MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230934

User Centered Rare Disease Clinical Trial Knowledge Graph (RCTKG)

Jeremy Parker YANG[a], Devon LEADMAN[b], Richard M. BALLEW[b,c], Eric SID[b], Yanji XU[b], Ewy A. MATHÉ[d], Qian ZHU[d,1]
[a] *University of Wisconsin-Madison, Madison, WI, US*
[b] *Division of Rare Diseases Research Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Bethesda, MD, US*
[c] *ICF International Inc, Rockville, MD, US*
[d] *Division of Pre-Clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, MD, US*
ORCiD ID: Jeremy Parker Yang https://orcid.org/0000-0003-2332-7893, Devon Leadman https://orcid.org/0000-0003-3471-7790, Richard M. Ballew https://orcid.org/0000-0003-4866-284X, Eric Sid https://orchid.org/0000-0001-7697-3026, Yanji Xu https://orcid.org/0000-0001-8033-3793, Ewy A. Mathé https://orcid.org/0000-0003-4491-8107, Qian Zhu https://orcid.org/0000-0002-4858-6333

Abstract. Drug development in rare diseases is challenging due to the limited availability of subjects with the diseases and recruiting from a small patient population. The high cost and low success rate of clinical trials motivate deliberate analysis of existing clinical trials to understand status of clinical development of orphan drugs and discover new insight for new trial. In this project, we aim to develop a user centered Rare disease based Clinical Trial Knowledge Graph (RCTKG) to integrate publicly available clinical trial data with rare diseases from the Genetic and Rare Disease (GARD) program in a semantic and standardized form for public use. To better serve and represent the interests of rare disease users, user stories were defined for three types of users, patients, healthcare providers and informaticians, to guide the RCTKG design in supporting the GARD program at NCATS/NIH and the broad clinical/research community in rare diseases.

Keywords. Clinical trial, rare disease, user story, data model, knowledge graph

1. Introduction

Of the approximately 10,000 rare diseases, [1] 90% have no effective treatments [2]. Bringing new treatments to market requires validation of their safety and efficacy, but clinical trials in rare disease space are notoriously challenging due to the small, heterogeneous, and widely dispersed patient population [3]. Deliberate analysis of existing clinical trials has thus been motivated to understand status of clinical development of orphan drugs and discover new insight for new trial design.

Knowledge graph (KG) has been widely applied in various biomedical applications, [4-6] owing to its merit of semantically managing big data, which offers an effective way

[1] Corresponding Author: Qian Zhu, email: qian.zhu@nih.gov.

for uncovering novel relationships between entities. Chen et al. recently demonstrated the value of KG in verbing clinical trials for supporting drug repurposing and similarity search [7]. Inspired by their work, we developed a user centered Rare disease based Clinical Trial Knowledge Graph (RCTKG) to serve the rare disease community.

2. Methods

We introduced a user centered clinical trial knowledge graph for rare diseases to represent clinical trials in standardized and semantic form for supporting computational analysis.

2.1. Data preparation

At the time of writing, we obtained 2,086 rare diseases that are associated with one or more clinical trials from ClinicalTrials.gov [8], and retrieved detailed information about each trail including clinical trial title/summary, eligibility criteria and etc., via the ClinicalTrials.gov API [9]. In order to link those clinical trials to rare diseases from Genetic and Rare Diseases (GARD) program at NCATS, we mapped GARD disease names/synonyms with those 2,086 rare diseases from clinicaltrials.gov based on exact name match.

2.2. User story creation

To limit the scope of this preliminary study and capture necessary information about clinical trials to the end users, three co-authors (RB, ES and QZ, who have pre-clinical and clinical pharmaceutical development, medical and biomedical informatics trainings) created user stories based on the needs from Patients, Healthcare Providers and Informaticians as our initial users of the RCTKG. They manually reviewed all attributes retrieved from the ClinicalTrials.gov and selected attributes that are applied to the created stories. At the end, they categorized the selected attributes into twelve different categories, ClinicalTrial, IndividualPatientData, Sponsor, Collaborator, Condition, StudyDesign, Participant, ExpandedAccess, Intervention, Location, PatientRegistry, Reference.

2.3. Data model definition

To formally represent data from clinical trials, we first defined a data model to capture semantics among those selected attributes and relationships between them. The data model is shown in Figure 1.

- **Primary classes.** The aforementioned twelve categories are corresponding to twelve primary classes part of the data model. An additional class of "GARD" was defined to capture information about GARD diseases.
- **Object properties.** Twelve defined primary classes are contributing complementary information to the class of "ClinicalTrial", thus, twelve object properties were defined to connect twelve primary classes to "ClinicalTrial". For example, "ClinicalTrial" class connects with its "Sponsor" via the defined object property of "sponsored_by".

- **Data properties.** Data properties were defined for each individual primary class. For instance, the class of "GARD" contains two data properties, GARDName and GARDId. Eighteen data properties are attached to the class of "ClinicalTrial", including BriefSummary, BriefTitle, CompletionDate, NCTId, OfficialTitle, Phase, OverallStatus, StartDate, StudyFirstPostDate, StudyType and etc.

Figure 1. Data model of RCTKG (the node colors are randomly rendered by Neo4j).

2.4. RCTKG development

We developed a knowledge graph (KG) in neo4j to accommodate the predefined data model to represent clinical trials. The primary classes are corresponding to the nodes, object properties are the edges represented in the KG. The RCTKG is publicly accessible via http://54.205.155.56:7474/browser/.

3. Results

In the RCTKG, there are 2,216,062 nodes and 8,437,108 edges. 2,086 GARD rare diseases are associated with one or more clinical trials. The number of nodes for each individual primary class are shown in Table 1. Notably, the total number of clinical trials is smaller than sum of clinical trials in different phases, because some clinical trials combine different phases. For example, "RAD001 and Neurocognition in PTEN Hamartoma Tumor Syndrome" (NCT02991807) is characterized as Phase 1/2.

Table 1. Statistical results of the RCTKG.

Primary Class Type	Number of Nodes
GARD	2,086
Clinical Trials (total)	182,938
•Phase N/A	96,797
•Phase 1	25,324
•Phase 2	40,669
•Phase 3	19,769
•Phase 4	11,260
Collaborator	63,767
Condition	179,632
Expanded Access	176,773
Individual Patient Data	76,586
Patient Registry	34,130
Study Design	179,354
Condition	179,632
Location	162,141

Participant	179,634
Sponsor	179,635
Reference	66,544

To demonstrate the use of the RCTKG for clinical trial landscape analysis, three use cases were performed. First, the top 10 rare diseases with the greatest number of clinical trials are retrieved and listed in Table 2, which aligns well with the fact that more trials developed for rare cancers compared to other rare diseases. Second, clinical trial distribution by intervention types is generated and shown in Table 3, which shows that most clinical trials aim at novel drug discovery for rare diseases. Third, the increasing trend of rare disease clinical trials in Phase III initiating in the past 20 years is observed and displayed in Figure 2. Clearly, increasing effort has been devoted to rare disease research in the past decades.

Table 2. Top 10 rare diseases with the greatest number of clinical trials.

GARD_ID	GARD_Name	# CT
GARD:0010964	Chronic graft versus host disease	47,217
GARD:0006544	Acute graft versus host disease	42,600
GARD:0000105	Oculocerebral syndrome with hypopigmentation	26,875
GARD:0006946	Lymphosarcoma	8,972
GARD:0003963	Neuroepithelioma	6,733
GARD:0006372	Erythema multiforme	6,695
GARD:0008226	Myeloid leukemia	4,060
GARD:0000632	Familial Alzheimer disease	3,942
GARD:0013445	Neuroendocrine tumor	3,605
GARD:0009364	Pancreatic cancer	3,567

Table 3. Number of clinical trials by intervention types.

Intervention Type	# CT
Drug	78,731
Other	34,241
Device	18,604
Procedure	16,983
Biological	14,341
Behavioral	11,503
Diagnostic Test	5,380
Radiation	5,256
Dietary Supplement	4,336
Genetic	1,392
Combination Product	817

4. Discussion

In this work, we introduced a user centered rare disease clinical trial knowledge graph (RCTKG) to convert and represent clinical trial data in a standard and semantic form. The resulting graph provides an efficient way to access information on clinical trials programmatically, thereby supporting further efforts in rare disease research. Of note, this work is preliminary and only a limited number of attributes from ClinicalTrials.gov were incorporated in the RCTKG. These attributes were chosen based on three different types of users (patients, clinicians, informaticians) by our subject matter experts. Globally, we found this graph to be useful in providing metrics for a clinical trial landscape analysis in rare diseases. Next, we plan to expand our graph by defining and incorporating other user stories to capture more clinical trial attributes, and by standardizing biomedical terms, such as conditions and interventions, to standard

terminologies, including SONMED, RxNorm for further data integration. Lastly, we recognize the ability to enhance the information content of our graph by leveraging unstructured data (e.g. eligibility criteria). While analyzing unstructured data from clinical trials is beyond the scope of this initial work, we plan to leverage advanced natural language processing (NLP) techniques to process the unstructured data.

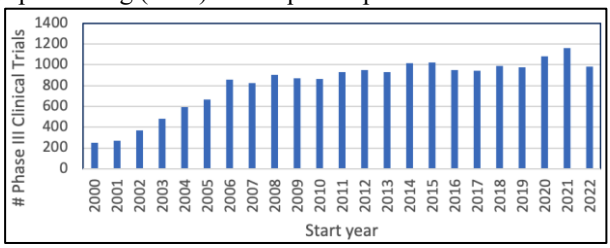

Figure 2. Trend of Phase III clinical trials initiated in the past twenty years (Notedly the number of trials in 2022 was not fully covered at the time of data retrieval).

5. Conclusions

We introduced a newly developed a rare disease based clinical trial knowledge graph, RCTKG to enable systematic landscape analysis of orphan drug development and inspire new trial design. We will integrate other types of rare disease related data with RCTKG for supporting various applications in rare diseases, including drug repurposing.

Acknowledgements

This research was supported in part by the Intramural (ZIA TR000410-03) and Extramural research program at the NCATS/NIH, and High-Value Datasets (HVD) program from the Office of Data Science Strategy (ODSS)/NIH.

References

[1] Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, Hamosh A, Baynam G, Groza T, McMurry J, Dawkins H, Rath A, Thaxon C, Bocci G, Joachimiak MP, Köhler S, Robinson PN, Mungall C, Oprea TI. How many rare diseases are there?. Nat Rev Drug Discov. 2020 Feb;19(2):77-8, doi: 10.1038/d41573-019-00180-y.

[2] Might M, Crouse AB. Why rare disease needs precision medicine-and precision medicine needs rare disease. Cell Rep Med. 2022 Feb;3(2):100530, doi: 10.1016/j.xcrm.2022.100530.

[3] Mellerio JE. The challenges of clinical trials in rare diseases. Br J Dermatol. 2022 Oct;187(4):453-4, doi: 10.1111/bjd.21686.

[4] Zeng X, Tu X, Liu Y, Fu X, Su Y. Toward better drug discovery with knowledge graph. Curr Opin Struct Biol. 2022 Feb;72:114-26, doi: 10.1016/j.sbi.2021.09.003.

[5] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023 Feb;10(1):67, doi: 10.1038/s41597-023-01960-3.

[6] Yang Y, Cao Z, Zhao P, Zeng DD, Zhang Q, Luo Y. Constructing public health evidence knowledge graph for decision-making support from COVID-19 literature of modelling study. JSSR. 2021 Sep;2(3):146–56, doi: 10.1016/j.jnlssr.2021.08.002.

[7] Chen Z, Peng B, Ioannidis VN, Li M, Karypis G, Ning X. A knowledge graph of clinical trials (CTKG). Sci Rep. 2022 Mar;12(1):4724, 10.1038/s41598-022-08454-z.

[8] Rare Diseases related Clinical Trials [Available from: https://clinicaltrials.gov/ct2/search/browse?brwse=ord_alpha_all.

[9] ClinicalTrials.gov API [Available from: https://www.clinicaltrials.gov/api/gui.