Leveraging the Formal Semantics of SNOMED CT for Research and Discovery in Rare Diseases Using a Graph Database Approach

Updated March 9, 2021

Target Journal: Journal of Biomedical Informatics, Research paper

Authors (names & order TBD): Rachel, Bob, Scott, Jay, Tom, Marci, Rima, Eric, Prajwal, Sandesh, Sigfried (?), Jim Moody (?), …. others….? Can also use acknowledgements section…

HIGHLIGHTS (3-5 bullets; required by the journal)

- Rare diseases have small numbers of cases and are generally understudied.
- The formal semantic relationships built-into SNOMED CT enable the exploration and analysis of clinical and research data to facilitate recognition of new and under-appreciated clinical patterns of disease.
- Methods and tools for using semantic relationships in analytics are missing.
- A tool is presented to integrate detailed research data with SNOMED CT relationships.
- There are many challenges to exploratory-type analyses with SNOMED CT and a one-size fits all tool is not realistic.
- Multi-disciplinary teams are required to leverage SNOMED CT semantics for data mining and knowledge discovery.

ABSTRACT

Background: Detailed clinical phenotypes are useful for understanding of the range of presentation and management of medical conditions, especially rare and emerging disorders. The formal semantic relationships built-into SNOMED CT enable the exploration and analysis of clinical and research data to facilitate recognition of new and under-appreciated clinical patterns of disease.

Objective: The objective of this project is to advance data exploration and knowledge discovery from a large dataset on children with Urea Cycle Disorders (UCD) using formal semantic relationships of SNOMED CT.

Methods: We imported data from a natural history study of UCD into a graph database along with the (n=#) semantic relationships of SNOMED CT. We developed an interactive data visualization tool that can graphically represent the prevalence and co-occurrence of multiple clinical findings (SNOMED CT concepts), and adjust the presentation for different levels of granularity in the SNOMED CT terminology. We iteratively refined this tool in a series of meetings with clinical and data experts to explore psychiatric and neurologic abnormalities in 8 different UCDs (i.e., diagnosis subtypes) and report our experience and lessons learned.

Results: Our work revealed specific challenges for the practical use of SNOMED CT in translational science, including selecting relevant SNOMED CT concepts and classes for display, managing active and retired concept relationships, selecting from multiple pathways of concept relationships built-into the SNOMED CT structure, iterating between data queries and display parameters, and applying inductive versus deductive reasoning paradigms.

Conclusion: Using semantic relationships in SNOMED CT in conjunction with a graph database approach facilitated the mining of rich natural history study data for knowledge discovery, but required a multi-disciplinary team of technical, data science, terminology and clinical experts. The insight derived from this case study has promising implications for improved recognition and care of rare conditions. From a data science perspective, this study underscores the importance of binding semantic-based terminologies to research and clinical data.

Commented [RR1]: Comments on title? I am thinking of keeping is short and cutting the "using graph database approach" but not sure if that phrase will pull in more potential readers…. Thoughts?

Also, we at about 9,300 words in the submission. There is no word limit in JBI but we should reduce as much as we can.

I will review and edit with an eye toward cutting and streamlining.

Please comment on content, ideas, and organization – do not worry about major editing on this draft.

Commented [RR2]: Scott and Jay – can you estimate the # of relationships imported in the DB? Alternatively, you could tell me which version and/or axes of SCT is in your gDB you shared for this.

1.0 INTRODUCTION

1.1 Detailed clinical phenotypes advance understanding and treatment for rare disease subtypes

In rare diseases, the number of cases is small by definition and the expression of diseases can vary substantially across patients.[1-3] Rare diseases can take years to diagnose and years to reveal their full effects, leaving patients with potentially related but unrecognized patterns of comorbidity that can seriously impact quality of live. Defining these clinical phenotypes (i.e., composites of observable characteristics or traits, such as comorbidities, morphology, development, biochemical or physiological properties, or behavior) is critical for the identification of *rare disease subtype*s and distinct etiology that may respond to different treatments.[4]

Commented [RR3]: Associated? Precise? Specific? Unique?

1.2 Articulation of Problem (that is biomedically or clinically motivated, per journal instructions)

Due to resource constraints, the data needed to characterize rare diseases are often not collected at all (i.e., research is not funded) and in clinical settings, detailed data are not typically collected in standardized or structured format. Clinical observations in research and clinical care are often reported as free text (e.g., in clinical notes or "other Physical Exam findings", where important information may be documented differently by different providers across differing disciplines). Even when controlled terminologies are used for coding observations, different observers across multiple sites may focus on different aspects of a complex presentation and document observations at different levels of detail (e.g., gait-specific-oddity vs. "abnormal gait"). *Semantic aggregation* can be used to integrate data that is conceptually related so that it can be processed and analyzed without content loss. Further, there are many ways the data instances in a data set can be 'conceptually related' and grouped for analysis, requiring an exploratory data-driven approach guided by a disease-specific knowledge of what is known and what needs to be known to be useful in the context of scientific and clinical problems.

Commented [RR4]: Should I cut this sentence? It distracts my transition to SNOMED intro in the next paragraph, but I want to introduce the exploratory and data-driven paradigm early….

SNOMED CT (The Systematized Nomenclature of Medicine - Clinical Terms) includes formal (explicitly asserted) semantic relationships between concepts that can be used to organize very specific data by common properties, enabling the semantic aggregation and description of data by concept groups at varying levels of detail. At present, there are no generalizable methods for using formal semantics in clinical data analytics, and that gap limits the use of existing data to discover clinical phenotypes for rare and emerging diseases. To address this gap, we designed a prototype data visualization interface to demonstrate the utility of formal semantic relationships in translational research. And value of semantic aggregation to find new patterns of disease expression that can trigger insight about mechanistic causes and potential cures.

We applied this approach to the challenge of defining new or under-recognized phenotypes in cases of Urea Cycle Disorder (UCD). Below we describe relevant background on UCD and SNOMED CT (including limitations of current analytics approaches.) After, we describe our experience using the formalized semantic relationships of SNOMED CT represented in a graph database environment for visualization and knowledge discovery in our dataset, and share our specific challenges and lessons learned. Finally, we suggest that the use of formal semantics is a nascent but vital area to support future work in data science, machine learning, and translational research.

1.3 Background on UCD, medical knowledge needs, and analytic challenge

The primary role of the urea cycle in the human body is to process bi-products of protein degradation, especially nitrogen through urea production and excretion, which avoids creation of toxic products including ammonia. Failures of the urea cycle involve mutations that lead to deficits in the production of any of 6 enzymes and 2 transporters that control the cycle resulting in one of 8 genetic subtypes. Failures lead to increased production of neurotoxic metabolites, especially ammonia, often leading to hyperammonemic crises, which are the hallmark UCD disorders. Severity (of the known) effects of UCD appear to depend upon the degree of deficit and where it occurs in the cycle (early (proximal) steps tend to be more severe than distal defects), but it is not clear that metabolic abnormalities or their clinical consequences are fully appreciated given the rarity and complexity of these disorders. The full range of presentation, severity, and etiology of the 8 different UCD genetic subtypes is largely unknown. A useful approach to answering these questions is to explore a large (ish) set of detailed clinical and molecular data to reveal detailed clinical phenotypes and use those to provide clues about where in the cycle the deficit is occurring ..

The Urea Cycle Disorders Consortium (UCDC, https://www.rarediseasesnetwork.org/cms/ucdc/), an international network of 16 academic centers, is funded jointly by NCATS and NICHD, to collect data and perform research to better understand UCDs and potential treatments. A longitudinal Natural History Study captures systematic data on patients with confirmed molecular diagnosis of one of 8 UCDs; data collected includes pathophysiology, morbidity, and mortality, as well as growth and development (including cognitive function), biochemical and nutritional status and quality of life.[5] The study protocol includes clinical assessments of neonatal onset cases every three months until 2 years of age and every 6 months thereafter. Late onset cases are evaluated every six months during childhood. Adult participants and those post-liver transplant are evaluated yearly. The study has generated more than 12 years of longitudinal data on over 800 patients, and represents the largest and most comprehensive data set on UCDs in the world. Using these data, the UCDC has addressed predefined research questions that have improved understanding of UCD risks and risk factors[4-7] and led to new treatments, including use of N-Carbamylglutamate (NCG) for NAGS [8]) and ongoing investigations of gene therapy for UCD[9], NCG in Short Term Hyperammonemia Treatment Trial[10,11], and nitrous oxide in ALD [12]). However, the UCDC has not yet made full use of the large and rich repository of coded clinical data, including physical exam findings and medical histories collected from thousands of clinical and research encounters (and coded by research staff in SNOMED CT and RxNorm (REF)), due in large part to limitations of traditional analytic tools and lack of guidance for semantic-based data analysis and exploration.

1.4 Relevant background on SNOMED CT

SNOMED CT is the world's largest clinical terminology, has approximately 340,000 unique clinical concepts, 800,000 synonyms, and over a million relationships and, and is used by more than 30 countries. A defining feature of concept-based terminologies, such as SNOMED CT, is that semantic relationships between concepts are formalized. Formal *synonymy* relationships assert that any number of terms (e.g., 'Pompe disease', 'GSD type 2', 'GSD-II') are semantically equivalent to a single concept (e.g. 'Glycogen storage disease due to acid maltase deficiency'). Other relationships are *hierarchical* (e.g., 'Pompe Disease' IS-A 'Autosomal recessive hereditary disorder', and IS-A 'Disorder of carbohydrate metabolism'). SNOMED CT also has *associative relationships* (e.g., 'Cleft palate' *Has-Occurrence* 'congenital', *Has-Morphology* 'Developmental failure of fusion'). The formal relationships of one concept are inherited by (i.e. can be applied to all) data instances associated with that concept. For example, a Pompe (diagnosis) in a patient record infers that the patient has an 'autosomal recessive hereditary disorder' which is 'congenital' because these semantics are formalized in SNOMED CT. When linked to instances of patient data in a dataset, the formal relationships in SNOMED CT can be used to group multiple concepts along some common property

Commented [RR5]: Sandesh and Bob -- Am I characterizing the state of the science correctly here? Should I include some references?

In any case, my goal is to establish need for exploratory analysis and definition of clinical phenotypes for each of the genetic subtypes….

Commented [RR6]: I do not have this quite right, I know, but this is a key transition. Need to establish – broadly – why we are looking at detailed phenotypes/clinical patterns…

Commented [RR7]: Rachel - Remove website in narrative and add as a citation.

Commented [RR8]: Rachel - Add these 2 conference references:

Patrick TB, Richesson R, Andrews JE, Folk LC. SNOMED CT coding variation and grouping for "other findings" in a longitudinal study on urea cycle disorders. *AMIA Annu Symp Proc*. 2008;2008:11-15. Published 2008 Nov 6.
Paper here:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656069/

Richesson, R., Syed, A., Guillette, H., Tuttle, M. S., & Krischer, J. (2007). A Web-based SNOMED CT Browser: Distributed and Real-time Use of SNOMED CT during the Clinical Research Process. Non-Serials, 631-635. https://search.informit.org/doi/10.3316/informit.783180454672166 (Original work published January 2007)
https://ebooks.iospress.nl/publication/11052

Commented [RR9]: Note - I need to update these numbers.

Scott –. I have never seen the number of relationships reported – can you help me estimate that? Alternatively, we can cut this and edit the sentence to fit the #'s we have… ;)

Commented [RR10]: Note: I cut this but it might be relevant background. Let me know if you think I should add it back in…

In the US, the National Library of Medicine (NLM) maintains a local extension and facilitates free use of SNOMED CT for public and commercial entities.

or dimension, and these groupings have potential to expose patterns (clusters?) in the dataset that were otherwise buried in more granular data. In practice, any given concept could have dozens of associated relationships through assertion and inference. However, many of these relationships are not clinically meaningful or useful for analysis. Methods are needed to filter out only those relationships that are relevant in order to recognize patterns (clusters?) and derive insights from the data that have clinical or scientific implication.

Commented [RR11]: Eric – Can you suggest better wording for this paragraph??

1.5 Approaches to Analytics with SNOMED CT

Despite the potential for SNOMED CT semantic relationships to support research and discovery, examples semantic-based analyses are actually sparse. There are several papers that leverage the formal semantics of SNOMED CT for auditing and error detection in the maintenance of the terminology (REFS), including comparisons of change or versions in SNOMED CT (ref – Yeshoa), however, there are few examples of using semantics for the analysis of (clinical or research) *data* specifically. Much of the literature about SNOMED CT in clinical environments relates to development of interfaces and approaches to automated coding of very detailed concepts, often for clinician-directed decision support or evaluation of its coverage for clinical text. Others have reported the use of SNOMED CT for organizing concepts in drop-down lists for EHR interfaces. Willet et al demonstrated the value and shareability of SNOMED CT-based diagnosis value sets to query EHR data from an academic medical center for population management and quality reporting. Several organizations have recently reported using the "intrinsic knowledge" of SNOMED relationships to define *intensional* value sets that describe certain clinical phenotypes of phenomena as a logical (or computed) collection of descendants (codes) of a broader concept of interest (as opposed to manually creating a curating a list of codes in an *extensional* value set), including the NLM VSAC, the UTSW (Willet), and the Observational Heath Data Sciences and Informatics (OHDSI) data collaboration. The use of intensional definitions can increase the completeness of relevant codes in a value set, and reduce the burden of maintain extensional value sets, which need to be updated as each terminology changes (e.g., new codes added or others removed.) (Cite Willet and NLM.)

Commented [RR12]: https://pubmed.ncbi.nlm.nih.gov/28566995/

https://pubmed.ncbi.nlm.nih.gov/29678093/

Commented [RR13]: https://pubmed.ncbi.nlm.nih.gov/30157499/

Willett DL, Kannan V, Chu L, et al. SNOMED CT Concept Hierarchies for Sharing Definitions of Clinical Conditions Using Electronic Health Record Data. *Appl Clin Inform*. 2018;9(3):667-682. doi:10.1055/s-0038-1668090

Commented [RR14]: https://www.nlm.nih.gov/pubs/techbull/nd18/nd18_vsac_intensional_definition_function.html

Commented [RR15]: https://ohdsi.github.io/TheBookOfOhdsi/StandardizedVocabularies.html

Current reports of SNOMED CT usage in clinical analytics generally use simple hierarchical (IS-A) relationships for querying populations of interest. Interestingly, it appears that all of these approaches use queries that are developed in a very top-down manner. In fact, even the guidance from SNOMED International (cite their guide to data analytics with SNOMED CT). What we do *not* see are applications where the data itself – or the SNOMED codes that appear in a large data set and at what frequency – drive the grouping to related codes. What we have not seen is a use of SNOMED encoded data to define the parents of interest – by traversing paths from a granular code that might appear in the data – either upward through parent codes to a broader super-type code, or more challenging – through a set of associative relationships to other concepts – each with one or more of its own parent hierarchies. This is likely due to the fact that a relational database is not an ideal fit for complex queries. In a RDB format, using SNOMED CT relationships to pull data requires complex queries that can be inefficient. These queries can be supported by a list of all relationships pre-computed in a *transitive closure* table. (REF) Despite that work around, to leverage the SNOMED relationships for querying data – or data mining, pattern recognition, etc – graph data base models are better designed to handle multiple relationships.

Commented [RR16]:
I need to revisit this paragraphs and streamline a little.

Consider adding idea of inductive reasoning and the challenges around that…
https://pubmed.ncbi.nlm.nih.gov/24434192/

Also, "data-driven analysis" is a different beast than question-driven analyses ….

1.6 Graph Databases in research and for SNOMED CT in particular

Graph databases (GDB) are schema-less (Not Only Structured Query Language, NoSQL) databases that use nodes to represent data entities and edges to represent relationships between them. gDBs can traverse liked relationships in a graph to discover new relationships. In that regard, graph-based queries can be

more efficient than SQL, which require multiple-join statements. Graph database models are well-suited for use with big data because the emphasis on relationships rather than structure allows for rapid addition of data and relationships without indexing. Others have demonstrated the use of graph database to detect patterns in a variety of fields, including banking and fraud detection and social network analysis. (REFS) The value of graph data bases has also been noted for data-driven clinical phenotyping [4] [5] and biology and genetics.

Dr. W. Scott Campbell and his team at University of Nebraska Health System presented the feasibility of using a graph database model for operational patient database (of academic medical center) built upon and around the semantic model of SNOMED CT and evaluated the model in terms of speed and accuracy of *clinical queries* of an operational system (>46,000 patient coded events and >2.1 million SNOMED codes).[6] They demonstrated that complex queries, including disjunction and negation, take longer implement and process in a relational database format. Their work, published in JBI, showed that complex queries (e.g., find "*all patients with a diagnosis of pneumonia caused by any influenza virus or human parainfluenza virus that subsequently consolidated into a particular region of the lung"*) can be quickly and accurately supported using a graph database version of SNOMED CT.

Our collaborative effort presented here includes the University of Nebraska investigators (SC and JP) and is an extension of the clinical use case (Campbell 2015) to facilitate knowledge discovery in an existing dataset. Our aims were to demonstrate the use of SNOMED CT, formal semantic relationships, and graph database model in a research context, and to evaluate its value in research discovery and analysis in a rare disease data set.

2.0 MATERIAL AND METHODS

2.1 Historical Approach and Formation of Research Team

To provide context for this paper (which presents our approach to semantic-based data-driven data analysis and our lessons learned), we will provide some history on the formation and evolution of the research team. The team and collaboration started with a strong data set (UCD study), a graph database version of SNOMED CT, and a number of experts motivated to "explore" the UCD data set and discover new patterns of disease or insights or opportunities to improve diagnosis or treatment of disease. Our team grew over several years and is represented by the authors and those listed in the acknowledgements. This work started with the goal to "roll-up" the data for meaningful descriptive statistics. (Annedotally, it was obvious at the start that there was variation in the level of detail in the SNOMED CT selected by researchers at distributed sites, and the large number of codes used only < 5 times was shown in a previous analysis. (REF)) The initial team included biostatistician (RM) and informaticist (RR), who had years of periodic discussions - mostly high level – about how to get value and leverage the SNOMED codes used for this important and detailed dataset on UCD. Another informatician (MB) joined the project and explored the quality of coding and the transfer into a GDB with the support of informaticist (WSC) and programmer (JP) from University of Nebraska. Over time, addition experts were added, including visualization expert (EM), data scientist/R programmer (TB), biostatistician (RI) and clinical expert (SN), and additional terminology/data vis expert (SG).

The team met sporadically frequently over the course of 2 years to identify strategy for comingling SNOMED coded data with known relationships, a plan for semantic analysis, and a scope/goal for the project. In a summer 2020 (get dates?), a clinical and research expert (SN) joined the team and brought

Commented [RR17]: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5389944/

Commented [RR18]: Rachel: think about adding a mention or references about

Ontologies.. including HPO were designed for this very purpose. (summarize Melissa's research and Monarch work.)

Commented [RR19]: Is it worth calling out the participation of Scott and Jay explicitly?

They did "inspire" this and help us….

Plus, their paper was published in JBI (same as where we are submitting). Which means, I think, that journal should be interested…

Commented [RR20]: This is an unusual section, but I want to give context around how long we did this, and how fast we moved (relatively) in past year-ish with Tom and Sandesh joining. (Their roles on the team are discussed later in the discussion.)

Do you all think this section "works" for this paper? Should we delete it, and just include in discussion?

Commented [RR21]: CITE: Marci Bowen's AMIA poster.

Commented [RR22]: Too much drama?

Commented [RR23]: … or has it been 3? Or 4? Seems like forever….. ;)

I will check my notes and clarify…

deep expertise in the biology, mechanics and presentation of UCD disease and an active program of research and working hypotheses on the biomechanics of UCD that would benefit from exploration of the extensive data in the UCD natural history study.

This paper describes the questions and information needs that this clinical and research expert brought to our collaboration and the UCD data set, and how we iteratively used and refined our semantic data visualization tools in response to his questions, collectively illustrating an unfolding story (described later in the results) of "knowledge generation". In the next section, we describe our data source, data preparation, our approach to formalizing and managing semantics, and development of our interactive data visualization tool.

2.2 Data Source

We use data from the UCD natural history study, which includes SNOMED-encoded clinical findings, to explore the feasibility of semantic-driven exploratory data analysis. The data was fully de-identified. The data set includes 753 patients with follow-up … range? The NH study protocol includes detailed capture of disease subtype (with genetic confirmation) at enrollment and detailed follow-up based on disease onset type (neonates … vs youth …) We explored a handful of structured variables in addition to all of the SNOMED CT codes reported (across all patients and visits) on the Behaviors sections of the Physical Exam and Medical History forms. The Physical Exam and Medical history forms included open-ended forms where research clinicians recorded free text for any finding or observation in the clinical visit and were prompted in addition to the free text to code in SNOMED-CT.

2.3. Research Data Preparation

A subset of data from the UCD NH study used for this exploration. Selected variables from a total of 753 participants were included in the dataset. Each participant had at least 5 years of follow-up data included in this data set. Variables imported were: UCD diagnosis, gender, age, all reported psychological and behavioral codes reported on the Physical Exam and Medical History forms (see appendix) at each visit, and date and type of (e.g. baseline, 3 month, 6 month) visit based upon enrollment in study. In addition, two other variables related to the severity of the UCD were included: 1.) presence or absence of a hyperammonemic event (with laboratory confirmation of low ammonia in blood), which is a disease-defining sequelae and serious event resulting in hospitalization or death, and 2.) presence or absence of symptoms accompanying that HA event.

As part of the original study, the research staff coded PE and MH findings in SNOMED CT, using a browser embedded in the online case report forms and tied to a terminology server. (REF) Each PE or MH finding was coded as a separate instance and the case report forms were designed to support the entry of up to 6 findings per body system. Both the search string and selected SNOMED CT code were captured in the dataset. In cases where the research coder entered free-text but failed to provide a code, a nurse informaticist (MB) with SNOMED CT training assigned SNOMED codes.

The analysis dataset includes 753 unique patients collectively representing 26,386 instances of 5,219 unique SNOMED CT codes. A clinician (MB) removed 85 of the 5,219 SNOMED codes due to lack of clinical relevance (e.g., navigational codes such as clinical finding) leaving 5,160 unique SNOMED CT concept codes.

The structured and open-ended variables for the data set are presented in Table 4 (or an Appendix).

Commented [RR24]: Bob – can you add any details here?

Commented [RR25]: Bob – can you go back and add relevant descriptors to the data you gave to Marci? I believe that we have 753 unique participants in our dataset but not sure what the original rationale for that was…. (completion of 8 study years of investigation, perhaps?)

Commented [IR26R25]: The number suggests that you have everyone enrolled alive up to the date of extraction

Commented [RR27R25]: Added note on March 2021: This paper is more about a Proof of concept for data visualizaton than the UCD-specific results. Soo…. Add what you can on the data but it can brief…

Commented [RR28]: https://pubmed.ncbi.nlm.nih.gov/17911793/

Commented [RR29]: Would a table of variable names and definitions be useful? Should we include in the appendix?

Commented [IR30R29]: Yes, I think that would be useful and we can separate the ones that are related to patient characteristics collected at baseline from diagnoses collected in follw-u

Commented [RR31R29]: Given length of paper, and our focus on innovation/POC/ discovery story (rather than reproducible research), I propose we put this in an appendix.

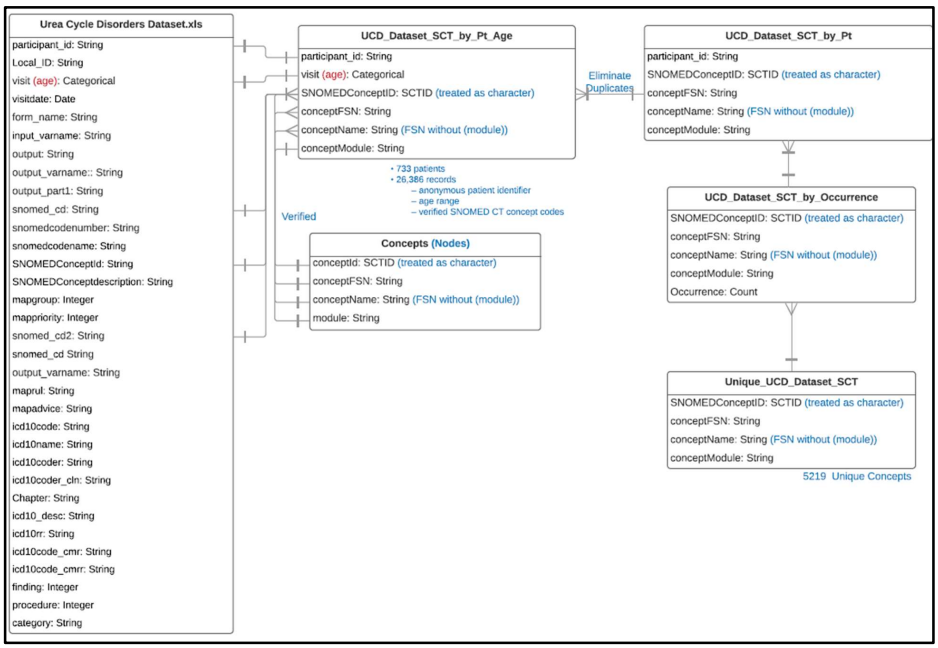
2.4 SNOMED CT Format and Preparation (Modeling Concepts and Relationships in Graph Database)

The UCD data described above was imported into Neo4J graph database containing concepts and relationships from SNOMED CT version RF2 using the following model, developed by Jay and Scott at University of Nebraska (REF).

Figure 1. Information Model

(links from patients → study visits → observations → SNOMED codes ; show cardinality, etc.)

(Marci has detailed documentation of her work (below); alternatively there might be something more of a figure from Jay's documentation. I am sure I have seen something in the past but cannot locate it now.)



Because there are so many concepts and relationships in SNOMED CT, we used the following process to select only the SNOMED concept codes that are a.) included in our UCD research dataset or b.) have an asserted relationship to a SNOMED CT concept in that is coded our UCD dataset.

To enhance visualization of our dataset, we further reduced the number of relationships by including only 20 specified SNOMED CT hierarchies (assessment scale, body structure, cell, cell structure, clinical drug, disorder, ethnic group, finding, medicinal product, medicinal product form, morphologic abnormality, observable entity, organism, procedure, racial group, regime/therapy, situation, specimen, staging scale, and tumor staging. The resulting file of relationships filtered by urea cycle disorders contained 16,490 records. The resulting file for the transitive closure filtered by urea cycle disorders contained 73,314 records.

Commented [RR32]: Jay – can you help me track down your information model for linking the SCT codes to the UCD visits? I this is our information model but it is pretty busy. I wonder if I should use something simpler… Or perhaps I can highlight the SNOMED-related attributes here…

Also, I need to take out the ICD10 stuff from this figure as it will just confuse…. (Marci and Bob and I used that in a very early phase of this project…)

2.5 Visualization Tool

Because of the size of the dataset and number of SNOMED relationships, the graphs had thousands of nodes and links and were essentially unviewable. Hence, a tool was needed to aggregate data instances into broader semantic groups based upon the relationships in SNOMED CT, which would reduce the number of nodes in a graph and allow clinicians / data viewers to view and understand the data and detect patterns / derive insights. To leverage the linked UCD and SNOMED data for visualization and interaction with clinical and data experts, a R-Shiny app was built to query and display data from the Neo4J graph database, using Neo4j's Cypher query language along with user supplied values from prompts presented by the Shiny app. The tool was designed and developed by a senior data scientist (TB) who interacted with data and terminology experts for initial requirements and clinical UCD experts for iterative development. Shiny provides functions for developing web pages for user interaction, while making available the entire R suite for back-end computation. RNeo4j provides an interface between R and a Neo4j database, so that queries resulting from user-specified inquiries are executed as a Cypher query against the graph database. visNetwork provides functions to generate nodes and connecting edges from queried, tabular data. The tool enabled users to select query specifications for variables (UCD subtype, proximal or distal subtype of UCD (based on the point in the Urea Cycle metabolism that the abnormality affects), sex, age, history of hyperammonemic event (an indicator for severe disease/flare), and reported medication (in RxNorm). In addition, users could query for SNOMED CT codes recorded as observations and findings on Physical Exam and Medical History forms using a dropdown list of SNOMED hierarchies that uses could navigate. Further, users can customize the parameters of the visualization, including the boldness of lines for the edges and colors and size of and placement of the notes. This Visualization tool has been previously described in (REF AMIA conference paper) and was used to support the iterative and collaborative and multi-disciplinary analyses we present in our case study here.

2.5 Approach to Collaboration and Dynamic Interactive Visualization of Data

Once the data was formatted and imported into the database, we invited a **clinical expert** (SN) in UCD that is an active physician and researcher engaged in identifying molecular abnormalities for specific UCD subtypes and identifying promising treatment approaches and possible drug targets to view the data and use the tool. The clinical expert is an investigator from the UCDC and contributor to the UCD natural history study and hence very familiar with the dataset. During 3 conference calls, we asked the clinical expert to view the data visualization and to talk-aloud regarding his thoughts and questions about the data being displayed, commenting on whether it was biologically plausible and asking what questions he might like to ask next. These calls were recorded with the consent of all meeting attendees.

The preparation for the sessions was that we had a large volume of SNOMED codes (limited to the behavioral and neurological codes) opened up the data using the R-Shiny data visualization tool. The tool was not ready for someone to drive (i.e. the interface was not fully developed), but our developer (TB) "drove" and we looked at data. We included the following variables: diagnosis (one of 8 UCD), whether or not they had HA events (an indicator of severity of disease), and all codes reported on the patient that were related to behavioral or motor symptoms. We selected this set for exploration because we wanted to leverage a clinical investigator that was doing a lot of basic science and animal studies around the different mechanism of disease subtypes. He had a desire to look at detailed clinical phenotypes and how he might use his clinical and biologic knowledge to identify or confirm (sort of ; validate perhaps) hypothesis around different biochemical mechanisms of action for different disease types based upon the clinical presentation of the cases (which are captured in our data…)

Commented [RR33]: Here is where we could mention all the tools we tried. The early days with Marci and Eric… And Jim Moody's advice… Would need to be one short sentence about existing products did not work for us…. (but we do not have to. Thoughts?)

Commented [RR34]: Marci, Eric, and Jim – could we/should we add a sentence here that mentions all your efforts in finding an appropriate tool…. Vos Viewer, Gephi, … I can't even remember them all. But, readers might want to know that even using existing tools was a conceptual and logistical challenge….

Curious if others think this is a useful piece of our story, or a tangent…

Commented [RR35]: Add this REF – we are presenting next week….

Using SNOMED CT Relationships for Data Exploration and Discovery in Rare Diseases - An Interactive Data Visualization Tool
T. Balmat, Duke University; R. Richesson, University of Michigan

March 24, 2021.
https://www.amia.org/summit2021/oral-presentations

Commented [RR36]: I need to mentioned that we looked an unique patients and all time points.

The data represent any instance of a SNOMED code for a particular patient at any point in their years of participation in the study…

Commented [IR37R36]: Yes, the "static" could be emphasized in the discussion. I would only mention in the dataset description that you are considering all diagnoses reported during follow-up regardless at which visit they were reported.

In addition, 2 UCD data analysts (RM and RI) actively participated in all calls, contributing their knowledge about the variables and codes collected in the study, as well as asking questions of the data. Other experts in data visualization (EM, SG) and SNOMED CT (RR) were on the calls to share relevant knowledge information. All participants on the calls asked questions about each visualization, in terms of what was included in a node (e.g., participants or visits; single instances of codes or groups of subsumed codes under a parent n concept), what the colors, boldness of edges, and different properties of the visualization mean. In addition, we had someone else (SG) able to review queries and counts on our calls to quickly verify numbers or answer questions such as "how many unique participants have codes for ….?" He used the same UCD dataset but linked to a set of transitive closure tables from OHDSI…

We had a total of 3 conversations, each lasting 1.5 hours all within the summer of 2020. These 3 collaborative sessions with the expert generated deep discussion, including questions about the data and requests for more data and interface components. The tool was iteratively enhanced to address these emerging questions and requests. In the next section, we present key observations and findings from these calls as our results below.

3.0 RESULTS

As described earlier, many aspects of pathogenesis and management of UCD subtypes are unknown. Armed with a rich and large dataset and a team of interested clinical and data experts, we applied SNOMED CT relationships to assist with data visualization and knowledge discovery. Our expert user asked questions and got us to a "story" that illustrates the use of formal semantics to assist in the development and confirmation of scientific hypothesis with potential implications for future treatment and management of UCS. In this section, we describe the results of our collaboration in terms of the design evolution of our tools, our challenges with SNOMED CT in this context, and our lessons learned.

3.1 Overview of visualization tool and user interaction – lessons learned

The R-Shiny tool created visualizations that looked like this:

[INSERT PICTURE that includes the query/filter configuration controls]

The SNOMED relationships were imported into the tool and displayed on a drop-down list. This list was visually overwhelming and cumbersome to manage. This is just prototype but for future, the usability of a SNOMED/semantic based visualization tool needs to enable users to visualize all the concepts and relationships that they might explore…. As with all data visualization, this is a challenge – giving users the information that they need (to help them to task or "think") without overwhelming them. (Share any other general usability issues we saw.)

It was customizable … but challenge was, we could not pre-determine the user interface and functionality, because it was dependent upon the expert and the questions, which evolve and involve different variables and relationships. So the variables and display parameters changed over the duration of our collaboration.

In addition to the interface design/query functionality, we had challenges (i.e., there are improvements to be made) around the display of results. Our work involved multiple conversations with a team of UCD researchers and data visualization experts to evaluate and adjust evolving query parameters and results displays. In using the tool, we discovered several important graph features that must be considered during

Commented [RR38]:
ZOOM RECORDINGS of meetings with Dr. Nagamani and examples to discuss:

July 10: https://duke.zoom.us/rec/play/7JQqf-ivpzM3HdeTsQSDUPV6W429LPqs0CFN-fIlmh61UXYCZgDzZrpHZ-Z8fZHZAuw5Xg5-_GwaEm1z?autoplay=true&startTime=1594407633000

June 12: https://duke.zoom.us/rec/share/w-d-K4Go2UlJTI3g0GfZflowRJTXT6a8g3AWqfAFy7lajxWLSmnGiMQq07PhaaU

June 5: https://duke.zoom.us/rec/share/59J_ApXKsWBJZZXB5UGHV6hwTtrYeaa81SEZrKBYnhuu6fxZQzK_vV6OT62M0kuk

Commented [RR39]: I need to go back to this section. Suggestions welcome.

Note that I do not want the focus of this paper to be on the Tool or the interface. I think it is more about the proof-of-concept for SNOMED or semantic-based data analysis, and also our challenges.

With that said, we have to show users something about our tool and pictures.

interpretation. We found that the size connected vertices must be considered when assessing significance of relative relationships. For instance, due to its heavier connecting edge, the relationship of proximal-UCD disorders to Attention Deficit Hyperactivity Disorder appeared more significant than that to Mood Disorder in our visualization. However, the heavier edge was explained by a greater number of observations in the first relationship than in the second, when in fact, the proportions of different UCD subtypes for each SNOMED concept were similar. Hence, there are unrealized opportunities to improve our interface to convey these data dynamics. To help identify strategies to improve these configuration and usability issues with the tool, it is worth examine some of the cognitive and conceptual challenges we encountered in this project.... (improve this transition to next section….)

In the next section, we present challenges in designing and using a tool on lots of data and for questions and purposes that were not known.

3.2 Challenges.

Our collaboration and experience with this work revealed several challenges that emerged from our multi-disciplinary data discussions and we present each below.

3.2.1 Challenge 1 – Lots of complex data makes visualizations difficult to design and present

The volume of data and relationships did create a feeling of overwhelm and was the subject of many conversations on how to get started visualizing our data using SNOMED CT relationships. This is understandable since the volume of data from SNOMED alone is huge, plus our UCD NH data. Over the course of our collaboration, we had to have multiple trials and errors to see which characteristics and level of granularity to use. There were many rounds of discussion about what data should be displayed and how, and then what was there. And how to interpret what is there. The data was just very complicated (plus we did not even address the longitudinal issue). There seemed to be no obvious or perfect way to display it…

There was really so much information in each visualization. The developer tried to build what the experts wanted but the initial direction of ("what differences in patterns of SNOMED codes exist between different UCD disorders or severity") was actually quite vague. Each call started with an iteration of the R-Shiny interactive app. All participants on the calls asked questions about each visualization, in terms of what was included in a node (e.g., participants or visits; single instances of codes or groups of subsumed codes under a parent n concept), what the colors, boldness of edges, and different properties of the visualization mean. (I could add some examples if needed, but the point is that there was a lot of questions just about what we were seeing and did it make sense.)

Interestingly, many of the questions about the data presentation were ultimately questions about the data. (Do we even have codes on xx in the data set?) We clarified our understanding of the data through a number of iterative questions.

Commented [RR40]: I can review recordings and find some specific examples if needed.

Eric mentioned once that data viz are often used to help understand data and identify data quality issues. So I think we should make this point explicit somehow.

3.2.2 Challenge 2 – Selecting from Multiple Pathways in SNOMED CT

Because SNOMED CT is multi-hierarchical (in contrast to strict classification systems such as ICD) a particular concept can have multiple parents – there are multiple possible 'pathways' in a graph that can be traversed. Decisions and appropriate interpretations of data visualization results need to be made regarding which SNOMED relationships to follow to build a path from a very specific concept to more broader concepts that are useful for data aggregation. An example is below.

Example – Tremor (from Tom – recording on 7/10/2020 ~ 7 minutes in)

Table 2: Query results for FSN description set 1

Description	SCT ID	Active	n-participants
[xother specified forms of tremor (finding)	NA	NA	0
benign familial tremor (finding)	192840004	0	1
dissociative tremor (disorder)	191713008	1	2
dysarthria (finding)	229685007	1	0
essential tremor (disorder)	632009	0	0
essential tremor (finding)	192839001	0	2
fine tremor (finding)	42800007	1	7
finger-finger test abnormal (finding)	250064001	1	0
finger-nose test abnormal (finding)	250063007	1	14
finger jerk finding (finding)	366462001	1	1
intention tremor (finding)	140873002	0	0
intermittent tremor (finding)	36637003	1	2
isolated facial tremor (finding)	230340001	1	1
isolated head tremor (finding)	230339003	1	1
muscle twitch (finding)	60238002	1	6
on examination - coarse tremor - flapping (finding)	163668002	1	1
on examination - fine tremor (finding)	163667007	1	5
on examination - intention tremor (finding)	163669005	1	5
on examination - tremor outstretched hands (finding)	417418002	1	12
repetitive rocking movements (finding)	30189004	1	2
resting tremor (finding)	25082004	1	12
tremor (finding)	308909003	1	0

Commented [RR41]: 2 things about this table. First, these are all the descendants of 'tremor' that appear in our data. Look at how many n-participants for each. (And imagine if we groups under the broader parent of tremor we would have more people with that observation.

Also, the presence of non-active concepts in our data…

Commented [IR42R41]: Yes, I like this

Need to identify the alternative paths and implied in this is the need to understand which path makes most sense in the context of the data and the particular question.

Commented [RR43]: THIS IS REALLY IMPORTANT…

We did not really have a question and that is the challenge with this. We have an ideas of what we want to explore but not sure exactly what we are trying to find. This is part of the paradigm shift for data driven exploration - but frustrating to developers.

They kept asking us "what do you want to find? Or "what is your question?" and the thing is we could not really express it. Wanted to see what we found in the data, but if you make the question too abstract, then it is hard to filter our the irrelevant SCT concepts.

On a similar note, we tried to make a comparison between 2 disease types (distal and proximal) or do 1x1 disease comparisons. Again, the problem is, if you have to limit the number of disease to compare in order to build a reasonable interface then you are limiting what you might discover …..

There are 2 different conceptual "paths" that can be generated from relationships (in this case "is – a" relationships) in SNOMED CT:

Clinical finding → finding of movement → involuntary movement → tremor → fine tremor

-and-

Clinical finding → Clinical history and observation findings → finding of movement → involuntary movement → tremor

As part of our process, we had to go back and forth between a number of tools to understand the SNOMED relationships alone and as they relate to the data in our study database. Our approach was a little awkward but effective. We would use the SNOMED International free online SNOMED browser to search terms and navigate up and down the trees. We had a person (Sigfried) that could tap the TC tables in OHDSI to give us counts, etc.) We found that these 3 tools were required to let us see understand the complex structure of SNOMED as well as to visualize our data co-mingled with SNOMED codes and relationships, and also see counts in tabular form of our data co-mingled with SNOMED relationships.

Also, an interesting finding / outcome of our collaboration was the realization that this choosing of concepts and paths is not trivial. Admittedly, the lead author and probably some others started this collaboration with the expectation that they data would drive the groupings – or in other words that relevant clusters of patients with semantically related codes would just be "revealed" from the data. But in reality, SNOMED does not have a standard semantic depth (e.g., the great grandparent code to a given code may be quite specific in one hierarch and less so than in another.) But also, just knowing which paths to follow or which broad concepts (examples…) are meaningful in the context of this. Also worth noting how vital clinical knowledge is to this. We had many iterations in our collaboration where an informaticist or data manager/analyst that are familiar with the research studies and UCD disorders tried to answer the questions but just could not. Only when we brought in our expert (SN) did we make some rapid progress on what to query and explore and how it might apply to working hypothesis and lines of thought about UCD mechanisms and phenotypic presentation.

Commented [RR44]: Really need to edit this better but wanted to capture it.

3.2.3 Challenge 3 – 6 Status (Active or Not)

Deprecated concepts were an issue. They were in our data and we saw some interesting patterns that we needed to research and explain. It is worth noting that the deprecated or retired concepts were in our visualization because they were in our data (presumably because those concepts were active at the time of the coding and became retired at some point in the study.) The question this posed for the team was whether or not we should delete these codes. Consensus was not to – they were coded because something was observed and recorded at a point in time during the study. We want to preserve that data. Most logical approach is to check the SNOMED change documentation to see what the chance was and why, and possibly identify codes for redirect. This is part of the U Nebraska model (ref) although we need guidance for additional use.

3.2.4 Challenge 4 - Determining the correct 'level' of semantic precision (i.e. which SNOMED concept) is meaningful in terms of addressing a translational research problem.

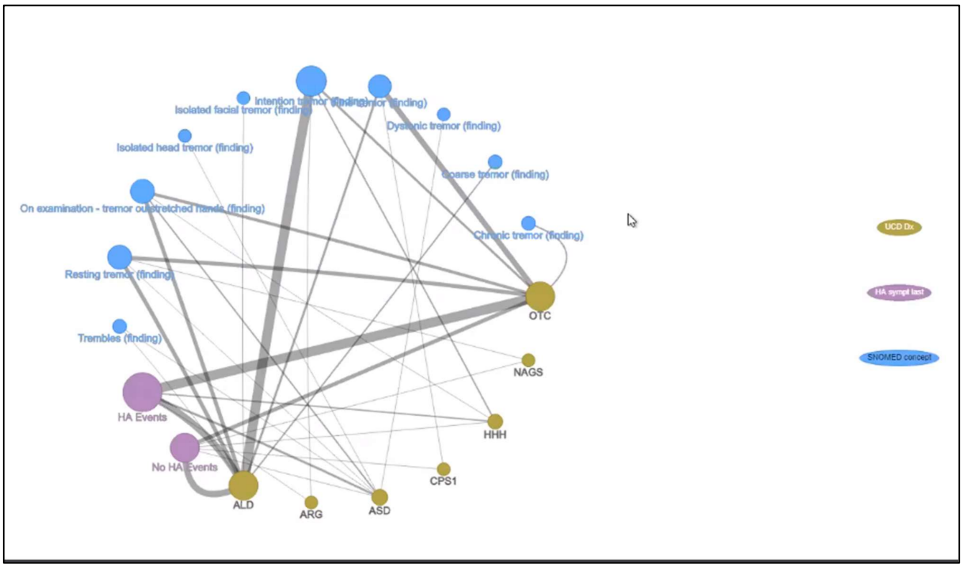
Commented [RR45]: This is our gold-mine of a challenge and really this section is the culmination of our efforts and should be the highlight of this paper. Curious if others agree or have ideas for how to frame this…

The clinical expert was interested in a line of inquiry around different types of disorders and what is in the database. He had an existing program of research comparing the mechanism and impact of different UCD subtype in animal models. Wanted to test and validate his hypotheses in context of the data .. Had general questions of interest but not a specific driving question.

We found several things in our data. First, we started to see differences, which were suspected by the expert, about the prevalence of different types of tremors in the datasets. Clinically, represented nuanced difference in the timing or etiology of the neurological symptom or comorbidity, which when combined with other data in the study about the severity and activity of disease, gives hints about …

We got to this figure which was very revealing…

Figure 1a: Visualization of tremor by all diseases



In the above diagram, we see the different types of UCD (yellow) and relationship to HA event (severity measure) and certain SNOMED codes (blue). The weights of the edges tell us that there are more participants with certain combinations of these different features but it is still hard to make sense of it. (This is the kind of thing our group spent many discussions on before Dr. Sandesh Nagamani joined the calls…)

We needed to drill down to find some significant/interesting patterns, and so we asked the expert to help us choose one or more diseases to compare. He was particularly interested in ASD vs ALD, which are caused by different biochemical defects in the urea cycle process. … Several questions/ data requests were suggested by the clinical expert and then translated into data queries by TB who was driving the tool. Eventually we came out with the visualization shown on the next figure.

Figure 1b: Visualization of types of tremor in ALD vs ASD, with HA events.

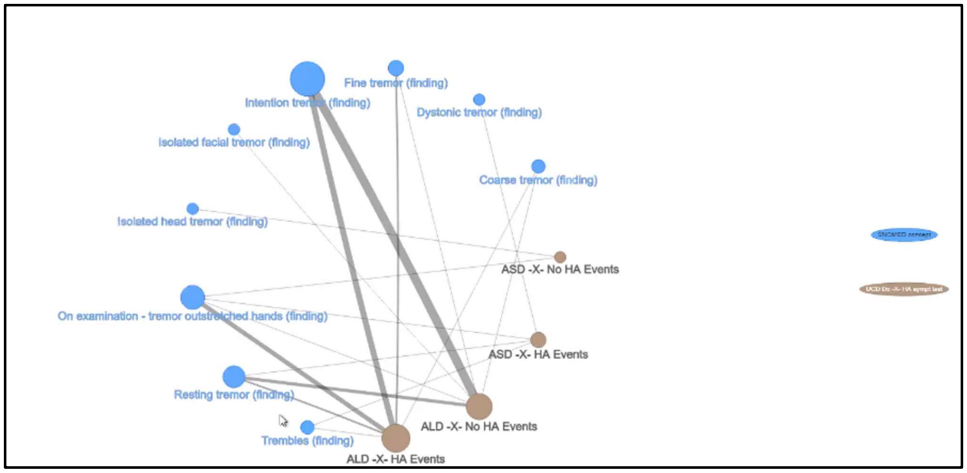


Figure 1 above shows all patients in the UCD data set with one of 2 different UCD disorders: ALD and ASD. Any patient with any UCD can experience a hyperammonemic event (a very serious set of varying clinical presentation accompanied with elevated levels of ammonia, severe enough to require hospitalization) , which is a hallmark of the disease. We created new nodes for each disease with and without the events, and wanted to compare the different types of tremors in the groups of patients with combinations of disease and events. Each node represents the number of *unique* patients in the dataset with the combination of events. On can hover over the node to see the # of patients included in each node. The size of the nodes are relative to the number of patients included so that the larger groups can jump out. The colors of the nodes were selected to distinguish the types of data included – blue nodes are all SNOMED codes, again the bigger node is more participants in the data set. The edges represent the #s of patients with both connecting nodes. Thicker lines mean more people.

Our collaborative discussion and iterative queries and visualizations with the data lead to an insight that complemented a line if inquiry by our expert. We present that insight as translational science "story" that illustrates the value and scientific impact of semantic-base data exploration to the understanding of rare disease. (See Box 1)

Box 1. The "story" is this:

In general, all UCD have some issue with the UCD which results in a failure to metabolize urea resulting in its accumulation of a toxic metabolite, ammonia, in the body. (The urea cycle is a cycle of biochemical reactions that produces urea (NH2)2CO from ammonia (NH3). The urea cycle converts highly toxic ammonia to urea for excretion.) Some of the neurological phenotypes seen in UCD is because of ammonia in the body. Ammonia is a general neurotoxin and responsible for many different neurological symptoms. Some specific UCD disorders (ALD, ASD, ARG) are expressed in multiple tissues *outside* liver. Even though they do not have toxic anemia (hyperammonia). The fact that ALD have no HA events but have resting and intentional tremors… this is a great visualization. Help us decide which of these neurological phenotypes are due to hyperammonemia (as a result of UCD and impacting the brain) versus a loss of other enzymatic function due to the specific disease/failure of the urea cycle metabolism. In other words, this visualization

Commented [RR46]: Should we add the hover numbers to the diagram, or add a small table to show how many are in different groups. The numbers are small, but in the end they show that we are digging deep into this data and making the most of an existing dataset.

Would require Tom to go in and generate new screenshot.

Commented [RR47]: Note – I need to present this very briefly here and also in a way that does not compete with Sandesh's other research and publication plans…

I will look more closely at the recordings and edit this section.

This was discussed on July 10 call around the 25-28 minute section.

helps us understand that some patients are showing neurological symptoms (e.g., tremor) not because of accumulated ammonia neurotoxin (as a sequalae of UCD) but because of loss of specific enzymes (due to the UCD defect) within the neurons. Specifically, the visualization shows that patients with ALD with no hyperammonemic events have resting tremors, which is interesting and unexpected. Also, if you look at resting tremors, there seem to be more people without HA events than with…. Confirms a hypothesis/line of investigation that his research team is currently exploring, but also gives us insight into the mechanics of disease for different UCD subtypes, and potentially in the future, targets for intervention.

"If I were to look at this as researcher, I would want to know if some of the resting tremors that occur because of loss of ASL with in the neurons, which may be unrelated to the hyperammonemia, and such things that we have done in the past have led to major discoveries for role of these proteins or what they make in the cell outside of their conventional roles in metabolizing." (27.09 minutes into the July 10 recording)

This visualization evolved from multiple iterations of discussion and questions driven by the clinical expert. It required not only an understanding of the SNOMED codes and subtle distinctions between different neurological symptoms (e.g. what part of the brain is responsible) and a deep understanding of the known (and unknown but suspected mechanism of action of the disease in the body), as well as familiarity with the disease population as a whole, the study scope, and the data in the data set.

3.2.5 Challenge 5 – Exploratory Questions by Definition are difficult to address

The story in Box 1 evolved over time and discussion. It was not immediate at the time we assembled our data and designed our query tool. Once the R-Shiny app was built, it became easy enough to ask investigators – what do you want to see? Find? Discover? And the process happened one step at a time, over several multi-hour calls. In retrospect and reflecting now, we believe that this was perhaps the primary challenge. The data is there. The discoveries should be waiting to be found but we have to filter through the noise to find them…. But how? Easier said than done.

Add idea of inductive reasoning and the challenges around that…
https://pubmed.ncbi.nlm.nih.gov/24434192/

We found that "data-driven analysis" is a different beast than question-driven analyses …. TOTAL PARADIGM SHIFT …

4.0 DISCUSSION

Limitations

There are several limitations to our approach and dataset. As outlined in the paper, the original goal was to combing our UCD NH dataset with SNOMED CT for discovery, but due to the challenges we noted above, the process took much longer, spread out over several years. The lead author did keep notes and recorded many meetings, but did not collect data specifically about our process or activities nor did we conduct a systematic exploration … So our lessons learned come from notes and recollection and may reflect recall bias.

Commented [RR48]: I need to trim this section on limitations. It is still too long and needs editing…

Please comment on the organization of paragraphs for the rest of the discussion, and add/remove ideas as appropriate.

Further, the data we use for this demonstration are from a research sample, and although it is arguably the largest sample in the world on UCD, it is not necessarily reflective of all patients with UCD. Everyone in our data set has a confirmed molecular diagnosis. Our data does not include: persons with undiagnosed UCD, a clinical diagnosis and no molecular confirmation of UCD, those who died, or persons with barriers to medical care or research participation. Our data includes only children. Also, this is just a subset of the extensive data collected in the study. The UCD Natural History study includes 8 (?) different forms and xx (>1000?)) items. Most of these are structured data. On multiple forms, many of which were structured to include known symptoms. (see appx for PE and MH forms.) Because of the presence of structured data, it is very possible that key data were entered as structured data and might not collected as free text or entered as SNOMED CT codes. (For example, all of these kids had UCD and that was recorded clearly on eligibility and baseline enrollment forms, yet many cases we see UCD diagnoses entered in as a SNOMED code under the MH or PE form. Our data is from a prospective observational study with many different sites. The data were collected prospectively and with a defined Manual of Procedures, but there is always variation (for both research and HER documentation) on who notices something (e.g., wears glasses, seems happy) and then documents it, and then at what level of detail (e.g. vision issues. The SNOMED coding was not done by experts and not validated. There was some training included in study protocol training but was minimal. (REF MEDInfo) The different levels of data documentation specificity and coding were actually expected and drove the choice for SNOMED CT as a reference terminology that could handle this type of variation using semantic aggregation. Finally, we did not explore any longitudinal or time component to this. That aspect undoubtedly adds more complexity to the data but will be important in the future to understand these relationships and how the disease progresses or presents over time and relevant to other clinical factors (collected in study as other variables.) Despite these limitations, our dataset presents a small set of clinical (– ish) data tied to validated diagnosis is ideal to demonstrate this proof-of-concept for semantic-based research analysis that can be used on other research data sets, or EHR data or clinical notes.

Findings – Reflection on Process and Approach

This collaboration began with the idea that we would find some "lightning bolt" sort of discovery with using SNOMED CT relationships that that did not materialize. But this is still terrific. We believe, at this stage of infancy of working with SNOMED data that it is nonetheless remarkable and valuable to use this unprecedented volume of data on this set of rare disorders to seed new knowledge based on finding evidence to support or refute concepts that should unfold based on knowledgeable inference. We used the basic and clinical science knowledge of our expert to postulate certain clinical patterns based on biochemical and pathobiological differences across these 8 disorders. However, the Urea Cycle research project was not designed to provide the kind of detailed information needed to explore specific features arising from steps in a biologic process. Instead, we have manifestations of pathology that would provide clues to help check the consistency of evidence against thought-based relationships. A turning point and an essential factor for the success of this was the addition of a clinical expert with DEEP domain knowledge – not just knowledge of the disease or familiarity with the UCD NH dataset, but someone with deep knowledge on multiple aspects of the disease (molecular and physical) and was motivated to discover new associations and patterns that are potentially clinically significant. Within a few short months we generated a visualization that was included in a research paper submitted to a clinical research journal. (REF submission)

Commented [RR49]: Move to results

Our experience highlights the challenges in using the extensive set of semantic relationships (knowledge) embedded in SNOMED CT. The formation of our group and extensive searching revealed lack of off-the-shelf tools for this kind of analysis. We developed an R-Shiny app that can (1) graphically represent the prevalence of multiple patient characteristics (e.g., disease information and reported SNOMED and Rxnorm codes), (2) graphically represent co-occurrence of multiple characteristics stratified by subgroup, and (3) adjust the result to different levels of granularity in the SNOMED ontology. This tool was valuable for our demonstration, but needs additional development to be ready for widespread use. As mentioned, we have issues with the visualizations (weights of lines and proportions in data vs counts). We are only graphically representing counts without showing confidence intervals in these counts. Also the customization issue is an ongoing challenge, because the optimal configuration of the tool depends upon the variables available (data set) and the nature of inquiry, leading us to believe that a generic "out of the box" tool for semantic-based data analysis for rare diseases data might not be feasible.

Two factors made us move fast – 1 the ability for an interactive tool. TO reduce the time to make and get queries. And 2, the involvement of a deep clinical expert. Even at the start of this with a data set, SNOMED knowledge, a (sort of experts) we could not make traction. The challenge with this project was that in supporting discovery it is difficult to design the tools that will find that. We kept asking investigators at different points in this to tell us what they were looking for .. but they really wanted to know "which important patterns are in the data?" but important counted on 2 things – first the number of instances but secondly on the clinical significance, plausibility or potential for impact and this is determined by investigators. In this sense, the addition of a deep expert was our turning point ..

Findings & Recommendations – SNOMED and Semantic-based data exploration

Our experience definitely illustrated benefits of SNOMED CT. Many examples of single instances of detailed codes that when combined (semantic aggregation) led to larger groups of patients with similar features. By re-organizing a sample of the UCD coded clinical encounter data into a *graph database* (graphDB) [13] that focuses on relationships rather than structure, we have been able to apply these methods to identify new UCD morbidity that traditional methods did not reveal. Our example of verifying known (or suspected) patterns of data may seem trivial but this is actually quite novel. We showed – through our "story" of challenge 4, that the use of this kind of deep data (an existing and rich natural history of disease dataset) could be used as a critical tool in the iterative development and validation of hypotheses related to the mechanism of disease.

We encountered 5 major types of challenges, each of which begs recommendations for the future… (Note: I will take the contents from the table below and put into narrative….)

Challenge	**Recommendation – Future Needs**
1 – Lots of complex data makes visualizations difficult to design and present	Recognize the complexity of data and reduce as much as possible to stay relevant to a question or goal…. Other ideas?
2 – Selecting from Multiple Pathways in SNOMED CT	Better interfaces that can allow all 3 functions – data visualization plus access to counts plus information about the relationships and hierarchies of SNOMED are important for the iterative development of data visualization tools.

	Ultimately, it would be nice if an viz interface could support the experts as they move around and interact in the data. (would allow more spontaneous idea generation and iteration …) but this is a challenge. There are complexities with the data itself, the SNOMED terminology, the interaction between them AND the biological and clinical domain.
3 – Many codes at Different Levels of Precision and Active or Not	a.) One or more clinical domain experts plus experts familiar with SNOMED CT are required to identify anomalies in data that might be due to SNOMED concept status or other relationships/features of the SNOMED concepts. b.) For any time of longitudinal data collection using SNOMED to code at the source, as we did for the RDCRN study, there will possibly be a need to address codes that are deprecated or have been moved around (representing new or changing relationships.) We are not sure that SNOMED distributes a computable version of this, but the graph database can handle multiple versions of SNOMED CT. Guidance in this area will be helpful for future investigations.
4 - Determining the correct 'level' of semantic precision (i.e. which SNOMED concept) is meaningful in terms of addressing a translational research problem	Ideas? Guidelines for starting? We did have to keep simplifying… 1x1 comparisons are much simpler.. Combining multiple variables into a node was a strategy…
5 – Exploratory Questions by Definition are difficult to address	**Can we give guidance on how to formulate the question? Or the logical thinking about it? How to form teams? (multi-expert teams)**

The biggest lesson learned … We represent a truly translational science perspective and highlight the need for CLINCIAL expertise and multi-disciplinary work… approaches to this problem that is human and technical. The use of semantic-based data exploration in rare and understudied disorders will likely requie a team science approach. … data experts plus clinical.. plus research ….

Implications for Research and Discovery in Rare Diseases

Our collaboration and insight (i.e. our "story") provides an example of using existing data for purposes that is was not intended. Further, this work demonstrates an approach using formalized and existing knowledge to mine and leverage and re-use existing datasets, and this has huge implications for understanding rare

diseases and for precision medicine. We see applications for this tool in translational science in all diseases, and a particular promise for rare diseases which by definition have far fewer data resources available. In research in general … Because rare diseases have fewer patients and often less researched, the challenge is to be able to make effective use of large-scale data based on recorded clinical observations to recognize patterns that define phenotypes. In rare diseases, especially, natural history studies are the first step in the research and drug development process. These studies come at great expense. They should be maximized ~ Moral imperative to look at datasets already collected. … While semantic-based data exploration is essential (perhaps) for rare diseases, these methods are not limited to small sample sizes and can apply to any condition. In fact, may have more impact if make a discovery on understudied conditions that affect many, many people.

This work also has implications for free text data analysis. SNOMED CT is most comprehensive clinical terminology, and is a logical choice as a reference terminology to codify concepts extracted from free text clinical notes. As healthcare organizations have incentive reduce health costs and improve patient outcomes, they will be increasingly motivated to look at clinical data (including free text data) and methods – including semantic-based – analytic methods. A recent JBI paper suggests a novel and game changing method for mapping text to SNOMED, including new codes. This is a breakthrough from previous methods which used string matching. The authors also provide excellent descriptive background on SCT and justification for its use. That we recommend to interested readers. The work builds the case for methods of advanced (semantic driven) analysis and that is what this work is about.

Commented [RR50]: Add these new references.

Future directions and next steps

The semantics and structure of SNOMED CT have tremendous potential (more embedded knowledge and detail than other coding systems) but it brings great complexity. Our work has showed us that there are no semantic rungs on a ladder of detail – eg.. concepts at level of depth 2 might be very specific or broad. So, cannot just go up certain levels. A key requirement / motivation for this work is to develop approach to manipulating the SNOMED CT codes to preserve the maximum level of semantics while making the data exploration feasible and useful. However, that 'maximum level of semantics' is context specific and database specific. So we need tools and approaches that can be replicated. … Further … The IS-A hierarchies have not only different levels of depth and detail but there are also challenges related to the multi-axial aspects of SNOMED CT. A given concept can have multiple paths… So in a graph DB scenario, it is hard to know which path to follow. Of course the latest challenge lies ahead and that is the lateral or associative relationships.

Moving forward, we hope to replicate our method in different datasets (e.g. EHRs and a clinical data for UCD) (~exploration goals/areas of inquiry) and reproduce and test our approach on new "questions" for other rare conditions. We would like to improve interface and somehow make easier for clinical experts to "Drive" the data. Ultimately, we envision that investigators and data scientists that use semantic-based data analysis feedback experience and results to SNOMED CT to improve the terminology and give guidance for the use of SNOMED CT in clinical analytics, research, and discovery. SNOMED International and others can further use / disseminate their experience and reach out to machine learning and translational researchers about this way to incorporate existing knowledge into data driven applications.

5.0 CONCLUSION

The formal relationships of SNOMED CT and other terminologies can be used to support analysis of clinical data to answer clinical questions and for discovering new knowledge (e.g., relationships / disease patterns).

The use of graph database technology and its inherent focus on semantic relationships represents an analytic paradigm shift that brings an opportunity to use modern data science and data visualization tools, making it possible to address questions that are extremely difficult or impossible using traditional relational databases.

But this is very difficult – often requires building the plane while flying it… trial and errors. Cannot pre-determine the patterns … Must bring in a data-driven aspects to know how level of semantic precision …

In our experience, because there are so many possible relationships in SNOMED CT, this cannot be done without a clinical expert that is focused on the …

The use of standard terminologies and open-source tools facilitates the scalability of analytic methods and applications, and can lead to improvements in diagnosis and personalized treatment of rare disorders.

Acknowledgements

This work was supported by the CTSA supplement (award ###) and the Duke University Office of Research Computing. The UCD Consortium (U54HD061221) is a part of the NIH Rare Disease Clinical Research Network, supported through collaboration between the Office of Rare Diseases Research, the National Center for Advancing Translational Science (NCATS), and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). We are grateful to the following for their significant contributions to this work: Bob McCarter, Rima Izem, W. Scott Campbell, Jay Petersen, Marcia Bowen, Eric Monson, Prajwal Viendra, and Sandesh C S Nagamani, MD. Jim Moody? Sigfried Gold?

Commented [RR51]: Update this. Remove confirmed co-authors from this section and be sure to include any important collaborators or supporters…

Resource Sharing

The code for the graph database is available at: List various Github sites…

REFERENCES

1. Batshaw, M. L., **Tuchman, M.,** Summar, M., Seminara, J. A longitudinal study of urea cycle disorders.

Commented [RR52]: Working on these.. I know refs are a mess but I will update and re-format.

In my move to Michigan I have lost my Endnote library and software… And merging content from multiple papers here..

I am switching to Zotero for my refs and will sort them out. Stay tuned…

In SI:Newborn Screening, Molecular Genetics and Metabolism. 2014;113(1-2):127-130.

2. Merritt, II, J. L., Seminara, J., **Tuchman, M.,** Krivitzky, L., et. al. Establishing a consortium for the study of rare diseases: The Urea Cycle Disorders Consortium. Molecular Genetics and Metabolism, 2010(100):S97-S105.
3. Daoud, Y., Seminara, J., Mew Ah, N., **Tuchman, M.,** Nagamani, S, Le Mons, C., McCandless, S., Gropman, A. The Urea Cycle Disorders Consortium: Highlights. *Molecular Genetics and Metabolism*. 2016 (117; 3):253.
4. Ah Mew, N., Krivitzky, L., McCarter, R., Batshaw, M., **Tuchman, M***.* Clinical Outcomes of Neonatal Onset Proximal versus Distal Urea Cycle Disorders Do Not Differ. *Journal of Pediatrics,* 2013 (162-2): 324-329.
5. Lichter-Konecki, U., **McCarter, R**. Glutamine as biomarker for quality of metabolic control and guidance for dietary management in patients with urea cycle disorders. *Molecular Genetics and Metabolism*. 2012(105; 3):333-334.
6. **Waisbren, S.**E., Gropman, A.L., Batshaw, M. Improving long term outcomes in urea cycle disorders-report from the Urea Cycle Disorders Consortium. *Journal of Inherited Metabolic Disease.* 2016(39-4):573-584.
7. Ah Mew, N.A., **McCarter, R.,** Lichter-Konecki, U., Tuchman, M., Daikhin, Y., Nissim, I., Yudkoff, M. Augmenting Ureagenesis in Patients with Partial Carbamyl Phosphate Synthetase 1 Deficiency with N-carbamyl-L-glutamate. *J*ournal of Pediatrics; 2014; (165- 2):401-403.
8. Caldovic, L., Ah Mew, N., Shi, D., Morizono, H., Yudkoff, M., **Tuchman, M**. N-acetylglutamate synthase: structure, function and defects. Molecular Genetics and Metabolism. 2010 (100) Supplement:S13-S19.
9. Wang, L., Bell, P., Morizono, H., He, Z., Pumbo, E., Yu, H., White, J., **Batshaw, M.**L., Wilson, J.M. AAV gene therapy corrects OTC deficiency and prevents liver fibrosis in aged OTC-knock out heterozygous mice.*Molecular Genetics & Metabolism.* 2017(120 – 4):299-305.
10. Children's Research Institute; Boston Children's Hospital; University Hospitals Cleveland Medical Center; University of California, Los Angeles; Children's Hospital of Philadelphia; Stanford University; Icahn School of Medicine at Mount Sinai; University of Pittsburgh; Children's Hospital Colorado. Short-Term Outcome of N-Carbamylglutamate in the Treatment of Acute Hyperammonemia; Trials.gov. **Mendel Tuchman***,* 2017.
11. Mew, N.A., McCarter, R., **Tuchman, M.,** Daikhin, Y., Nissim, I., Yudkoff, M. N-carbamylglutamate Augments Ureagenesis and Reduces Ammonia and Glutamine in Propionic Acidemia, Pediatrics, 2010. 126:1.E208-E214.
12. **Nagamani**, S. C.**S**.; Campeau, P. M., Shchelochkov, O.A., Premkumar, M.H., Guse, K. et al. Nitric-Oxide Supplementation for Treatment of Long-Term Complications in Argininosuccinic Aciduria. American Journal of Human Genetics. 2012; 90(5):836-846.
13. Robinson, I., Webber, J., Eifrem. E. Graph Databases, O'Reilly, 2nd ed, 2015.

1. Muenzer J. Overview of the mucopolysaccharidoses. Rheumatology (Oxford) 2011;**50 Suppl 5**:v4-12 doi: 10.1093/rheumatology/ker394[published Online First: Epub Date]|.

2. Thomas JA, Beck M, Clarke JT, Cox GF. Childhood onset of Scheie syndrome, the attenuated form of mucopolysaccharidosis I. J Inherit Metab Dis 2010;**33**(4):421-7 doi: 10.1007/s10545-010-9113-7[published Online First: Epub Date]|.
3. Bruni S, Lavery C, Broomfield A. The diagnostic journey of patients with mucopolysaccharidosis I: A real-world survey of patient and physician experiences. Mol Genet Metab Rep 2016;**8**:67-73 doi: 10.1016/j.ymgmr.2016.07.006[published Online First: Epub Date]|.
4. Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. N Engl J Med 2018;**379**(15):1452-62 doi: 10.1056/NEJMra1615014[published Online First: Epub Date]|.
5. Seminara J, Tuchman M, Krivitzky L, et al. Establishing a consortium for the study of rare diseases: The Urea Cycle Disorders Consortium. Mol Genet Metab 2010;**100 Suppl 1**:S97-105 doi: S1096-7192(10)00046-6 [pii]

10.1016/j.ymgme.2010.01.014[published Online First: Epub Date]|.
6. Campbell WS, Pedersen J, McClay JC, Rao P, Bastola D, Campbell JR. An alternative database approach for management of SNOMED CT and improved patient data queries. J Biomed Inform 2015;**57**:350-7 doi: 10.1016/j.jbi.2015.08.016[published Online First: Epub Date]|.
7. ONC. Interoperability Standards Advisory. Secondary Interoperability Standards Advisory 2019. https://www.healthit.gov/isa/.