SCILHS i2b2 PCORnet Common Data Model (CDM) documentation - for public release

Version 1.5.2 - 2/26/15

Jeff Klann, PhD

Introduction

this i2b2 ontology representation of the PCORnet Common Data Model (CDM) [PCORnet ontology] serves as the basis for standardized data across sites. Interoperable querying is supported using SHRINE. We have also developed a transform to create a datamart in the CDM structural format using data mapped to the ontology, which is used for pan-PCORnet queries.

Requirements

- Adapting this ontology to local data is done through a "local mapping" process, which does 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 the spreadsheets included in this release.
- It is not necessary to limit what is in your repositories to only the items in this ontology, and it is possible to have multiple other ontologies existing in parallel to this one.

- Not all possible terminologies in this ontology will be used. We commonly see ICD-9, DRG, CPT, and sometimes HCPCS. Unsupported terminologies are inactive in the ontology (visualattribute contains an I). You can leave these out of your concept_dimension. Note that even though ICD-10 is inactive, it is included in the ontology.

This ontology is running live (backed by an evolving version of the demodata). Go to https://www.i2b2.org/webclient/ and change the username to pcori. (The password is still demouser.) An increasing number of queries run.

Included Files

Release_152\Oracle\TABLE_ACCESS.txt and Release_152\SqlServer\TABLE_ACCESS.txt: Table access entries for the ontology, in Oracle and SQL Server formats.

Release_152\Oracle\SCHEMES.txt and Release_152\SqlServer\SCHEMES.txt: Schemes table for the ontology, in Oracle and SQL Server formats.

Release_152\[Oracle|SqlServer]\PCORNET_*.txt: Current SCILHS production version of all tables in the PCORnet CDM v1 spec. All of the ontology files are in pipe-delimited format. The largest are zipped. These are to be imported into corresponding tables in the database.

SCILHS i2b2 CDM Documentation v1.5.2.pdf: This document describing the ontology, the process of loading it, and the mechanism for creating local mappings.

Demographics_table_0709.xslx: An Excel file representation of the demographics table, with documentation and hints for creating local mappings for the demographics section.

Vitals_workbook_093014a.xslx: An Excel file representation of the vitals table, with documentation and hints for creating local mappings for the vitals section.

Encounters-noDRGs.xslx: An Excel file representation of the non-DRG portion of the encounters table, which are to be imported as new columns in the visit_dimension. Contains documentation and hints for creating local mappings.

Encounters-DRGs.xslx: An Excel file representation of the DRG portion of the encounters table, which are to be imported into the fact table. Contains documentation and hints for creating local mappings.

SCILHS_Encounter_ETL_Guidance_v11.xslx: An Excel file documenting PCORI's encounter table and how we expect it to be mapped into i2b2 in SCILHS. Note that we expect everything but DRGs to be added to the visit_dimension as additional columns. This will be

helpful to non-SCILHS sites wishing to have out-of-the-box compatibility with our ontologies and transforms.

Tasks

Installing the ontology
1. The ontology is broken down into six tables. You will need to create these tables using the standard ontology table structure with an added PCORI_BASECODE column. We have provided sample scripts to create these tables in the Oracle and Sqlserver folders (create_*_metadata_tables.sql), which you can use if you like.
2. The ontology is supplied as a set of pipe-delimited files, named pcornet_*.txt. Choose the sqlserver or Oracle folder depending on your database platform. Strings are double-quoted and null values are empty. There are no linebreaks in the strings, and all characters are standard ASCII. Import the ontology files into separate tables named according to the file names. (These are the tables created in step 1 above.)
   - pcornet_demo.txt - Demographics table - import into pcornet_demo. If you used the Demographics Excel spreadsheet to define your mappings, you will not need to import this file.
   - pcornet_diag.txt - Diagnosis table - Import into pcornet_diag
   - pcornet_enc.txt - Encounter table - Import into pcornet_enc
   - pcornet_enroll.txt - Enrollment table - Import into pcornet_enroll
   - pcornet_proc_nocpt.txt - Procedures table - Import into pcornet_proc.
     i. This version does not have CPT codes because one must have a CPT license to use CPT. Contact SCILHS if you have a license and would like the full version.
   - pcornet_vital.txt - Vitals table - Import into pcornet_vital
   - Note that the ontology contains one non-standard column, PCORI_BASECODE.
3. Import the TABLE_ACCESS table, also included as a txt file. You can replace the rows in your existing TABLE_ACCESS table.
   - You need to add entries in your TABLE_ACCESS table in order to perform queries. We recommend you use the provided file to ensure compatibility with the SHRINE pass-through mappings and forthcoming SMART-i2b2 configuration. It is fine to keep other entries for non-SCILHS ontologies in your TABLE_ACCESS table.
4. Optionally import the SCHEMES table, also included as a txt file. You can replace the rows in your existing SCHEMES table.
   - The SCHEMES table is used for searching by code when building queries in the webclient and workbench.
   - Note that you will need to change the `c_keys` in the SCHEMES table to match your local codes after you finish mapping.

5. Create indexes on some of the columns below, for each table (listed in order of importance below). If you used the sample scripts in step 1, you do not need to perform this step.
   - `c_fullname, m_applied_path` - these are the defaults
   - `c_hlevel` - this improves performance a lot
   - `m_exclusion_cd, c_synonym_cd, c_visualattributes` - improves performance, in decreasing order of significance
   - `sourcesystem_cd` - does not speed up ontology cell queries, but helpful when creating local mappings
6. After you do your mappings (and before you can run queries), you will need to create updated concept and modifier dimensions. You can delete the data from your existing concept_dimension and modifier_dimension tables prior if you prefer. Either way, you will need to generate entries for both tables from each of the six ontology tables. Example SQL SELECT statements are below. You can use these to generate exports to insert into your concept and modifier dimension. Modify these as needed for each of the six tables.
   - ```
     select c_dimcode AS concept_path, c_basecode AS concept_cd, c_name
     AS name_char, null AS concept_blob, update_date AS update_date,
     download_date as download_date, import_date as import_date,
     sourcesystem_cd as sourcesystem_cd, 1 as upload_id into
     CONCEPT_DIMENSION from dbo.PCORNET where m_applied_path='@' and
     c_tablename='CONCEPT_DIMENSION' and c_columnname='concept_path' and
     (c_columndatatype='T' or c_columndatatype='N') and c_synonym_cd =
     'N' and m_exclusion_cd is null and and c_basecode is not null;
     ```
   - ```
     select c_dimcode AS modifier_path, c_basecode AS modifier_cd,
     c_name AS name_char, c_comment AS modifier_blob, update_date,
     download_date, import_date, sourcesystem_cd, 1 upload_id into
     modifier_dimension from PCORNET where m_applied_path!='@' and
     c_tablename='MODIFIER_DIMENSION' and c_columnname='modifier_path'
     and (c_columndatatype='T' or c_columndatatype='N') and c_synonym_cd
     = 'N' and m_exclusion_cd is null and and c_basecode is not null;
     ```
7. At the end you might want to recompute stats on your database, especially if you are running Oracle:
   - ```
     BEGIN
     ```
   - ```
     DBMS_STATS.GATHER_SCHEMA_STATS('<schema_name>',
     DBMS_STATS.AUTO_SAMPLE_SIZE);
     ```
   - ```
     END;
     ```

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 release all changes to each table before you engage in mapping, but some changes are out of our control.

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.

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.

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:

- **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 the rest

The above sections document mapping all sections in CDM v1 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:

```
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 != '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

About the ontology

What's in the ontology?
- *Core ontology* - based on Dan Connolly's code to generate ontologies from the PCORnet CDM spec, currently v1 of that spec.
- *Lori Phillip's ontology trees on BioPortal* - ICD-10 2014AA, ICD-9 2014AA *(ICD-10 is inactive in the ontology because we are not aware of sites that use it)*
- *Partners' RPDR* CMS-DRG tree and *Beth Israel's* MS-DRG tree (not sure which version)
- *Nathan Wilson's* HCPCS tree (not sure which version)
- 3-digit zip codes are derived from the i2b2 demo data
- Age tree is a more granular version of that found in the demo data
- Some clarifications (e.g., meaning of the procedure code types) from the coordinating center appear in the tooltips.

What's not in the ontology?

* *LOINC v236 and SNOMED Clinical Findings v1.2* are available on BioPortal, but they are not included in this version of the ontology. We suspect sites are not using SNOMED, and the language from the coordinating center indicates CPT and HCPCS are preferred over LOINC for procedures at present: "Only billed procedures should be included in the PROCEDURE table. The ORDER concept may be incorporated into future phases of the CDM." If you need LOINC, please contact us and we would be happy to incorporate it.

* *CPT* is not included in the public release because we do not have a way of ensuring licensing. It is available in private release to those who attest they have a license.

Changelog

6/13/14 - initial release
6/24/14 -
- Changed export format (brackets no longer surround dates, null values are empty rather than "(null)", file type is ASCII)
- Set all update_date fields to be non-null

- Corrected foreign characters, long hyphens, and backquotes that were invalid (content changes were mostly in the ICD-10 tree, a few in ICD-9 and HCPCS, the rest were in tooltips)
- Fixed invalid c_path and c_symbol entries (not required for querying, SHRINE, or SMART - used by some tools like Lori's mapping validator).

6/25/14 - Released TABLE_ACCESS in same export format as main table.

6/26/14 - Minor changes to ontology
- BMI no longer greyed out
- Only top-level folders are cases
- Modifiers have been excluded from top-level folders (they are unqueryable anyway)

7/10/14 -
- **7/1/14**
- Age buckets had some errors in basecode, trailing slashes, c_path, and c_symbol
- Some metadata XML was missing the XML header (but only for hidden elements)
- Line breaks removed from metadata XML (to ease importing)
- **7/7/14**
- Biobank flag = No is hidden now; this should not have been queryable
- **7/10/14**
- Made all terms editable
- Added version of tooltip column without linebreaks

Released new ontology to sites

7/22/14 - 399,921 rows
- Added 129 ICD-9 codes that were retired by 2014AA but are still used.
- Fixed an error in the dimcodes - they matched c_path instead of c_fullname
- Added pcori_basecode column.
- Removed pipes (|) from C_BASECODE.

7/29/14 - LCP
- Fixed the leaf to folder modification resulting from retired terms.

Released as patch to sites

8/4/14 -
- Fixed three HCPCS synonyms
- Remove linefeeds from tooltips

8/7/14:
- Removed start date from Enollment/Encounter-based so it will work on all db platforms.
- Split ontology release into six tables.
- Made uncommon/unused terminologies inactive (SNOMED, LOINC, ICD-10).

8/12:
- Removed duplicate ICD9:645 and ICD9:386.00 and renamed Sporting Injury from ICD9:E899 to ICD9:E889

8/13:
- Inserted CPT_2014AA from Bioportal into the procedures table

8/15:

- Updated sourcesystem_cd from 'Integration_tool' to 'Integration'

8/18:

- Added schemes table
- Version concepts added
- Inactive visual attributes propagate to children
- CPT procedure modifiers folders is now actually a modifier folder
- Deleted basecodes from internal CMS-DRG groupings - it never appears in data
- Fixed top-level DRG basecodes
- Added schemes to scheme-less basecodes
- Made dates non-null in ontology (if a date was null, it was set to the update date)

8/19/14:

- Fixed errors in dimcodes on Diagnosis\PDX and the version concepts.
- Removed modifiers from the schemes table.

Released new ontology v1.4 to sites (8/19/14)

9/30/14:

- Modifiers had C_FACTTABLECOLUMN='concept_cd' instead of 'modifier_cd'.
- chart:n is now a computed value that returns 0, and chart:y returns all patients
- Enrollment table is fixed and reorganized - only encounter-based enrollment is now active and it is a computed value so no data needs to be entered
- Vitals -> Height had a bug and was partially set up as a modifier
- Added vitals metadata_xml for normal ranges and unit conversions.
- Made NI inactive in vitals

Released new vitals and enrollment tables v1.5 to sites (10/3/14)

11/16/14:

- Modifiers had C_FACTTABLECOLUMN='concept_cd' instead of 'modifier_cd'.
- Made NI inactive - it is not computable
- Merged in around 700 retired CPT codes

Released new procedures table v1.5 to sites (11/18/14)

11/24/14:

- Made NI inactive - it is not computable
- Added MS-DRG tree and modified the DRG tree to reflect the two coding systems
- Changed all non-DRG queries to query the visit_dimension
- Changed top-level items to cases (containers)
- Bugfixes on some encounter items, especially the null flavors

Released new encounters table v1.5 to sites (12/1/14)

- Diagnosis modifiers had C_FACTTABLECOLUMN='concept_cd' instead of 'modifier_cd'.
- pcori_basecode issues fixed (as defined by ontology_fix_script_v5, released with the transform)
- Made version concept appear at end of list
- Changed C_COLUMNDATATYPE='T' for BIOBANK_FLAG:Y (queries do not run correctly if not) - this was part of a patch on 8/27/14.
- Made NI inactive on all remaining tables.

- Set HISPANIC yes and unknown to use dimcodes in the patient dimension by default.
- Made changes to the pediatric age ontology per suggestions by Nate Apathy at Cerner Research:
  - 2 months: changed label from "2 months old" to "02 months old" for sorting order
  - Changed "1 months" to "1 month" & "1 days" to "1 day"
  - 0/1/2 months: changed type to folder so days display under in hierarchy
  - modified month ranges to be inclusive toward greater number vs. lesser number, to match format for years (e.g., "1 month old" means 0-1 months). This also fixed a bug in the Oracle version that caused months to not query properly.
  - Note that days still follow the non-standard format ("1 day old" means 0-1 days), for backward compatibility.

Released new CDM ontology tables v1.5.1 to sites (2/13/15)

- More fixes to the pediatric age ontology. Now using all of Cerner Research's modifications. Of note, "days old" now follows consistent format with other ages.
- HCPCS tooltips (including modifiers) now follow the i2b2 standard tooltip format
- HCPCS modifiers are now correctly placed under one modifiers tree (rather than several separate trees for different classes of modifiers)
- The encounters .txt file mistakenly contained an unnecessary column with linefeeds in v1.5.1. This column has been removed.
- In Encounters and Procedures, cleaned up the Data Source modifier - Unknown and Other are now hidden, and No Information is active and by default returns all records
- In Demographics, cleaned up No Information - now always active and by default returns the count of patients with null in the specified column
- Fixed Ethnicity: greyed-out Non-Hispanic, which is a negated term and thus doesn't make sense in an i2b2 ontology; fixed an issue on Hispanic with c_columnname in Oracle
- Updated the schemes table

Released new CDM ontology tables v1.5.2 to sites and to public (2/26/15)