Using SNOMED CT Relationships for Data Exploration and Discovery in Rare Diseases - An Interactive Data Visualization Tool

Tom Balmat[1] and Rachel L. Richesson, PhD[2]
[1]Duke University, Durham, NC; [2]University of Michigan, Ann Arbor, MI

Introduction

Electronic clinical and research data coded with standardized terminology can accelerate the discovery of detailed clinical phenotypes, which are vital to understanding the pathology and management of rare and emerging disorders, particularly multi-symptomatic disorders with varying severity and patterns of disease. The formal semantic relationships in SNOMED CT can support the exploration and analysis of data to recognize new clinical phenotypes, but tools for using these relationships in clinical analytics or research are lacking. The objective of this presentation is to demonstrate a data visualization tool that leverages the formal semantics of SNOMED CT to advance data exploration in a large research dataset on children with rare Urea Cycle Disorders (UCD).

Methods

Data source. The NIH-funded UCD natural history study includes data on over 800 participants, each with a confirmed molecular diagnosis of one of 8 different UCD subtypes followed over 12 years.[1] Data from UCD study used in our data visualization include patient identifier (anonymized), type of UCD diagnosis, visit date, structured information on hyperammonemic (HA) events (markers of disease exacerbation that are the hallmark of UCD), clinical observations from biannual physical exam and research visits. Medical history and physical exam observations were coded in SNOMED CT by research staff at each study visit, as described in (2). We imported the UCD study data into a Neo4J graph database along with the semantic relationships (i.e. the “knowledge base”) of SNOMED CT, building on the work of Campbell et al.[3] We used their information model to link study participants with multiple research visits, each with multiple observations linked to SNOMED CT codes.

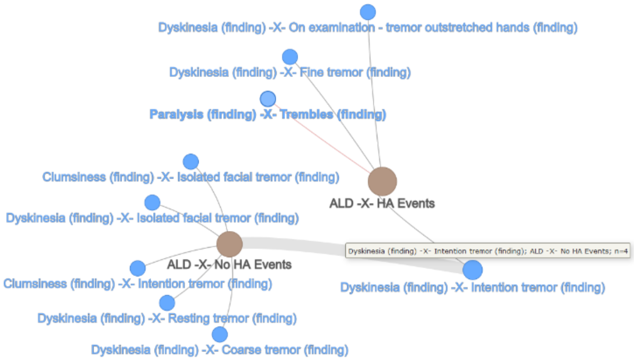
Design goals and project team. The UCD consortium[4] investigators want to explore and make full use of the rich data collected in the UCD natural history study. We collaborated with these clinical and data experts to explore the UCD and SNOMED CT data using interactive graph visualization. Because of the size of the dataset and number of SNOMED relationships, the graphs had thousands of nodes and links and were essentially unviewable. Hence, the primary design goal for the tool was 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 other patterns to be detected. To meet this goal, we developed a dynamic data visualization tool to explore the prevalence of psychiatric and neurologic abnormalities across UCD diagnoses.

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. The tool was developed in R using the Shiny, RNeo4j, and visNetwork packages. 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. In the next section we describe the architecture of the tool and some interesting challenges that we encountered.

Results

To use the visualization tool, a researcher selects SNOMED CT concepts from a hierarchical list patterned by the SNOMED CT browser (5), and additional covariates (UCD subtype, history of HA events, sex, age) to be represented as nodes in the graph. The R script queries the UCD data in the Neo4j graph database and uses it to construct node and edge descriptor tables. The descriptor tables are processed by visNetwork to produce a graph presented in a sub-window of the viewer’s web page. The user reviews the graph with various roll-over labels, zooming methods, and graph reformatting operations (bipartite, tripartite, radial edge bundled). Node and edge sizes, weights, and roll-over labels indicate numbers of distinct associated participants.

A feature of visNetwork places nodes with high mass (many participants) near the “gravitational center” of other, related nodes of lesser mass. This produces sub-networks of high mass concepts surrounded by associated diagnoses and conditions. Once a graph is rendered, nodes can be repositioned using selectable methods that affect attraction (causing connected nodes to be “dragged" along). Attraction can be disabled, so that nodes can be repositioned without affecting others. Nodes can also be programmatically repositioned, and it may be useful to implement alternative algorithms, such as for collecting nodes by type (Concept nodes in one region, Participant nodes in another, Prescription nodes in a third region). Features are implemented to subset a graph by either highlighting a node's nearest neighborhood or by truncating the graph to the neighborhood of a selected node. Nodes can also be “exploded,” such that instead of subnetting to the selected node, a graph is rendered using the children of the node. This drill-down method is useful in examining detailed concepts or diagnoses that produce high mass relationships observed at higher levels. Multiple node subsets are also possible, based on user specified node or edge filters.


Dyskinesia (finding) -X- On examination - tremor outstretched hands (finding)
Dyskinesia (finding) -X- Fine tremor (finding)
Paralysis (finding) -X- Trembles (finding)
Clumsiness (finding) -X- Isolated facial tremor (finding)
ALD -X- HA Events
Dyskinesia (finding) -X- Isolated facial tremor (finding)
Dyskinesia (finding) -X- Intention tremor (finding); ALD -X- No HA Events; n=4
ALD -X- No HA Events
Clumsiness (finding) -X- Intention tremor (finding)
Dyskinesia (finding) -X- Intention tremor (finding)
Dyskinesia (finding) -X- Resting tremor (finding)
Dyskinesia (finding) -X- Coarse tremor (finding)

We have used this tool with disease experts to explore the co-occurrence of multiple characteristics stratified by subgroup in the UCD natural history dataset. 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 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.

Discussion

The tool is interactive to allow iterative, drill-down queries and questions to be asked of the data, such that researchers can explore subsets of observations as new associations are revealed. This is a prototype tool but has been favorably reviewed by clinical investigators and data experts familiar with UCD and the research dataset. Preliminary experimentation has identified important patterns of association between UCD diagnoses and SNOMED CT concepts, such as for motor dysfunction and involuntary motion (tremors), that align with clinician intuition. A more formal evaluation of this tool and the value of SNOMED CT in the exploration of this dataset is forthcoming.

Our podium presentation will highlight specific benefits and challenges to using SNOMED CT relationships to display complex biomedical data in a manner that allows clinical experts to detect new and validate suspected relationships. This work demonstrates an approach using formalized and existing knowledge to mine and leverage and re-use existing datasets. 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.

Acknowledgements This work was supported by the UCD Consortium 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, Marcia Bowen, W. Scott Campbell, Jay Petersen, Eric Monson, Prajwal Viendra, and Sandesh C S Nagamani, MD.

References

1. Batshaw, M. L., Tuchman, M., Summar, M., Seminara, J. A longitudinal study of urea cycle disorders. In SI: Newborn Screening, Molecular Genetics and Metabolism. 2014;113(1-2):127-130.
2. Richesson, R., et al. A web-based SNOMED CT browser: distributed and real-time use of SNOMED CT during the clinical research process. Stud Health Technol Inform 2007: 129(Pt 1): 631-635.
3. Campbell, W. S., et al. An alternative database approach for management of SNOMED CT and improved patient data queries. J Biomed Inform 2015: 57: 350-357.
4. 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.
5. SNOMED International. SNOMED CT Browser, 2020. URL https://browser.ihtsdotools.org