Field Mapping v2 (Fall 2018)

This doc describes ~~the range of field mappings types encountered so far, and a schema or template to describe field mappings and transformation functions~~ is a design wish-list for coming work.

In the context of harmonization, a value in a source variable, file/column, or table/field needs to be mapped to a concept available in OHDSI. Values of a field should be transformed to an appropriate or standard unit when numerical, and to concept values when the variable is categorical. Untransformed or free text fields are also discussed.

The notion of a named visit common to many clinical studies is difficult for its own reasons, but presents some issues upon representation in OHDSI.

# PARAMETERIZED MIGRATE\_FUNCTIONS.Py

The bulk of the data in a clinical study are measurements or observations taken when a patient visits a clinic periodically. The distinction between measurement and observation is that a measurement’s value is either a number or a (discrete) concept. An observation can take a free-text value. Observations can be numerical, but if an instrument was involved in finding the number, it’s thought of as a measurement. Ultimately the concept table within OHDSI describes the domain of the value which indicates the table.

## Transformation TYPES

### Numerical

* Multiplicative factors:
  + Decimal\_to\_percent
  + Years\_to\_Months
  + Years to days
  + Cm\_to\_in, in\_to\_cm
  + Kg\_to\_lb, lb\_to\_kg
  + Centiseconds\_to\_milliseconds
* Multiplicative with offset
  + Fahrenheit to Celsius (haven’t seen body temperature as a lab measurement)
* Offset only
  + Celsius to Kelvin (never seen it)

### CATEGORICAL

* Yes\_no\_to\_concept
* Sex\_to\_concept
* Race\_to\_concept
* Btx\_to\_placebo

### AS-IS

Some as-is values are free text, many are dates.???

## ODDITIES/ERRORS

There are presently some mappings to observation.value\_as\_string that contain dates and are a hack to deal with visits and study related milestones like randomization and end of study. They are addressed as more of a table level thing like supporting the visit table. The others are curios.

* These are all in error and should be concept mapped to yes/no.
  + Best.pcsf.pcmihs to SNOMED-428752002 “Recent myocardial infarction” to observation.value\_as\_string
  + Best.pcfs.pcangina to SNOMED-4557003 “Preinfarction syndrome” to observation.value\_as\_string
  + Best.pcfs.pcpregnt to SNOMED-169449001 “Trying to conceive” to observation.value\_as\_string
  + Best.pcfs.pccormorb to SNOMED-3998192003 “Co-morbid conditions” to observation.value\_as\_string

# Study-Specifics and DATES: PERSON.py

Person.py started out as a way to populate the OHDSI person table, but grew to include other issues as well.

## DATES vs offset DAYS

* Calculate\_year\_of\_birth
  + BEST is odd because it has dates and we need the regex
  + Others subtract age from study start

YOB is used in OHDSI’s person table. Study start isn’t clear for each study and even so, the age isn’t correct. We may as well just use 2000-01-01 as a base, the study start for all participants in all studies, to convert between age and DOB.

## IDS

* Get\_study\_person\_ids
  + Need **field** and **table** to select distinct from on study side
* Get\_all\_person\_ids
* Convert\_person\_id\_to\_ohdsi
* Convert\_person\_id\_to\_study
  + Need a **regex** to extract the int part,
  + So far just a **constant** to prepend to recreate the study id once the shift is dealt with

## PERSON

* Populate\_person
  + INSERT INTO person (person\_id, gender\_concept\_id, year\_of\_birth, race\_concept\_id, ethnicity\_concept\_id)
  + Date, sex, age, race column names and table.
  + TOPCAT’s race value is weird and involves some math and two columns because it splits out ethnicity. In fact, if we represent in the person table as study-native or as close to, we can use the race and ethnicity fields there. Still yuk, it’s not harmonized.
  + BEST does ethnicity too.
* Factory

## OTHER

* Get\_id\_field\_name (for queries)
* Get\_date\_column\_for\_table (branches by table name, but ONLY for BEST. Ignored otherwise !!!!) ??
  + This may be about incomplete implementation.
* Use\_date\_column\_on\_select (defaults True)

# VISIT http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:visit\_occurrence

There are two questions regarding the date of a value (measurement or observation etc.). First is what date to use. These anonymized studies don’t include the data of randomization and measurements are recorded as an offset to that date. It’s that offset that’s most important, so adding the offset to 2000-01-01 so that we can get the relative value into a date field would work.

The second issue is how to access values by visit rather than date. If we populate the visit\_occurence table with concepts related to the offset, say in month-long increments, we can use the visit\_occurrence table to get to data by the visit type.

The fields that seem to offer the most hope for utility of the visit\_occurrence table with regard to clinical study visits are the visit\_concept\_id and visit\_type\_concept\_id columns. I suppose you could have a study-specific ontology of terms that describe the visits within a single study. More useful for combining studies would be an ontology that unifies the concepts. We need a survey of visit types/names among the studies we deal with. The motivating use-case for now is the change in BP we are considering for the NHLBI data pilot. The starting BP is assumed to be early in the study like randomization. We’d like the later BP readings to be about the same amount of time later, say, a year.

The concepts for the visit\_concept\_id domain in OHDSI are pretty basic:

* ER Visit,
* Inpatient Visit,
* Outpatient Visit,
* Long-Term Care Visit.

The reason for the visit, interesting when an outcome like CV Hospitalization is of interest, goes in visit\_type\_concept\_id??? (could be a local decision)

## clinical STudy VisiT types proposal

These are all types of visits from clinical studies with a parent concept of Clinical Study Visit Type.

* Randomization
* Visit @ month ( have ontology include concepts for up to 150 years of 12/months/year?!!)
* End Of Study/Death

They don’t enumerate more/less standard clinical study visits: randomization, titration, monthly and yearly. This doesn’t include common one-time visits for measurements or examinations that aren’t performed regularly or only in the case of certain events.

These are issues regardless of the use of OHDSI!

## VISIT SOURCE COLUMNS

Date Columns: these are typically days since randomization. Need the column name and unit for each table:

* BEST 2708
  + Mort1 (death) why put date of death into observation????????
  + Clab1 visit number in days, 1,120, 30
  + Wd
  + Lab1 visit number in days? 1,120,30
  + Cvh2
  + Cvs visit number in days
  + Clab2 visit number in days 1,120,30
  + Qol3 visit number in days
  + Lab2 visit number in days 1,120, 30
  + Cvh1
  + Diab
  + Ecg visit number in days 1,120,30
  + Muga visit number in days 1,120,30
  + Ame visit number in days
  + Qol2 visit number in days
  + T????
  + Pe visit number in days
  + Br
  + Eos
  + pcsf
* HFACTION: 2130
  + Analysis (none)
  + Cecdeath (none, death) why put death\_days into observation?
  + Cpx cpxdays ?
  + Labs ???
  + medhx2 (none)
  + randvit (none)
  + studcomp (none?)
* SCDHEFT: 2521
  + Basecrf (none)
  + baseline\_new (none)
  + death (none, death)
  + ecg (none)
  + endpt\_new (none)
  + rdemog (none)
* TOPCAT:
  + q002, visit name from visit column, could derive times from these names
  + t003 all BASE
  + t004 all BASE
  + t005 all BASE
  + t006 all BASE
  + t008 all BASE
  + t010 all BASE
  + t011 all BASE
  + t030 all EOST
  + t079 only 530 entries (death?)\d
* ACCORD: 10251
  + F01\_inclusionexclusionsummary : (none)
  + accord\_key : (none)
  + otherlabs\_201604 by visit name in “visit” column
  + concomitantmeds by visit name in “visit” column
  + lipids\_201604 by visit name in “visit” column
  + F07\_baselinehistoryphysicalexam by visit name in “visit” column
  + bloodpressure\_201604 by visit name in “visit” column
* AIM-HIGH: 4273
  + Conmed by visit name in “visit” column, could configure a rough time based on the names
  + patient (none)
  + lab by visit number in visit column: units???
  + Incl (none)
  + demo (none)
  + hx (none)
  + pe (none)
* ALLHAT: basic\_with\_outcomes XXXX
* BARI2D: 2368
  + BARI2D\_BL (none)
  + BARI2D\_LONG by visit number in column “fuper” units???
* ATMOSPHERE/PARADIGM: test

# OUTCOMES, EOS and RANDOMIZATION

Death and End-of-Study dates ended up in the observation table and belong in Death or Condition\_Occurrence.

See also the discussion under visits, regarding the reason for admission.

## RANDOMIZATion AND non-DEATH EOS

**Where to put this stuff if not observation?!**

You could argue that randomization is some kind of a visit, and I’ve mentioned elsewhere here that a clinical visit ontology should include randomization.

EOS in the cases of last non-death contact involves a visit, right? Can you correlate the date we have for EOS with a visit somewhere? ALWAYS? It may be most explicit to have an EOS visit, then you can look it up by that visit type just like randomization or any other date you might be interested in. The EOS visit may not have any data associated with it, but that shouldn’t be an issue.

## Death

* All-cause death 🡪 Death (no qualification, no filtering where clause)
* CV death 🡪 where Death.cause\_concept\_id == (CV Death)

There is a mapping for best.mort1 that goes into observation. Not sure what uses it.

## HospitaliZATION

“Admitting Diagnosis” appears in Condition\_Occurrence with a start-date and corresponding visit whose type implies admission (Inpatient Visit or Outpatient Visit).

* Hospitalization
* CV-related hospitalization

# OTHER ISSUES RELATED TO SCHEMA

## MULTI-TENANT

Study harmonization ends up involving the community of Science eventually. Data is studied to write papers that get read. Readers will want to know methods, which in this work includes the harmonization. Another angle is that of a second study on previously harmonized data. The harmonization scheme might be re-used or at least read. In the absence of standards, or even the presence of developing standards there may be more than one standard active for a study in the same platform instance (database) at the same time. The schema needs to support that. Because the whole point is for the data to reside in OHDSI, OHDSI needs to support more than one harmonized version of a study. There are *issues* with that.

### Multi TENant MAPPING

Study\_to\_ohdsi\_mapping is keyed by study as if there is only one mapping per study. This needs to change and would ripple to the following tables: ohdsi\_calculation\_function, ohdsi\_calculation\_argument, events\_mapping. Adding user\_id to the key then limits you to one mapping per study/user though it might be nice to know who owns a particular mapping. Study/user/instance is one way to go, the other would be a surrogate key tied to studies and users and instances in another table…. \*\*\*\* TBD \*\*\*\*

### MultI TENANT DATA (OHDSI)

Multi tenant data in OHDSI isn’t so easy if you want to avoid modifying a schema that isn’t part of source code you want to own. We avoided that path by dividing person\_id up into ranges instead of adding a study\_id to the OHDSI schema. We could continue that path and add a second dimension to the person\_id ranges. Instead of reserving blocks of a million person\_ids per study, we could reserve blocks of 100k to users within a range reserved for a study. That only makes room for 10 versions of a mapping.

Other options are to leave OHDSI as it is and divide out at the instance level. It may be too awkward or involve too much code change, but you could put different instances of OHDSI into different schemas. Finally, since the seam between mapping data and OHDSI data is narrow enough, you could consider just running different databases for each different mapping set. Even when harmonized, study data lives apart from other study data only to be joined in the form of an analysis matrix. So if you had different mappings on different studies, the harmonized study data could be on different servers and it wouldn’t cause too much a problem.

The more interesting part of the question of different servers is how to get the data onto them, especially when consider the Platform. One way would be different SQL connections to different servers. Another would be to use an API and have different connections. Either way, there’s some architectural level data that needs stored. \*\*\* TBD \*\*\*

## Source Table STRUCTURE

Labs tables are often in melted form, so selecting from them requires a column and value to use in a where clause to identify what would be the column.

## ADDITIONAL DestinatION TABLE FIELDS

OHDSI has a few other fields that can be filled in the destination tables:

* Measurement\_type\_concept\_id ??
* Operator\_concept\_id ??
* Unit\_concept\_id
* Range\_low, range\_high ?
* Visit\_occurrence\_id \*\*\*\*\*!!!!

## New ONTOLOGY: CLINICAL\_VISIT\_TYPE

As described above. Some visits have specific purposes, others are just concepts for monthly boundaries.

## UCD-KAO Ontology updates/Changes (TBD)

## IMPROVED DEstination TABLE SUPPORT, CONCEPT.DOMAIN\_ID

Condition, Measurement, Observation, Device, Procedure

### Modifications to EXISTING MAPPINGS

There concepts from each of the domains Condition (88), Device (4), Drug (57), Observation (13), Procedure (16), Race(1), and UCD.Kao D (45) that are not in like-named tables. UCD.Kao D should probably be deprecated and replaced on a case-by-case basis with the existing domains. Drug implies use of the drug\_exposure table (see below). Work for the NHLBI pilot put a simpler description of concomitant meds into the observation table. An incremental step may be to put the simpler values into drug\_exposure where they belong, without tackling the exposure interval. Just move them as-is.

## DROP to\_table from STUDY\_TO\_OHDSI\_MAPPING table. To\_COLUMN?

The table a concept’s value goes into in the OHDSI world can be inferred from the domain\_id in the OHDSI concept table. This means man of the values that are erroneously in the observation table need to be corrected \*\*\*.

Can something similar be done for the value column used? Can we tell from the concept if the value is string, concept or number? Nothing obvious in the schema.

# UI for ENTERING NEW STUDIES – UPLOAD (later/deferred)

# DASHBOARD: INTEGRATING PeaX, RSTUDIO, MAPPING DEFN, PIPELINE

# CEDAR INTEGRATION

## CEDAR for MAPPING TEMPLATES and MAPPING DATA STORE

This is a point of group integration, dependencies and greater risk for getting blocked. I’d like to move forward in a way that involves minimal duplication of effort, yet maintains the option for running without CEDAR…that you can enter the data with UI that exists today if the CEDAR integration slips.

## CEDAR for STuDY LEVEL METADATA TEMPLATES AnD STORE

Not my circus beyond superficial integration. \*\*\* DISCUSS \*\*\*

# ADVANCED TOPICS

Some studies involve challenges whose semantics go beyond schema. I don’t see how a schema design alone could deal with these issues.

## ALLHAT (TBD)

ALLHAT for example involves many purposes that make the data appear like more than one study, but because of the overlapping nature of the studies and the study population, you can’t completely separate them. Our awareness of these issues started with the fact that a patient’s death status depends on which sub study of data you are looking at. One ended earlier and has earlier dates for End Of Study than others, so the same patient may appear to be known as a live and dead on the same date.

## RACE or RACE-ETHNICITY? (TBD)

Most do just race and consider Hispanic a race. TOPCAT considers it an ethnicity applicable to any race. If the result is to be harmonized we need to reconcile this.

## Drug\_EXPOSURE (TBD)

For the NHLBI pilot we dealt with concomitant meds as a simple Boolean value without regard to the exposure period. A more precise approach would make use of the Drug Exposure table that has places for the start and end dates and times of exposure as well as other details we may not have.

## PrOCEDURE\_OCCURRENCE (TBD)

Some concepts made it into either observation or measurement that should be in procedure\_occurrence: revasc SNOMED-8126608 and SNOMED-174911007, PCI SNOMED-415070008 and angioplasty SNOMED-41339005. SNOMED 167094009 “plasma glucose level” and SNOMED 394034006 “coronary heart disease medication review” don’t seem to be the right concepts because the concepts are procedures but I don’t think that’s what they were meant to be documenting.

# REPORTING (TBD)

The reports need to reflect the organization here and need to show local decisions like cause of admission being stored in the visit\_type\_concept\_id.

A search for synonymic concepts would help find places where different concepts where used that should be the same.

# RESULTING SCHEMA/CODE changes

## additional fields for study table

* For migrate\_functions.py and person.py
  + Id\_field, Id\_table
  + Id\_prefix.
  + Birth\_year\_column, birth\_year\_table
  + Sex\_column, sex\_table
  + Age\_column, age\_table
  + Race\_column, race\_table
  + Ethnicity\_column, ethnicity\_table

## CATEGORIZATION or EXTRACT CHANGES

If outcomes aren’t in observation, their categorization\_function\_metadata will have to change to where it is now.