BASECODE, provides a reference for the standard code (please do not change it).
2. **Local data is mapped to LOINC codes that are missing from the lab tree:** Create additional LOINC children of the laboratory grouper terms in the ontology table. For LOINC codes that are not included in the table, sites will need to create child terms of the grouper nodes. The approach is this: If five LOINC codes fit in a grouper category, create five children underneath that grouper code and queries will automatically gather all the local codes. Children should have a fullname and dimcode that corresponds to the PCORI ontology, with a basecode corresponding to your local system's LOINC fact code.

3. **Local data are not mapped to LOINC codes:** Identify the LOINC codes in the tree that correspond to local codes. If there is a 1:1 relationship, change the c_basecode as in option 1 above. If not, create children under the LOINC code as in option 2 above.

All three of these approaches are similar to the approaches previously used for other domains.

If the majority of your data are local codes (non-LOINC), it might be simpler to map using the Mapper tool we distributed when Diagnoses were released. It can be downloaded here: https://community.i2b2.org/wiki/display/NCBO/PCORI+Mapping+Tools+version+1.0 . Follow this procedure to use the mapping tool:

1. Set up and install by following the instructions in the software_and_data ZIP file at the link above.
2. Next, update the mapping tables to use the labs table. The prepare_lab_mappings.sql script in the Google Drive will assist you. Before running, make the minor changes for your configuration noted in the comments. It also supports importing a partial mapping if you have one. Note that only a SqlServer version is provided, but the code was written to be very portable.
3. Next, run the Mapping tool in the workbench to perform manual mapping and mapping verification. (See the Word document at the link above.) It is useful to manipulate the project_ont_mapping table through SQL during the mapping to speed up the process as you find patterns in lab names. You might need to restart the workbench if you make changes in SQL.
4. (Optional) Run the utility script (#5 in the SQL above) to verify you have some mapping to every lab category.
5. Use the workbench mapping tool to generate the integration table.
6. Run the script at the end of the SQL file (#6) to reinsert pcori_basecode and pcori_specimen_source into your integration table. DO NOT use the script at the end of this document (for diagnoses).
7. Your integration table is now ready to be used in place of the stock pcornet_lab table!

A new version of the Mapper tool that uses MetaMap to partially automate mapping will be released soon.

Second, you will also need to map modifiers if your data have this information (result priority and result location: lab or point of care). Edit the c_basecode you are using for the modifiers in your i2b2. (This corresponds to modifier_cd in the fact table.)

Third, you will need to address query metadata and unit conversions, as was done in labs:

- If you do not have units in your fact table, make sure the default unit in the metadata_xml field matches the units your data are stored in. Please do not change the default unit - it is assumed by SHRINE. If your data are not stored in this format, you will either need to convert your data or turn unit conversions on.

- If you have units in your fact table, make sure you enable unit conversions in i2b2 and make sure your unit codes match those in the metadata_xml in the vitals ontology. In this case, all terms with numeric values must have a valid units code.
- If your normal range is significantly different than that in the SCILHS labs table, please change it and contact us so we can synthesize these differences for the SCILHS SHRINE ontology. (There is likely to be some variance among the leaf nodes but not the groupers.)

More information on metadata XML is here: See here for more information:
https://community.i2b2.org/wiki/display/DevForum/Metadata+XML+for+Medication+Modifiers

Finally, it is very important you send us either your labs ontology or a changelog after mapping is complete, so that we can create a SHRINE ontology consistent with what sites are actually using!

Mapping Meds (CDMv3)

The SCILHS Meds table contains a 2012 RxNorm tree organized by VA Drug Class (NDF-RT). Also we have merged in all the NDC codes in the UMLS mapping, as children of RxNorm codes. Mapping meds is different depending on whether the medication is due to a prescribing event or dispensing event. (We assume most events are prescribing events in your local data; however, you might have dispensing events from pharmacies that are part of your hospital system.)

Prescribing events:

a. Set the modifier_cd in your fact table to RX_BASIS:01 (for prescriptions to be dispensed) for RX_BASIS:02 (for prescriptions to be administered in the hospital). If you already are using modifier_cd, you will need to add an additional modifier row - please refer to the i2b2 documentation.
b. Your local meds must be mapped to RxNorm. If your meds are in NDC, our tree might eliminate most of your work. Use the standard mapping approaches, described below.

Dispensing events:

a. Set the modifier_cd in your fact table to RX_BASIS:DI. If you already are using modifier_cd, you will need to add an additional modifier row - please refer to the i2b2 documentation.
b. Your local meds must be mapped to NDC. It is possible you will need to add NDC codes to our tree. Version 2.0.2 adds new NDC codes to the medication tree; you only need to update to this if you have NDC codes in your i2b2.

Mapping meds:

- If your meds are already mapped to RxNorm (prescribing) and NDC (prescribing or dispensing) or there is a 1:1 mapping, modify the the c_basecodes in pcornet_med to match your local codes. You might need to add NDC codes that are missing.

- If multiple local codes map to a single RxNorm or NDC code, add local children as you have done in other domains. Prescribed meds do not need to be children of NDC codes, just RxNorm. You can use the mapping tool as per the instructions for labs above. Substitute the labs SQL script in those instructions with the included prepare_med_mappings.sql script. Be sure you have copied the integration table you created for labs to another table or it will be deleted!
- Be sure to set your modifier_cds in the fact table as described above.

When mapping is finished, please run the totalnum script on the meds ontology and send us your meds ontology. We are using this for QA, and it is especially urgent if you have added any NDC or RxNorm codes, or we cannot build a compatible SHRINE ontology!

Mapping the rest

The above sections document mapping all sections in CDM v1 and a growing amount of C except enrollment. Enrollment is currently a computed value that should not require modification. You can modify it if necessary. It uses the following method:

1. **Create computed terms for values that can be derived from data in your fact tables.** For an example, see Enrolled->Basis[Encounter]. This returns all patients with an encounter since 1/1/2000 and can be further date-constrained by the query tool.

More complex solutions are possible but please inform us if you take another route so that we can align our parallel work that is dependent on the ontology.

Other notes

- Terms in the ontology are marked as editable (visual attribute ends in ‘E’). This was done to facilitate editing in the i2b2 workbench, but sites will probably want to make them read-only (for security) once mappings are complete. Simply remove the ‘E’ from the visual attribute.
- After using the Mapper tool’s integration step, run the following script to correct the pcori_basecode column - this version has a bug fix for SQL Server since v2.0. (A version of this is included in the prepare_lab_mappings script, where it works correctly on the pcornet_lab table):

  SQLServer:

  ```
  update integration
  set pcori_basecode = project_ont_mapping.destination_basecode from
  project_ont_mapping
  where integration.c_basecode = project_ont_mapping.source_basecode
  and integration.pcori_basecode is null
  and (project_ont_mapping.status_cd is null or
  project_ont_mapping.status_cd != 'D')
  and integration.c_path = project_ont_mapping.destination_fullname
  ```

  Oracle:

  ```
  update integration
  set pcori_basecode = (select
  project_ont_mapping.destination_basecode from project_ont_mapping
  ```

```
where integration.c_basecode = project_ont_mapping.source_basecode
and integration.c_path = project_ont_mapping.destination_fullname
)
where exists (select 1 from project_ont_mapping
where integration.c_basecode = project_ont_mapping.source_basecode
and integration.c_path = project_ont_mapping.destination_fullname
and project_ont_mapping.status_cd != 'D' or
project_ont_mapping.status_cd is null)
and integration.pcori_basecode is null
```

Information

SCILHS Mapping Policy

SCILHS sites are required to make their i2b2 data accessible through the i2b2 ontology representation of the PCORnet Common Data Model (CDM) [PCORnet ontology], which SCILHS is using for SHRINE queries and to create a datamart at each site in the CDM structural format for pan-PCORnet queries, as required by our milestones. This ontology serves as the basis for standardized data across sites.

Site Requirements

- Sites are required to make this ontology compatible with their local data through a “local mapping” process, which will not require changes to ETL processes but only changes to the ontology database table (and concept_dimension/modifier_dimension). This local mapping process is detailed in this document, and we will provide additional documentation, examples, and other help as we progress along this path.
- Sites should also use the ontology as a guide to what data is required in your i2b2 repository. If you determine some data is not in your current i2b2 instance, it must be added to the i2b2 instance if it is readily available. If adding it would require significant effort and it is not a core element (not e.g., diagnosis, procedure, and DRG codes), you might be allowed an exception. Please inform the SCILHS leadership.
- Sites do not need to limit what is in your repositories to only the items in this ontology, and you are welcome to have multiple ontologies for local investigators.
- Sites do not need to support multiple terminologies within each tree. We expect most sites to use ICD-9, ICD-10, DRG, CPT, and possibly HCPCS. Unsupported terminologies are inactive in the ontology (visualattribute contains an I). You can leave these out of your concept_dimension.
- Dates are also required: start date for all facts and an end date (discharge date) for inpatient encounters. (Currently, this is the start and end date in the visit_dimension table.) There are hidden terms in the ontology that explicitly represent these dates, as a guide for implementers.

SCILHS Requirements

- We will provide additional details on the local mapping process.

- We will continue to provide a trivial 1:1 SHRINE mapping that will support distributed queries against the PCORnet ontology.
- We will continue to provide an ETL tool to create a datamart at your site in the CDM structural format that is compatible with this ontology as long as the local mapping process is followed correctly.
- The SMART cell will be configured out-of-the-box for the PCORnet ontology.

List of Labs to Map

Below is a list of all lab types that need to be mapped. Each of these is a folder in the ontology with several LOINC codes. (Note that the tree also contains HIV tests, which are inactivated because queries on these data without specific approval is illegal in Massachusetts. You do not need to map these codes.) In alphabetical order:

```
Activated partial thromboplastin time (aptt)
Albumin
ALP
Alpha-1-fetoprotein
ALT
AST
Basophils
Basophils in peritoneal fluid
Bicarbonate
Bilirubin
Bilirubin.non-glucuronidated
C reactive protein
Calcium
CD4 Count
Chloride
Cholesterol in HDL
Cholesterol in LDL
Cholesterol non HDL
Cholesterol Total
Cholesterol/HDL Ratio
Creatine Kinase Mb
Creatine Kinase Mb/creatine Kinase Total
Creatine Kinase Total
Creatinine
Cryoglobulin rheumatoid factor
eGFR
Eosinophils
Eosinophils in peritoneal fluid
ESR
GGT
Glucose
Hematocrit
Hemoglobin
```

Hemoglobin A1c
Hepatitis B virus core
Hepatitis B virus core Ab
Hepatitis B virus core IgG
Hepatitis B virus core IgM
Hepatitis B virus surface Ab
Hepatitis B virus surface Ag
Hepatitis C virus Ab
Hepatitis C virus IgG
Hepatitis C virus IgM
Hepatitis c virus rna
Hiv 1 rna
International Normalized Ratio
LDH
Lymphocytes
MCH
MCHC
MCV
Monocytes
Mycobacterium tuberculosis Mitogen stimulated gamma interferon
Mycobacterium tuberculosis stimulated gamma interferon
Myelin basic protein
Natriuretic peptide b
Natriuretic peptide.b prohormone
Neutrophils
Neutrophils in peritoneal fluid
Neutrophils.segmented in peritoneal fluid
Phosphate
Platelet Count
Potassium
Protein
Protein in peritoneal fluid
Protein in Urine
RBC Count
RDW
Reagin Ab in CSF
Reagin Ab in serum
Rheumatoid factor
Sodium
Triglyceride
Troponin I Cardiac
Troponin T Cardiac (qualitative)
Troponin T Cardiac (quantitative)
Urate
Urea nitrogen
WBC Count