GPC HackathonTwo Meeting Notes
Jan 22-23, 2015
San Antonio, TX

by Laurel Verhagen, scribe
review by Dan Connolly, chair, in progress

See also: HackathonTwo in GPC trac wiki with agenda, preparation, etc.; gpc-dev for contributions to and comments on the record.

Names, ticket numbers in section headings indicate follow-up items.

Day 1 - Thu, Jan 22

Introductions, Opening Remarks - GPC Phase 2 LOI

Attendees briefly introduced themselves, giving name and affiliation.

Russ Waitman reviewed the GPC Letter of Intent (LOI) for phase 2 of PCORNet.

- CDRN Phase 1 Goals: Engagement, governance (clinical trial readiness), collaboration (analysis ready data)
- CDRN Phase 2: analysis ready datasets, embedding trials in the health systems, oversight framework, insurance data (work to get CMS data directly to funnel back to each site; anthem healthcare) and other data partners, extend work to unstructured notes, increase collaborating with PPRN (phenotyping, etc.), connect with CTSA infrastructure.
    a. New sites: Indiana and Missouri
- Aspirin trial/proposed future model
    a. Typically $5000 to enroll someone in a clinical trial; this aims to do it for $250-300
        i. Cohort definition
        ii. Review by clinicians
        iii. Send it out
        iv. Consent at Healthy Heart Alliance (PPRN)
        v. Randomize
        vi. Integrate an alert back into the health record
        vii. Watching data accumulate
            1. Watch for bleeding events
            2. Watch SSA for death records
        viii. Patients could go back to Healthy Heart to enter PRO

AM Session 1 - GPC CDM ETL (#225, #145)

Nathan Graham presented on #145 and related tickets:

- GPC CDM ETL

1. Approach is to generate the ETL represented as i2b2 metadata; map GPC common paths into CDM paths. ref #145, #146
2. Example with demographics > hispanic
    a. Slide 3: similar query result using the CDM ontology
3. Slide4: inserted UMLS paths for ICD9
4. Slide5: written in SQL, python script for fast test iterations, loads the pcori terms csv file
    a. PCORI to GPC csv - maps GPC i2b2 paths and PCORnet CDM paths
    b. generates concept_dimension
5. Slide6: two paths map to the local concept code
6. Slide7: maps the CDM path to the GPC path
7. Slide8: not dependent on local concepts, the paths handle the mapping

8. Question: dimcode vs. concept dimension
    a. Started using dimcode, wound up with one to many mapping, which didn't work well; using concept dimension allows many to many mappings.
    b. Alex asked about using wildcards in the dimcodes; has a couple of slides that he could share to show this.
    c. Dan: Alex, have you worked out the SQL to handle this?
    d. Alex - yes, two concept paths pointing to the same dimcode, two different basecodes, two paths, mapped to the same dimcode
    e. Alex may provide an example of this work late on Friday
    f. Nathan is open to input/feedback
9. Slide9: then build the data
10. Slide11: SQL files for each table to select against the i2b2 PCORnet tables
    a. Python script does this for KUMC to build the CDM with the SQL files
11. Slide 13: Issues in gpc-pcornet-cdm bitbucket project
12. Dan: so we had a volunteer (MCW) and some interest (UTHSCSA) … progress?
    a. GeorgeK from MCW didn't get to this
    b. Angela Bos tried this yesterday; terms aren't fully aligned; didn't link to facts properly. Entries were added to concept dimension, but couldn't run patient sets.
    c. Phillip had done the first iteration of the code; created a mapping file; it seemed to work; created his CDM tables based on i2b2 for the Annotated Data Dictionary
        i. Nathan - this is a fork of the project for the Annotated Data Dictionary

After the presentation, attendees tried out the approach in workshop/hacking mode.

JRC offers to give example of CDM ETL screw case with polyhierarchy such as snomed. (cf discussion with Nathan…) (#225)

Reeder points out an issue with CDM ETL as presented: incompatible with GPC demographics paths "implemented" using patient_dimension. *Nathan noted this in ticket/145#comment:21.*

Angela asked questions about maintaining local codes/paths that are of use to local researchers. Could take a two step approach: map to GPC; map to PCORI

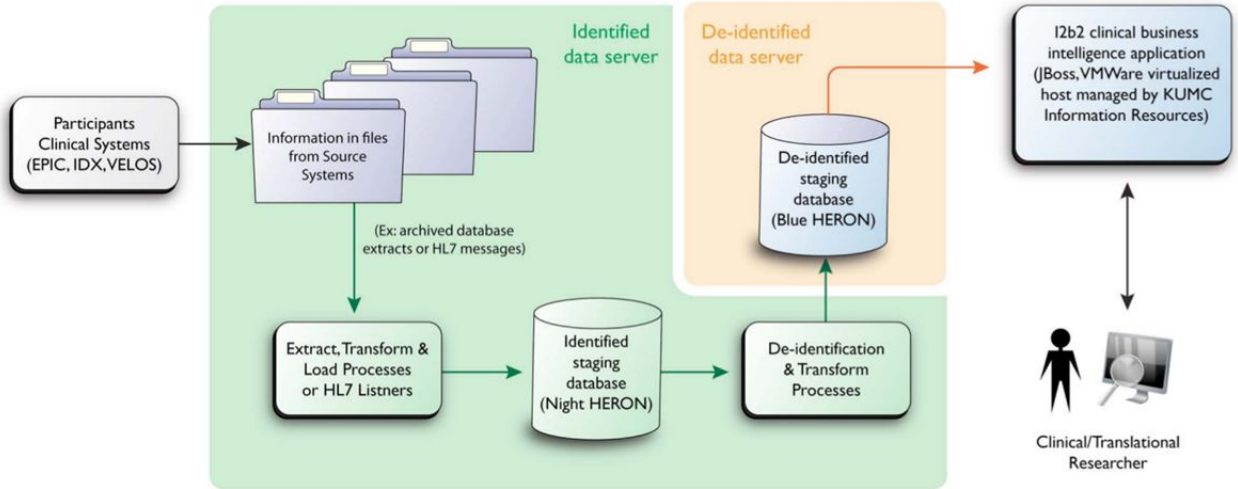
AM Session 1 - Heron Code Sharing

- Dan asks people to run heron code; it works with the specific KUMC use case, but sometimes other sites have issues.
- Dr. Waitman wants to push code sharing as it will benefit GPC for phase 2; how can we make this happen?
- Alex comments sites that will benefit are non-KUMC; sites should set up. The people who have problems should do the development with an emphasis on not burdening KUMC. How can we do this without disrupting KUMC?
- Dan mentions that other sites have something work; the benefit isn't in switching to using heron; not clear on the path to integrate.
- What are obstacles to using the heron code?

a. Angela:
    i. Infrastructure differences - (database) service names had to be updated
    ii. versions of i2b2 - mock i2b2 source has hardcoded strings to Harvard versions
    iii. Staging data pre-heron isn't well documented; especially if it was written 2 years ago
    iv. Other data issues - curated data files; EPIC file is different locally, how do I generated that curated file locally?
b. JRC: Lot of work from extracts into common/standard codes; not in heron (mapping is later with heron); expect to share that with GPC
    i. Dan mentions KUMC can't take that approach until their hospital picks up the mapping work.
c. Eric @ Marshfield (from chat):
    i. One obstacle we hit at MCRF is we are a non-Epic site. So any integration needs to be refactored to use our source systems. Occasionally we are able to utilize logic specific to calculated terms, but normally we end up developing in house. Use Windows/SQL server.
    ii. So, documentation on the code can be more important to the code itself at times.
d. Hubert (from chat):
    i. Differences in Clarity versions also causes some differences -- our infrastructure also drove some of the directions that we took.

identified i2b2 (Mosa@MU)

- Jim Campbell: UNMC is Bringing up an identified and de-id'd version. Heron code seems focused on de-id'd side.
    a. Dan: actually, it does both. Perhaps that wasn't well communicated. Recall [HERON architecture]:



b. Alex mentions that: Blue heron (de-identified) code can be used with Night heron (identified) data. A front end is now pointing to
    i. Heron code develops an identified repository, then the de-id'd version (blue heron); originally only had a front end on blue heron (de-id'd)
    ii. Night heron originally didn't have the metadata concept dimension
c. Tom asked if sites were working on identified versions? Several sites
    i. Tom mentions IRB challenges to approving a database for identified data.
d. Missouri offered to talk about governance for identified i2b2.

AM Session 2 - Breast Cancer Survey Finder FIle (UMN #110, Iowa, MCRF #217, UTSW #110, KUMC #227, MCW #221)

1. Dan asked Tamara if any sites are having trouble finding data elements for the query
    a. No
2. Elements not in tumor/NAACCR?
    a. Demographics / Language
    b. Demographics / Vital Status / Deceased
    c. Demographics / Vital Status / Deceased SSA
3. Date shifts (#101)
    a. Approach to handle de-identification, identification
    b. Process set up originally did not retain a connection to identified data; if you wanted identified information after reviewing de-id'd, then re-ran/requested the query against an identified source.
    c. Alex has an idea to handle this; worked it out via napkin on 1/21. Obesity will be different than BC; IRB will be complete, unlike BC.
    d. George K to confirm with Dan Hale that IRB will be complete prior to the Obesity data request #221
        i. postscript: George: I talked to Dan and confirmed that yes, this is the intention.
    e. Governance with BC?
4. Installing DataBuilder or workaround process
    a. KUMC - done
    b. CMH - n/a
    c. Iowa - unknown; date 1/20
    d. Wisc - data sharing agreement needs to be readdressed; technical issues with SQL server; health concerns; interim solution
        i. Hubert suggested alternate methods to make the python easier to install
        ii. Dan agrees that it's ugly to install.
    e. MCW - done
    f. MCRF - technical issues Dan to help Joe, getting close; IRB/approval to release question;
    g. UMN - interim solution; builder is on their list; waiting on approval to release file
5. Submitting the resulting data set (#211)

    a. Received 5;
    b. Need 4 (UMN #110, Iowa, MCRF #217, UTSW #110)
6. **QA - KUMC (Tamara, …) will do some work here (#227)**; Wendy is the statistician
    a. May result in questions/new requests back to the sites
    b. Tamara mentions that it may be necessary to re-run the queries with a different date if the methods core finds that sites don't have data through May 2014.
    c. From BC meeting 1/21: Tamara asked about possibility of sites providing a diagnosis date distribution chart?
        i. Betsy wants a chart to see how many were diagnosed in January 2013 vs. 2014; aggregate data
        ii. Can sites share the number of bc diagnoses by month (not shifted)?
        iii. Dan to discuss this with Russ.
7. Step (f.g.) to filter by real date got integrated into the original request.
8. re-ID and verify cases with tumor registry
9. Site HB provides mail contact information to the site BC PI/lead
10. All WP-1 steps must be complete by March 2.
11. Remaining Breast Cancer Work plans:
    a. Workplan 2 - Sending surveys
    b. Workplan 3 - Getting survey back
    c. Workplan 4 - Investigator using the survey data

AM Session 2 - Obesity, BMI percentile (#210)

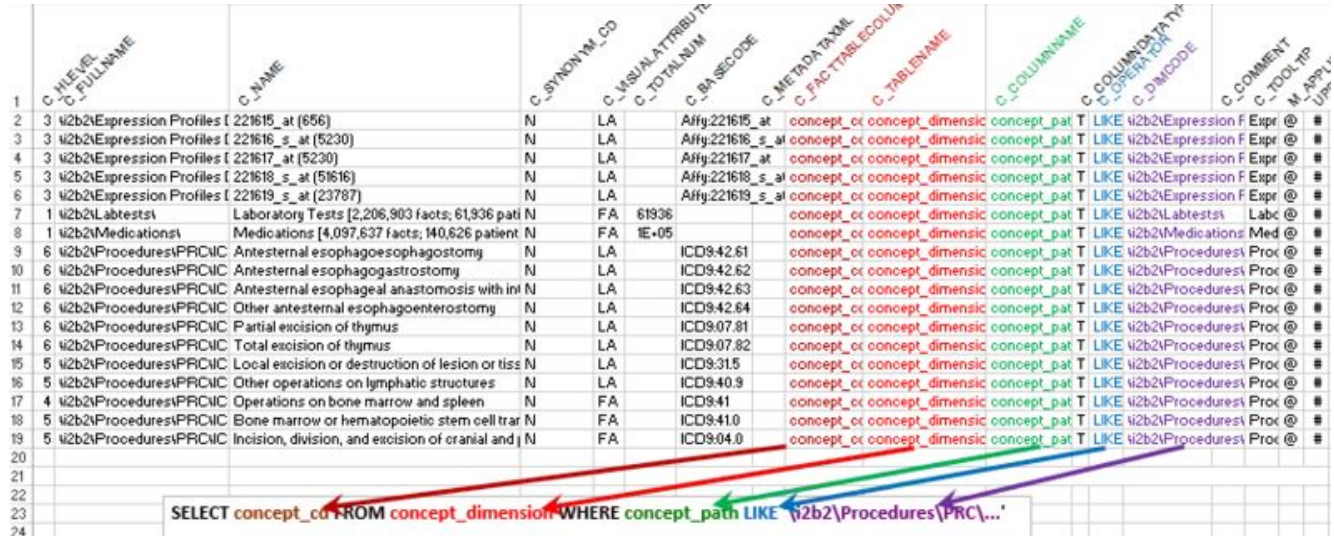
1. Working from a spreadsheet; download it from google docs; if you open it, could mess up the formatting
2. Sent two files to GPC dev for implementing BMI percentiles in an EMR-agnostic manner.
3. SQL dump generates a lookup; asked if anyone had reviewed the files distributed at 1:30AM; no from attendees.
4. Run the SQL file in the schema where you want the CDC lookup to live. Questions?
    a. Jim: why bins, not the actual percentile?
    b. Alex: This was what I could access from CDC; thinks there is a way to calculate and convert to z-scores
5. Next script generates a lot of temp tables; need to update from blueheron.
    a. pr1 = percentile 1
    b. joining from observation fact to patient dimension (birthdate); sex recoded to 1 or 2 (might need to update this);
    c. selects fields of interest; goal to filter patients to make sure they're between 2 and 20 years old.
    d. All BMI observations from kids are replaced with percentile.
    e. Inserts new concept into concept_dimension that the metadata tables can target.
6. Alex: comments on this process?
    a. Nate A: Is there a standard ontology for this section?
    b. Phillip: not yet

7. Alex: Does anyone know how to access BMI from EPIC or other EHRs?
    a. Nathan: I looked and didn't find an easy way to do this, so I used the CDC tables etc. instead.
8. Alex: What would it take to implement by site?
    a. Jim: All pediatric care is at a separate institution; is this worthwhile? Mentions having 0.02-0.03%;
        i. Dan Hale mentions that those sites would use an equivalent population under the age of 49.
    b. Alex: This could make an obstacle for participation in this type of study.
    c. Phillip: same issue
    d. Discussed across sites; comments or issues should be sent to Alex.
    e. @@All sites with significant pediatric data will do The Right Thing
    f. Decision that this should be LOINC coded. Has the reference codes.

AM Session 2 - Terminology Mapping Strategies

Alex Bokov presented a strategy related to CDM ETL.

9. i2b2_terms.xlm (from the metadata schema)
    a. Explained columns and queries



    b. If path contains a % and you use ORACLE, this could be interpreted as a wildcard.
    c. Could end a branch with a %, which would match all codes downstream
    d. Returns concept codes
    e. Codes are the only thing you need to link the search terms to the entries in observation_fact
    f. Dimode doesn't need to match c_name, etc.
    g. If you have an existing concept and you want to add a new ontology that matches that concept with a new path, just add an entry that updates
        i. c_fullname (path)
        ii. c_basecode (does not need to match concept code)

h. Find a way to map to a shared set of paths/basecodes
    i. Jim: if we use standardized codes across sites, queries can also run across sites.
    ii. Example: Medications at KUMC are mapped to RXCUI, but observations use native EPIC codes.
i. Next steps: Alex, NathanG.; update tomorrow morning.

PM Session 1 - Terminologies

Demographics: Email, Enrollment (Reeder #87, Campbell #229)

1. Demographics: Added language, vital status
    a. (Dan C. checked: zip is gone on babel)
2. JimC: Are any of these deployed as Observation_Fact variables that allow you to enter in a value for the term? LOINC coding and then the value associated with the observable.
    a. Phillip: was using the standard i2b2 approach with a list instead of a tval char
3. Email address
    a. TomM mentions that Umberto was requesting the ability to define a cohort with this variable for Phase 2. Is it possible to add?
    b. Tom: I gather studies are coming to GPC that want a “we have an email address on this person” flag.
    c. Group decision to add email to Demographics. (Reeder to add to babel, CDT)
4. Enrollment
    a. UNMC added “Two encounters or Wellness” - addresses the active definition: two visits in three years separated by 30 days or more.
        i. JimC: or if someone comes in for a Wellness visit, it’s okay for one visit to occur.
        ii. JimC will distribute SQL for this. (#229)
    b. This is on the list for consideration. *(Dan C. to ensure there’s an open ticket: #229)*

Diagnosis Modifiers (Reeder #90)

5. Dx - ICD9DM
    a. Added modifiers where there are multiple sources
    b. **All agreed** to; Phillip just created the metadata for them

Medication Modifiers (#199)

6. Medications
    a. Still talking about modifiers for this; med_admin (Kansas uses this), ordered meds, UTSW uses current meds (as prescribed at a visit), rx filled (surescripts data)
    b. Are there other modifiers that sites need? Could add if it makes sense.

i. Jim asked about inpatient; Phillip shows that Ordered includes two leafs (inpatient and outpatient)

c. Dan - this isn't required; is it worth the time?
   i. JRC: I expect our (UNMC) researchers to ask for this
   ii. ours too (CMH)
d. Alex - isn't there a need for history vs. active?
e. Phillip - the intent was to distinguish the source; high level, easy to get
f. NateA: What does current Meds mean?
   i. Phillip: anything that was included on the encounter.

Vital Signs (Reeder #23)

7. Vital Signs
   a. Currently have GPC measures, which includes a couple vital signs
   b. Proposing to take what Bonnie provided; flat vitals hierarchy; map to them for version 1. Consistent with the LOINC codes provided by Jim.
   c. Dan - didn't we agree to just use Clinical Measurements? #23 is closed
   d. JimC - mentions working with Epic to get LOINC coding associated with flowsheets.
   e. Need decision between flat vs. extensive hierarchical option
      i. Important because we need to know which path to use for obesity query.
      ii. Reviewed an alternate strategy used at CMH.
      iii. JimC - most important decision is a standard basecode.
      iv. Dan C.: for prospective trials, we need standard codes in the EMR, but for near-term retrospective query requirements, we're interchanging i2b2 queries, which only include paths. So is it OK to make a decision on vitals ontology v1 independent of codes?
      v. Jim C.: if we agree on the codes to go along with the paths, I have no preference between various pragmatic organizations - flat, hierarchy.
      vi. Group asked for opposition to the flat model - WISC
      vii. Group asked for opposition to the hierarchical model - none
      viii. Group agrees to promote the Clinical Measures / Vital Signs up one level. Will use LOINC.
         1. All present agree. No one present disagrees.
         2. GPC Vital signs will follow this for paths.
         3. Noting the prioritized list.

PM Session 2 - Data Quality (Mish, KUMC #159)

Tom presented Hackathon 2 Quality Discussion.

1. Quarterly response - next due in March
   a. JimC - PCORI said we needed a plan for quality; in the proposal, Jim said the quality data will be the reports we worked on for cohort characterization. That's why they prototyped this based on the obesity work.

b. For the last submission, Tom spent lots of time entering results into REDCap
c. **Proposal that each site will enter own results into REDCap**
    i. Code should dump a CSV to import into REDCap
    ii. Get each site a login for KUMC REDCap?
        1. Dan: using redcap.kumc.edu was an interim measure; the cloud-based GPC REDCap instance should be available in this timeframe (#159)
    iii. Group agrees that this will be the process for March
d. Queries/output for March
    i. Who can help write the queries?
    ii. Phillip - can PCORI CDM mapping be required
    iii. Tom - yes.
    iv. #193 GPC Obesity Worksheet - starting point; this will be similar to the REDCap form; run query for each term.
        1. Jim: Notice this starts with an Enrollment population (not all patients)
        2. Set up paths for all from standard LOINC and PCORI
            a. Requires updates based on GPC decisions.
        3. For sites that want to just use SQL, specified modifier codes.
        4. Python or SQL to distribute? No python. **Just distribute SQL code. (Mish to follow up)**

2. Quality metrics
    a. Demographics - how about we revise to match GPC demographics finalized today?
        i. Drop religion.
        ii. JimC: Need to include Enrollment population.
        iii. **Group agrees to revise; will need to add Enrollment**. (see enrollment section above) (#229)
    b. Data Analysis
        i. Male/Female: General population is 55% Female; 45% Male; should we be checking site/site for outliers/data issues?
        ii. Begs the question of how to check for coding errors.
        iii. Alex - should we be sharing issues we identify locally?
        iv. Dan - MCW mentioned they have basic checks that the counts that go in match the counts after loading. MCW indicated they will share.
        v. JimC - may need more validation on the CDM valueset.
        vi. Start building quality checks for GPC if we can't get this from DSSNI/PCORnet?
            1. Dan mentions that there are tickets with **data-quality** keyword
    c. Tom - proposes that **if there is a quality issue, create a ticket; before the ticket is closed, gpc-dev should decide if the issue should be integrated into quarterly quality checks.**
        i. Example - #196 Active Patient/Enrollment Issue

ii. **Add methodology in the quality checks to include what sites use for Enrollment. (Mish)**

PM Session 2 - Encounters (#120, #155)

1. Dan C. asks: what does financial encounter mean in i2b2?
    a. KUMC -
        i. Pat_inc_csd_id (?)
        ii. HSP_account_ID in Epic for hospital scrubbed data
        iii. KUMC tried everything that happened to a patient on a day (patient day)
        iv. But...Not sure at this point if it solved the problem; no duplicate key issues.
    b. Tom - what do other sites do?
        i. Cerner - doesn't have the problem; uses an encounter note
        ii. Marshfield - patient, date, facility, encounter_type
            1. Jim questions if this will properly handle inpatient/admits. Multi-day admissions should be one encounter.
    c. Jim mentions that PCORI provides logic:
        i. ambulatory: multiple visits to the same provider on the same day are one encounter; multiple visits to multiple providers on the same day are two encounters.
        ii. Hospitalizations would one encounter
        iii. All seem to be open to changing to a defined standard.
    d. Hubert distributed relevant
        i. for ref: Monday, November 03, 2014 3:55 PM fom Wanta has SQL for encounters.
        ii. Nov 11 code from Hickman pcori_encounters.sql
    e. PR: We can align GPC Encounter terms with PCORNet CDM columns independent of deciding how to correlate encounters
        i. DC: OK; I'll split the encounter ticket so that alignment with CDM is separate from same-financial-encounter functionality.

PM Session 2 - Text Deidentification (Jacquie @ MCRF)

Jay Urbain presented…

1. Background on work. Found two de-identification packages;
    a. MIST - lots of work to set up, three languages (java, scala, python); not multi-thread; not as high as published results. Complicated to deploy. Didn't work well in the i2b2 challenge.
        i. Name entity recognition wasn't poor… at least with MCW data.
            1. Missing a patient name was the same as missing a phone number, zip code. Most important part is name… not zip code, so results were misleading

    b. Also evaluated against Stanford NLP group, Apache's Open NLP, (tool released soon written in perl)
    c. Developed a multi-thread tool with solid name entity recognition. Performance is 100s of records/second
    d. 97.3% accurate overall
2. Process
    a. Patient
    b. Text Preprocessing
    c. Blacklist de-identification (words the system misses; around 12)
    d. Whitelist de-identification (medical procedures named after people)
    e. Regular Expression Processing
    f. Date Shifting
    g. Named Entity Recognition
    h. De-Identified Patient
3. Available on bitbucket
    a. https://bitbucket.org/MCW_BMI/unstructured-notes-deidentification
    b. need data to evaluate; also has software to create a de-identified fabricated set.
4. Group mentioned priming the de-id tool by passing in metadata like MRN to pre-screen.
5. Phillip - how are the notes being delivered to users?
    a. George - Text search on a field - returns count of matching patients.
6. Jacquie - how does it differentiate between "will" and "Will"
    a. Jay - surrounding words, capitalization, training, 40-50 probability models it's trained on, sentences in a column format, each word in context has a label. Conditional rain of field.
    b. Phillip - mentions a place that used medical publications for training purposes on the medical terminology.
7. Dan - what was the process to get this deployed at the institution
    a. Had permission to used a paid tool already; which made it easier to get approval.
8. Phillip - can you search for specific note types
    a. George - yes, you can drag over the types and then do the search.
9. Matt - got the program working on note text.
10. Phillip - what's this based on?
    a. Based on Stanford. Focus was speed.
11. **Marshfield (Jacquie volunteers to work on this).**

DAY 2

AM Session 1 - Usable LOINC Lab Hierarchy - (Apathy #158)

Nate Apathy lead discussion; presentation materials:

- Usable LOINC Lab Hierarchy

1. CMH uses an assay-level only (individual lab consult; not panels, orders)
    a. Paths are moderately complex
    b. Information is available on #158
2. GPC current ontology
    a. Flat design; too many child records
    b. Codes are difficult to navigate; easy to search, simple paths
3. UNMC - Lab
    a. Includes orders and panels; still have lots of leaf nodes
    b. Codes live in multiple locations within the hierarchy; multiple paths
    c. JRC: (not sure how to note this…)
    d. Panels can be very site specific with Epic and Cerner depending on the build team.
4. How to handle panels/discussion
    a. Is an individual lab different than a lab that's part of a panel? Is there enough value in adding this level of detail?
        i. Is a sodium panel a sodium panel and a sodium panel?
    b. JimC - UNMC maps to the same lab code if the test uses the same materials.
    c. NateA - this suggests not including panels; priority is the actual LOINC code; the code tells you specifically what was done (unit, instrument, reagent, etc). Can decide on generic codes or each site can provide specific codes used at their site.
    d. NateA - the folders are ranges of LOINC codes and expresses the collection of codes.
    e. JimC - same approach could be used to populate panels that are important to a clinician? If there are multipe LOINC codes, bring them in under chem panel, but collapse them.
    f. NateA - Are you anticipating users dragging over the "chem panel" term?
    g. JimC - Most important that a user knows that they have information on sodium, if all is in one method and it's not under chem panel, but that's where everyone expects to find it…. then that's a bad thing. At least for common panels. There aren't many of those.
    h. NateA - if there are multiple codes for the same conceptual test, it will clutter the view within the folder.
    i. JimC - Most important for GPC might be glycohemoglobin. Make sure those are organized. Make sure that we can see there is data within this area across sites, even if the LOINC codes are slightly different.
    j. TomM - I hear that someone wants all "glycohemoglobin"; doesn't even care about the values, just want all and want to know if it's high/low.
    k. Dan - how is this not available?
    l. Dan - I hear a lot of researchers say "give me hemoglobin a1c"; they know what to get
    m. JimC - recommends rolling out a combination of hemoglobin a1c tests to ensure that we can pull all of the codes possible across sites.

n. Phillip - I have an option to show as well. Someone with the CTSA act project did work grouping CDM labs.
o. NateA - should we focus on CDM labs first?
p. Dan - do we have a list of GPC requested terms?
q. JimC - there are codes for obesity that are not part of the CDM.
r. JimC - CMH and UMN removed all LOINC codes from their hierarchy that aren't present in their dataset. May be possible easily build a list of all LOINC codes across sites.
s. NateA - we hide terms that aren't present.
t. Dan - can be an issue. If you delete the terms, the query won't run.
u. JimC - curating a hierarchy of all LOINC is a big pain.
v. NateA - our ontology table is enormous, but the list that's seen is shorter.
w. PR: in the Regenstrief hierarchy, do the terms occur in multiple places or just once?
x. NateA: just once; the Regenstrief hierarchy is a mono-hierarchy; we expect users to search and be happier with a hit in just one place in the hierarchy; consultant piece is to look by test, not panel.
y. NateA - issue with loading LOINC panels can be that it doesn't return patients with the panel… just patients who have had any of the components.
z. Dan - does Regenstrief include if tests are quantitative, units, etc.
aa. Jim - it's possible to access/download that information about the LOINC codes along with units.

5. Design decisions
    a. Code agreement
    b. Should GPC…. a.) normalize to the same codes or **b.) keep local (site specific) LOINC codes and map to common codes where the CDM requires it in the GPC LOINC ontology**
        i. **no objections to option B.**
        ii. JimC - proposes to load LOINC, then compare counts across sites to see if there is already agreement.
        iii. NateA - the folder strategy might avoid the need for comparing
        iv. Some discussion of variance/mismapped codes/fat fingered lab requests
        v. Phillip - Action then is to build out LOINC with metadata to query by value
    c. Do we want to keep local hierarchies? Keep proprietary LOINC hierarchies with different paths, but then agree on CDM compliant paths?
        i. JimC - Need to go with that. The standard GPC paths can be used for queries.
        ii. Phillip - my basecodes are my componentIDs (not LOINC)
        iii. Dan - simplest method is a mono-hierarchy, where each code only occurs once.
        iv. Hubert: Jim's comment is good - even if the complete hierarchy is not exposed to users it would be compete for shared queries across GPC, and would include all of the CDM LOINC codes as well

v. Proposal: use the CMH approach with a mono-hierarchy
    1. Details: current state slide from Nate's presentation
        a. Assay-level only
        b. No orders, panels
    2. JimC - why would we need that if we've agreed to all use LOINC?
    3. Tamara - Will people be searching by panels?
        a. JimC - it's not a requirement for a GPC shared query; it is a local requirement.
        b. Tamara - if needed, will panels be brought in? Researchers at KUMC do search by panel. Don't know if there will be GPC researchers needing panels.
        c. Dan - revisits the point that searching by panel will return results for the components.
        d. Tamara - probably okay for GPC.
        e. Phillip - I think this is very straight-forward
    4. Angela - is this useful enough for researchers without the panels?
        a. NateA - we push the same ontology for all clients; never had pushback on the LOINC ontology; has had pushback on NDC and others. Typically researchers search for a specific test.
    5. Tom - Can there be version numbers added to the transformation code used to build the ontology?
        a. JimC - since labs will never be removed, this shouldn't be an issue.
        b. Dan C.: but it costs very little, so it's worth adding just in case.
        c. NateA: easiest strategy should be to generate the metadata and distribute the results.
    6. Decision: **Adopt the CMH strategy**
    7. **Follow-up question from Jim: does this include Metadata XML?**
        a. **NateA - yes, it's on his to do list!**

AM Session 1 - NLP/Text Notes Code Sharing

1. Since Jay's presentation yesterday, did some work.
2. Dan - there's a way to download a zip of the source.
3. Process - George unless otherwise noted
    a. Included test program to populate mysql data with names, date shifts
    b. Will need to update the testDeID.sh script
    c. Required fields: id, note_id, note_text
    d. Threads: Jay recommends increasing the thread count by 1-2 over the number of cores available. How many records should a thread process at a time: if you have

a lot of memory, then you can allocate more space to each thread. Basic parameters should be good 90% of the time. Will be slower with 2 core.

e. Run on multiple servers; copies to Obervation_fact, observation_blog, tval char is set to binary, which allows i2b2 to search by character. Text searching needs to be turned on in ORACLE. Need that for this to work.
f. Feel free to add issues.
g. Wiki will be updated with better documentation in the future.
h. Once building outside their environment
    i. Matt - should this be available to GPC?
    ii. Should this be shared in the cloud:
    iii. DB platform question from the group.

4. Questions
    a. Dan - GPC date shifting approach is 0 to -365 (cf #73)
    b. Jay - needs to be a ticket; software may not handle 365, just did +/- 15 to shift within a month
    c. Dan - Asked for an update on Matt's install;
    d. Matt - was tuned for MySQL; got it to work with MySQL on his laptop over the VM suggested by MCW. Didn't want to use the VM, put it in Eclipse on his Mac, then tried to tunnel back to a data source at KUMC (Oracle?). Changes to the code were required. Committed code for Oracle support. Fiddled with it; it worked; Showed Russ; Russ was happy. Saw it work for 1000 notes; saw some nice replacing on a data source.
        i. NLP missed some things. Suggestion with knowing the PHI and adding per-record/text line blacklist would catch some of the issues he saw.

AM Session 2 - Federated login (Mish #159)

1. #159, Number of accounts required for REDCap data sharing is more than is manageable at KUMC, including researchers
2. Need to figure out how to get federated logins working for REDCap.
3. Which system do we want to use?
    a. Shibbolith? Incommon?
4. Phillip - I've worked with Shibbolith
5. Justin - not handled by their team at UMN; no experience
6. Tom - I have a person looking into Shibbolith problems in Madison; **Tom volunteers UW.**
    a. Tom will connect with his person on Monday
    b. Dan will ping Tom after the gpc-dev call in the afternoon on Tuesday, Jan 27.
    c. Tom already wants to federate a couple of departments into the REDCap instance.

AM Session 2 - Building Analytic Datasets

Some of the stuff Alex was talking about is implemented in the same repository as the data builder. Dan will demonstrate what the code does; this will help Alex. VinceL (KUMC) has developed a prototype interface.

Using heron_extract to reshape data for use in REDCap

Connolly demonstrated...

1. built.db is a sqlite file from the DataBuilder plugin / DB Browser for SQLite
    a. Browse Data; reviewed a few tables.
        i. Job table and variable table are the inputs: drag over job file, adds concepts which are the variables.
        ii. ignore underscore dt
    b. No questions asked on this background.
2. Overview
    a. DFBuilder.py, edc Summary.py (electronic data capture summary)
    b. Finds patients and builds a cross section table with views.
    c. Generates redcap_metadata (data dictionary for REDCap)
        i. Patient number, patient form, demographics form, etc.
    d. Also generates redcap_data
        i. All elements out of i2b2 is rendered on this form.
        ii. one row for every patient and variable (or can organize around encounters);
        iii. first date and last date that something occurred in the data
            1. Example: if aspirin was dispensed 100 times during a week, then this would tell you 100 times for the frequency, beginning date of the week and last day of the week.
        iv. python takes the REDCap API and pushes it into REDCap
        v. This is what KUMC uses to push data into REDCap
3. Alex: can we look at the cross_section again? we don't understand how the frequency is used?
    a. Angela - is this based on encounters?
    b. Dan - the consumer decides if the data should be organized by encounter or patient.
    c. Angela - No encounter grouping then?
    d. Tom - it's the pat_inc then?
    e. Dan - it's just the encounter logic.
4. Phillip - could you show the steps to generate the BC file? Without Jenkins
    a. Dan - CDW export is what you have
        i. First mocks up the export
        ii. command line arguments for by_encounter or not; identified (would put MRNs in REDCap) or not. keep_redundant (if you put in asprin and medications, can have redundant or not)

iii. shows variable table for testing purposes
iv. code values must be alpha numeric for REDCap, so it's HEX encoded. (SQL.hex)
1. Alex: asked whether this is what's pulled out of REDCap if the researcher exports.
2. Dan: yes - no motivation for anything different.

b. Phillip - what's the next step to load it into REDCap? Create project, load data dictionary… etc?
i. edc_summary.py
ii. Dan - give it summarized_dc, project_key for API
iii. Loads in chunks for performance.

c. Phillip - can I get the 5 steps documented? `typewriter font` for commands; *`italics`* for stuff that varies from run to run.
i. Given: *`breast-cancer-utsw.db`*
ii. Summarize: `python edc_summary.py deid` *`breast-cancer-utsw.db data-dictionary.csv`*
iii. Make a redcap project based on that data dictionary; get a API key (needs write)
iv. Upload: `python edc_summary.py upload` *`breast-cancer-utsw.db X123`* `--url` *`http://redcap.right.place/API/`*

d. Alex - Showing the current data builder interface may help; there are 3 fields:
i. Top - drag a patient set
ii. Middle - drop terms that will be variables
iii. Bottom - (will finish later)

Analyzable Data (Bokov #228)

Alex presented

1. Observation_fact isn't analyzable, because each case is it's own row; most routines aren't built to handle this format. Not analyzable in R or SAS. (can hack, but not the job of the statistician)
2. Example of analyzable data; can plot one column against another. Similar to by_encounter/per visit.
3. Fit A Regression Model; textbook, statistical use case, testing for significant relationship between one variable and one or more variables.
4. DataBuilder's SQLite file
    a. Per patient dataset is a per visit dataset with only one visit/special case.
        i. The one row per patient model compresses the detail.
    b. Domain experts (stats people), if you have two visits for a person and two blood pressure records, then the average blood pressure might be okay; maybe the final blood pressure after an event - lots of ways to compress a time series into one data point.
        i. Which to choose?

    ii. Comparing median measures are meaningless.
c. Recommendation: preserve time structure. treat time as a variable.
d. One row per visit dataset, facts as columns is the best way to accommodate the max number of analytic use cases.
    i. If several rows per patient is an issue for the researcher, that user should talk to a statistician.
e. Phillip - this is why I like to give them the SQLite file; then they can do whatever they want with it.
    i. Example: want all right before the surgery? what does before mean? 5 days? 30 days? 365 days? Depending on the question, any of those could be the right answer.
    ii. Need a conversation with the clinician and stats person to figure out exactly what they want.
f. Dan - Visit is one of the least reliable variables in Heron… may work, may not. if you want to combine tumor and health data - it doesn't work.
    i. Alex - in this case, I mean records by patient date. This is manual; have to hope that values didn't change within the same date/time that would result in a loss.
g. For each data domain, need to gradually develop rules for handling modifiers and aggregation logic.
    i. Example: if someone pulls allergens, there may be modifiers
    ii. Indicator variables: should some data elements just be a "yes" the person got that panel?
    iii. Alex plans to clean some of this up and distribute it (see screenshot below); need to review.
    iv. JimC: we should consider the metadata_xml as a guide for how to place the data in a file.

Query Type	%|ALLERGEN:%	%|FAMILY HISTORY|DIAG:%	ICD9:% and %|DX_ID:%	%|COMPONENT_ID:%	%|FLO_MEAS_ID:	%|MEDICATION_ID:	%|PROC_ID:	DEM|ETHNICITY:%	DEM|%	%|PAT_ENC:
stem node	1	1	1	1	1	1	1	1	?	1
stem node, mod	1	1	1			1	1			
leaf node, numeric				1	1	1				1
leaf node, any	1	1	1	1	1	1	1	1	1	1
leaf node, numeric, mod						1				
leaf node, any, mod	1	1	1			1	1			
leaf node, cat			?	?	?					
leaf node, cat, mod						?				
Additional Fields										
MODIFIER_CD	2	2	1	1*	0	2	1	0	0	0
VALTYPE_CD	0	0	0	2	1	1	0	0	0	1
TVAL_CHAR	0	0	0	1	1	1	0	0	0	1
NVAL_CHAR	0	0	0	1	1	1	0	0	0	1
UNITS_CD	0	0	0	1	1	1	0	0	0	0

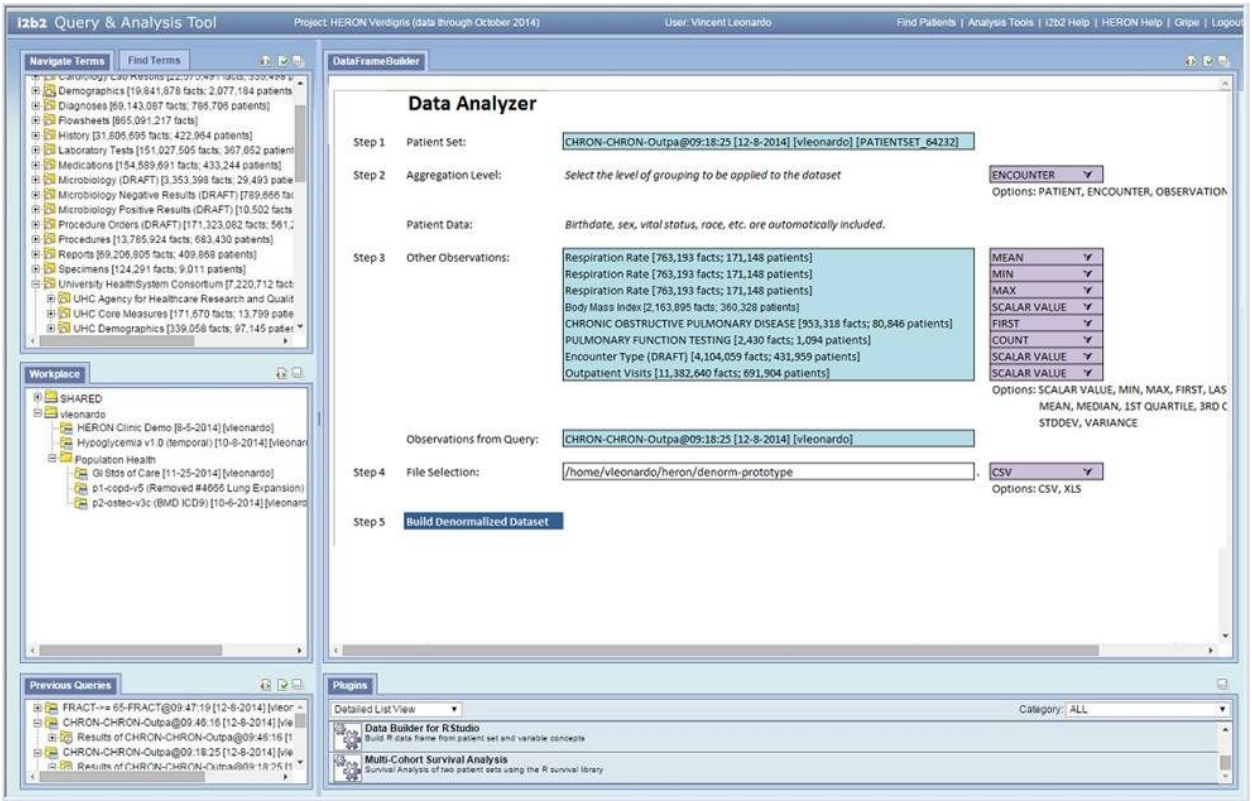
Questions:

What does the generated SQL in the log look like for each type of query?

What should be the default actions by data-builder within each visit for each set of observations pulled up by that query?

Data Analyzer User Interface

Vince Leonardo @ KUMC presented some ideas.



- example output table

1. An idea to provide a new plugin tool to i2b2 -- that would generate user-customized analysis datasets
2. Formulate the dataset for any concept that you want to pull out, identify the level of aggregation.
    a. Example with 3 aggregated values: min, max, median
    b. Can pick many types of aggregation. This would be exported as an actual xls
    c. Could get someone started without the completely detailed observation fact.
3. Output
4. Biggest benefit is just to request the level of detail to include
    a. Aggregate at the encounter level
    b. Aggregate at the patient level
5. Alex - if you're aggregating by date, how would you aggregate an ICD9 code
    a. Vince - in this example, aggregating at the encounter level, asked for the ICD9 code; need to know what level is appropriate.
    b. Alex - it may not be well defined. Meant aggregating the patient date;
    c. Vince - asking for the first one for the encounter.

d. Alex - aggregate by date means all the data for a given patient for a give date and put it on a row.
e. Vince - 3 levels available - patient, date, none (all rows appear independently)
    i. Alex - proposing patient+date
f. Alex - likes the aggregation level, because it forces the researcher to consider this question in a visual way that isn't confusing
g. Vince - yes, helps them define what they want in the final dataset. Translates that to the user interface.
h. Alex - my approach was having to turn the observation_fact, etc. into data; this offloads some of the stuff that should be done by the user to the user... instead of trying to make assumptions and do this for the user.
i. Vince - yes, how do you normalize for a user? First, last, average, minimum, max? That lead the idea of pushing the design process into i2b2.
j. Alex - do you have time for this development?
k. Vince - not decided.
l. Phillip - is this worthwhile, given the number of variations of data that will be coming through? Unless you look at the full dataset, how will you know what you want? Won't you need the full file to figure out what you need? Is it better just to give them tools to analyze the SQLite file? Most people don't know what they're looking for…
    i. Dan - we can put this over the DataBuilder?
m. Alex - none of this changes the need for writing the functionality to allow the exploration that Phillip mentioned.
n. Phillip - a Cohort Characterization tool would be useful
    i. Top 10 dx from a dataset
    ii. Top 10 labs from a dataset
    iii. Dan - There's a ticket for that! Anyone is welcome to take it. #138
o. Alex - who's worked on i2b2 plugins?
    i. Dan

PM Session 1 - EMR Integration

Jim McClay and Courtney from UNMC presented:

- Implementation of Research in the EHR
  Courtney Kennedy, Clinical Research Advisor
- CER in EHR
  Courtney Kennedy, RN and James McClay, MD

1. Incorporating research into practice - Jim
    a. Milestones - collect patient generated information to the recruit for clinical trials, integrate into EHR, randomize
    b. PRO Survey
        i. Replies: 6 of 10 institutions (missing KUMC, MCRF, UMN, UW)

        ii. Results: all have REDCap and a patient portal
        iii. All can capture PRO in a structured format
        iv. Different strategies to administrate and capture
    c. To capture PRO with GPC, should agree to use REDCap for a uniformed mechanism
    d. Engagement to address the divide between the health system and research
    e. Workflow - how to deploy to all sites?
        i. Build at one site in REDCap, then share
2. In Operation/Implementing of research in the EHR - Courtney
    a. Background as a Clinical Research Advisor
    b. Objectives of integration:
        i. Integrate research + standard of care for trials
        ii. integrate clinical/billing operations + research compliance
        iii. integrate data collection in clinical workflows for trials
    c. Reviewed research tools at UNMC

PM Session 2

Attendees discussed and worked on various topics independently and in small groups.