Title: Developing an Oncology Knowledge Graph
Lead: Michael Watkins - UChicago - michaelwatkins@bsd.uchicago.edu

Introduction:
As data interoperability has risen to the forefront of clinical trial design and RWD capture, community awareness of clinical data standards has never been higher. Clinicians and data scientists alike understand that bespoke data modeling leads to complex and manual downstream harmonization. However, the resultant proliferation of clinical data standards does not fully realize data interoperability. There is a “last mile” need for computational approaches to data mapping and semantic reasoning that can leverage these standards to semi-automate the task of data harmonization.

Perhaps the most difficult aspect of interoperating over data bound to different terminological standards is that concepts are rarely exact matches and are usually partially equivalent in an ill-defined way. Knowledge graphs are a mainstay for semantic reasoning in many other industries and consist of concepts (nodes) and relations (edges). By encoding these concepts and relations in a graph representation, such as the Resource Description Framework (RDF), a reasoning language like SPARQL can query this knowledge graph and provide a user with a precise and computational relationship between two concepts.

Aims:
1. Curate a set of RDF triples that encode the relationships between oncology-related concepts from NCIt, SNOMED-CT, ICD-O, Disease Ontology, and Uberon.
2. Combine those sets into a small but linked proof-of-concept knowledge graph.
3. Develop SPARQL queries that can access the knowledge graph for given clinical terms.
4. Instantiate those queries within a data mapping demo that takes in C3DC data and annotates it with additional concept bindings from the knowledge graph.

Outline:

9/29

9:30–10:30a	Introduction to skeleton materials, RDF authoring, and terminologies (all)
11a–1p	Curate RDF sets (groups)
1–1:30p	Introduction to skeleton materials and knowledge graph architectures (all)
1:30–3p	Develop the knowledge graph (all)
3–3:30p	Introduction to skeleton materials, C3DC, and SPARQL (all)
3:30–5p	Develop SPARQL queries and sandbox C3DC data (groups)

9/30

9a–12p	Combine top queries into a deliverable demo (all)
1–2:30p	Report-Out Session

Things that will enhance participant experience:
Github account, Python experience, familiarity with oncology terminologies