**UCD – SNOMED – Neo4j Project**

*Update, December 19, 2019*

**Objective:** To complete this project by creating a working demonstration that uses a combination of tools (specifically, some data query/visualization (?) tool or process to augment the Neo4J graph database that we have) to leverage the SNOMED CT semantic relationships and display them to experts in such a way that they can discover patterns in the data that might reflect previously unknown relationships about the disease. These patterns can then drive the design of studies to test these relationships using traditional statistics on data exported from the graph database.

A successful demonstration can consist of one or two phone calls with clinical/disease experts who will comment on our data visualization, and ask subsequent questions that we can iteratively query and display until they find a novel association that we can test using traditional statistics. In addition, we will develop a paper (target to the *Journal of Biomedical Informatics*) to clearly explain our data set, the value of SNOMED CT, rationale for the tools, the steps we took to process and analyze the he clusters and visualization to show the discovery process and enable others to replicate and extend this work. (Note from RR: the paper might not cover a visualization but just show data on value of SNOMED and RxNorm for queries…)

**Status:** We have the Neo4J graph DB up and running and are looking to add an interactive query or visualization (?) piece (given that we cannot afford the Neo4J Bloom viz tool) that can help us present data and relevant SNOMED relationships to disease experts that can identify patterns that have clinical or biochemical plausibility. The challenge is finding someone that is familiar with these tools and the semantic relationships in SNOMED CT and other terminologies.

**Next steps:** *(Rachel to coordinate, December 19, 2019)*

1.) Succinctly summarize steps to date and what the database looks like. (See *Dataset and Database Description* sections below)

2.) Define clinical questions/exploration topic/goals for demonstration (what is important for UCD researchers to know or explore? (See *Clinical/Research Questions* section below)

3.) Define an approach to develop the demo and evaluation, and list steps. (See *Approach* section below)

4.) Define team and contributions. (See *Team Roles & Expertise* section below)

5.) Choose tool for next step – R, Gephi, Vos Viewer, or Pajek; or other approach

6.) Find help with these tools and develop budget and timeline.

7.) Confirm budget reallocation / plan with Children’s National (*Bob McCarter, PI*) and identify requirements/process for payments.

***Dataset description:***

We have a data set from the UCDC Natural History Study. This data set contacts 816 participants, each enrolled in this study and confirmed to have a diagnosis of one of 8 UCDC Disease Subtypes (OTC, ALD, ASD, ARG, CPS 1, …, and NAGS). Each participant record contains longitudinal data from a number of visits (at least 2 per year depending upon the age of the child) over the course of 8-10 years. At each visit, complete physical exam data is collected as free text (search terms) with SNOMED CT codes. Approximately 2,000 SNOMED CT codes were generated from free text findings (in cases where they were missing codes) and verified for accuracy (Marci Bowen, 2018). The final data set in Neo4J includes 5,219 unique SNOMED CT codes (xxx occurrences) for the 816 participants.

The dataset contains over xx variables (See [Data Dictionary](https://duke.box.com/s/0520l5mcwqbnl40qa2ykvyfu7m88ad33) for complete list) representing the following general areas: Current Eligibility, Enrollment, Diagnosis, Participation, Clinical Status.

***Database description:***

This data has been imported into a Neo4J graph database that contains versions of SNOMED CT and RxNorm. The information model is described here: [DUKE UCD NEO4J Graph Database Info Model 20190411.docx](https://duke.box.com/s/mca2snj8bo0dcy02rag3v38ldtpy7vzw) Essentially, most of the patient variables are included as attributes to the patient nodes, and the SNOMED and RxNorm codes are included with edges both to patient and visits. We

imported the UCDC data (structured elements such as … list (total # of records). And the SNOMED clinical finding and RxNorm medication codes applied to records at each visit. This is the information model and then we imported SNOMED CT and RxNorm into the graph database for …

**Table 1. Edge definitions (“relationships”) in our Neo4J Graph Database**

|  |  |  |
| --- | --- | --- |
| Edge Name | Nodes | Definitions |
| HAS\_VISIT | Participant to Visit nodes | These edges connect a participant to each of the visits associated with that participant. |
| HAS\_SCT | Visit to ObjectConcept nodes (Neuro/Behav, etc) | These edges connect visits to the set of SNOMED CT diagnostic codes assigned to the participant as a result of that visit. These are terminology nodes. |
| HAS\_RX | Visit to RXCUI/NDC nodes | These edges connect visits to the set of RXNORM or NDC pharmaceutical codes assigned to the participant during that visit. |
| P\_SCT | Participant to ObjectConcept nodes, (Neuro/Behav, etc) | These edges connect participants to the distinct set of SNOMED CT diagnostic codes which have ever been assigned to the participant. |
| P\_RX | Participant to RXCUI/NDC nodes | These edges connect participants to the distinct set of RXCUI and NDC codes for any pharmaceutical prescribed for the participant from any visit.  SNOMED CT-specific edges: ISA, HAS\_CAUSATIVE\_AGENT, … *Note: There are a large number of edge types defined for the SNOMED CT nodes which provide detailed semantics to SNOMED CT code definitions. See the information model document for additional details.* |

***Clinical/Research Questions:***

We are iterating this still. For general background, I am including the most recent ideas and clarifications here.

From December 11, 2019, Bob said… Our general areas of interest are:

1. Identify patterns in Psychosocial/psych diagnoses by UCD type (and by most severe UCD disorders or groups)
2. Identify risk factors that predict development of symptoms (including abnormalities in female heterozygotes w/ OTC (most prevalent UCD subtype)

And new idea from Rima on December 19, 2019 is here.

I summarize them to these 2 areas:

a.) Identify patterns in psych diag by UCD type. (add a time component - Sigfried suggests some tools on this…)

b.) determining time trends (over calendar time or over age) of drug utilization for seizure-related meds and diagnoses.

***Approach:***

1.) Build upon and articulate what we have accomplished so far and integrate it into another tool to show the value of what we are doing. (i.e., although the graph database makes this easier, we can still do all the same things using traditional tools. What can we use to show the added value of the semantic relationships here?? The assumption is that a visualization tool can allow users to see clusters or patterns of related codes that they would not see on a data report.)

2.) New – December 2019 – Rachel will work on the JBI paper just using Neo4J and CYPHER queries as the reportable product. She will work collectively with Jay and Prajwal to make this happen.

***To advance a true demo and tools for research team:***

3.) Prajwal will use either R, Gephi, Vos Viewer or Pajek (or other tool, TBD) to develop and show (“drive”) a “interactive query”/visualization in a specific area (described in next step 3). [Alternative is to use the Neo4J tools, Bloom, to visualize the graph data. If we have access to a qualified Neo4J consultant, they might have access to this tool. (That would meet our short- term goal for a demonstration of the semantics of this, but would be less reproducible for academic medical centers, given the high cost of enterprise Neo4J Bloom tool.)]

4.) Tom can lead / help with a R-Shiny app as he described to me on December 12, 2019: …

a.) Identify patterns in psych diag by UCD type. (add a time component - Sigfried suggests some tools on this…)

b.) determining time trends (over calendar time or over age) of drug utilization for seizure-related meds and diagnoses.

***To demonstrate move from data “exploration” (above) to a true analytic question (which could validate our approach), use these steps:***

4.) Divide data into severe and less severe groups. (Note that we already have Neo4J queries for this. Most severe group is called ‘proximal’ I think. See Jay’s Cypher queries at the end of this doc: [DUKE UCD NEO4J Graph Database Info Model 20190411.docx](https://duke.box.com/s/mca2snj8bo0dcy02rag3v38ldtpy7vzw) ) Query and export the data so that it can be visualized and cluster analysis using other tool (TBD).

5.) Using other tool (TBD), import the data from the Graph DB and present visualizations of the density of SNOMED codes that appear between the 2 groups. (The *challenge* is to maximize the use of the SNOMED semantic relationships (IS-A) to aggregate related but granular SNOMED concepts into clusters that are meaningful somehow….)

4.) Get the input of clinicians on this visualization. While on a call. They can look at the visualization and see what makes clinical or biochemical sense.

5.) Develop a clinical question of the pattern (“Are xx type of symptoms/features (a broad grouping) more prevalent in severe cases versus non severe cases?”) The focus will be on psych disorders and new clinical phenotypes for certain UCD subtypes.

6.) With the clinical question, we extract the data from Neo4J using a Cyper (SQL-ish) query and import that into a SAS data file for traditional testing. The key, and the exciting point of this, is that we will actually be querying paths in the gDB (using the semantic relationships in SNOMED CT) to extract these groups of symptoms.

7.) To demonstrate the value of the SNOMED CT relationships and the GDB query mechanism, we will report the number of different SCT concepts that are in the different groupings we test (showing the value of “semantic aggregation”).

8.) We will report the results of the comparison (all analyses) to experts for comment. Might be topic for a clinical article that UCDC experts might lead.

9.) Report this exercise as a case study (in Journal of Biomedical Informatics) and include some reflection and evaluation data. Present this as an approach for data mining in future rare disease studies.

10.) Create steps or a plan that would enable data analysts and/or clinical experts to query and manipulate the data themselves. (Stretch aim – area for future funding.)

Overall aim of the project and above demonstration and our write up: Graph DB and associated analyses can be used to identify clinical phenotypes associated with increased disease severity in a population of children. This requires manipulating the SNOMED CT codes to preserve the maximum level of semantics while making the data exploration feasible and useful.

***Team Roles & Expertise***

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Insitution/Role | Role on Project | Tasks |
| Rachel Richesson | Duke | Project Mgt / strategy |  |
| Bob McCarter | Childrens National | PI (epidemiologist/statistician/”domain expert) | Identify clinical and analytic questions; liaison with UCD clinician researchers; coordinate payment |
| Rima Izem | Childrens National | Co-Investigator; biostatistician | Identify clinical and analytic questions |
| Jay Pedersen | Univ of Nebraska | Neo4J expert, programmer | help with any CYPHER questions; define process for updating data in Neo4J |
| Scott Campbell | Univ of Nebraska | Co-Investigator; implementation and evaluation of graph DB and research database; SNOMED expert | Help with paper framing and writing |
| Prajwal Vijendra | Duke | Graduate assistant/data scientist | Perform Neo4J queries and plan/develop/execute demo |
| Tom Balmat | Duke – Office of Research Computing | Programmer; R experience and others | Help plan/develop/execute demo ; demonstrate R Shiny app |
| Eric Monson | Duke Libraries | Advisor | Advisor – data viz and tools |
| Jim Moody | Duke Sociology | Advisor (if needed) | Advisor – Cluster analysis; Pajek, social network analysis |
| Sigfried Gold | Univ of MD | Advisor and co-investigator | Data Viz,. Longitudinal analysis; terminology experience – esp visualizing the SNOMED relationships |
| Marcia Bowen | Duke alum | Co-Investigator and did tons of preliminary work; Advisor (if needed) | Share her methods; respond to questions as needed. …. |
|  |  |  |  |
|  |  |  |  |

NOTES:

Univ of Nebraska– Scott Campbell can advise on strategy and contribute to the JBI paper, which will be a follow-up from his paper here: <https://www.ncbi.nlm.nih.gov/pubmed/26305513> .

Sigfried Gold

Researcher, Consultant, Developer

COMBINE Fellow (NSF Research Traineeship)

Visualization for Health Data Analytics

Here's the longitudinal stuff:

<https://hcil.umd.edu/eventflow/>

<https://www.cs.umd.edu/hcil/lifeflow/>

I be interested in talking with Jay more about the neo4j queries they use to find related snomed codes. And did someone else on the call said something about following not just is-a relationships but all the other sorts of relationships between concepts?

***Budget:***

Bob will confirm that we have funds and a process to continue this work. That will impact the project plan and the timeline below.

***Proposed Timeline: (Need to update dates after Jan 16, 2020 call)***

|  |  |  |  |
| --- | --- | --- | --- |
| Step | Target Date | Person Responsible | Estimated # of hours |
| Finalize approach, strategy, and staff | January 15, 2020 | Bob and Rachel | N/A |
| Plan and start demo (iterative) | January 15, 2020 | ?? |  |
| Prep demo call with Bob and Rima (UCD *data* experts) | January 16, 2020 |  |  |
| Revise & iterate |  |  |  |
| Demo call with UCD *clinical* experts | March 1, 2020 |  |  |
| Revise & iterate tool | March 15, 2020 |  |  |
| Traditional analysis to confirm / augment findings | April 1, 2020 | Bob & Rima |  |
| Finalize JBI manuscript and submit | May 1, 2020 | Rachel (will lead; all will co-author) | N/A |

**APPENDIX - Notes and resources**

Duke Box Resource Folder:<https://duke.app.box.com/folder/52754915270>

SNOMED Resources (to inform the design of Neo4J queries that demonstrate semantic aggregation…)

SNOMED CT Browser - <https://browser.ihtsdotools.org/>

SNO-OWL – (video here: …..

RxNorm / and related tools to group RxNorm codes.

Documentation for RxNav: <https://rxnav.nlm.nih.gov/RxNavViews.html>

Example: get all RxNorm codes for meds that treat seizures/disorders

RxMix : <https://www.nlm.nih.gov/research/umls/user_education/quick_tours/RxMix/RxMix_Pre-Built_Workflows.html>

OUTPUT (12/19/2019):

| term\_type | RXCUI | name | SourceId | SourceName | Relation |
| --- | --- | --- | --- | --- | --- |
| IN | 167 | Acetazolamide | 167 | Acetazolamide | DIRECT |
| PIN | 91406 | Acetazolamide sodium | 91406 | Acetazolamide sodium | DIRECT |
| IN | 719 | Amobarbital | 719 | Amobarbital | DIRECT |
| IN | 719 | Amobarbital | 719 | Amobarbital | INDIRECT |
| PIN | 1314242 | Amobarbital Sodium | 1314242 | Amobarbital Sodium | DIRECT |
| PIN | 1314242 | Amobarbital Sodium | 1314242 | Amobarbital Sodium | INDIRECT |
| IN | 1325 | Barbital | 1325 | Barbital | DIRECT |
| PIN | 235404 | BARBITAL SODIUM | 235404 | BARBITAL SODIUM | DIRECT |
| IN | 1739745 | Brivaracetam | 1739745 | Brivaracetam | DIRECT |
| IN | 21241 | clobazam | 21241 | clobazam | DIRECT |
| IN | 2598 | Clonazepam | 2598 | Clonazepam | DIRECT |
| IN | 3322 | Diazepam | 3322 | Diazepam | INDIRECT |
| IN | 1482502 | eslicarbazepine | 1482502 | eslicarbazepine | DIRECT |
| PIN | 1482501 | eslicarbazepine acetate | 1482501 | eslicarbazepine acetate | DIRECT |
| IN | 4136 | Ethotoin | 4136 | Ethotoin | DIRECT |
| IN | 4177 | Etomidate | 4177 | Etomidate | INDIRECT |
| IN | 24812 | felbamate | 24812 | felbamate | DIRECT |
| IN | 72236 | fosphenytoin | 72236 | fosphenytoin | DIRECT |
| IN | 72236 | fosphenytoin | 72236 | fosphenytoin | INDIRECT |
| PIN | 82806 | Fosphenytoin sodium | 82806 | Fosphenytoin sodium | DIRECT |
| PIN | 82806 | Fosphenytoin sodium | 82806 | Fosphenytoin sodium | INDIRECT |
| IN | 28439 | lamotrigine | 28439 | lamotrigine | DIRECT |
| IN | 6470 | Lorazepam | 6470 | Lorazepam | INDIRECT |
| IN | 6585 | Magnesium Sulfate | 6585 | Magnesium Sulfate | DIRECT |
| PIN | 1311625 | MAGNESIUM SULFATE ANHYDROUS | 1311625 | MAGNESIUM SULFATE ANHYDROUS | DIRECT |
| PIN | 1311259 | MAGNESIUM SULFATE MONOHYDRATE | 1311259 | MAGNESIUM SULFATE MONOHYDRATE | DIRECT |
| IN | 1310578 | Medium chain triglycerides | 1310578 | Medium chain triglycerides | DIRECT |
| IN | 6960 | Midazolam | 6960 | Midazolam | INDIRECT |
| PIN | 203128 | Midazolam Hydrochloride | 203128 | Midazolam Hydrochloride | INDIRECT |
| PIN | 142441 | Midazolam Maleate | 142441 | Midazolam Maleate | INDIRECT |
| IN | 7440 | Nitrazepam | 7440 | Nitrazepam | DIRECT |
| IN | 7909 | Paraldehyde | 7909 | Paraldehyde | DIRECT |
| IN | 8004 | Pentobarbital | 8004 | Pentobarbital | INDIRECT |
| PIN | 203085 | Pentobarbital Sodium | 203085 | Pentobarbital Sodium | INDIRECT |
| IN | 1356552 | perampanel | 1356552 | perampanel | DIRECT |
| IN | 8134 | Phenobarbital | 8134 | Phenobarbital | INDIRECT |
| PIN | 82077 | Phenobarbital Sodium | 82077 | Phenobarbital Sodium | INDIRECT |
| IN | 8183 | Phenytoin | 8183 | Phenytoin | INDIRECT |
| PIN | 71227 | Phenytoin sodium | 71227 | Phenytoin sodium | INDIRECT |
| IN | 187832 | pregabalin | 187832 | pregabalin | DIRECT |
| IN | 8782 | Propofol | 8782 | Propofol | INDIRECT |
| PIN | 9004 | Pyridoxal Phosphate | 9004 | Pyridoxal Phosphate | DIRECT |
| IN | 684879 | pyridoxine | 684879 | pyridoxine | DIRECT |
| PIN | 203164 | Pyridoxine Hydrochloride | 203164 | Pyridoxine Hydrochloride | DIRECT |
| IN | 69036 | rufinamide | 69036 | rufinamide | DIRECT |
| --- | | | | | |
|  |  |  |  |  |  |

First time user? Check our [RxMix tutorial](http://rxnav.nlm.nih.gov/RxMixTutorial.html).