Leveraging Multiple Public Unstructured Oncology Data Knowledge Compendium for Advanced Queries, Search and Predictions

Vishakha Sharma[1], Alex Thomas[2], Vishnu Vettrivel[2], David Talby[2], Antoaneta Vladimirova[1]
[1]Roche Diagnostics, Roche Information Solutions, Santa Clara, CA; [2]John Snow Labs, Seattle, WA

Abstract

In healthcare product development and research, teams invest huge amounts of time to study through publications and other relevant resources. There is a need for a novel solution to efficiently and reliably extract information from multiple clinical resources, in addition to generating new insights which can only be achieved through structuring textual information and accessible intelligent synthesis across multiple relevant resources. We created a cloud-based solution where data from heterogeneous sources is structured, integrated and harmonized, and users can easily leverage the combined database to answer domain-specific questions and generate insights efficiently in a targeted way.

Materials and Methods

Knowledge graphs provide us the advantage to encode and leverage relationships in addition to concepts in the context of heterogeneous data. We leveraged graph and NLP AI techniques to build a domain-specific knowledge graph. We extracted the biomedically-relevant subset of wikidata, and augmented it by adding more entities and relationships from the biomedical literature (PubMed), clinical trials (clinicaltrials.gov) and NIH grants. We leveraged domain-specific named entity recognition (NER) models to identify and include rich biomedical entities.

The data is stored in data store (the graph itself), search indexes (the documents), and database tables (derived data for the visualizations).

We used an embedding model of terms and MeSH entities in order to create the scatter plot of related terms in the trend visualization. The trending terms are looking at year-over-year percentage increase in occurrences in the select set of documents.

The biomarker model is produced using features from a TransE-L2[5] embedding and a classification model. In order to make the problem tractable, the possible pairs are limited to those connected by a fixed set of paths.

Results

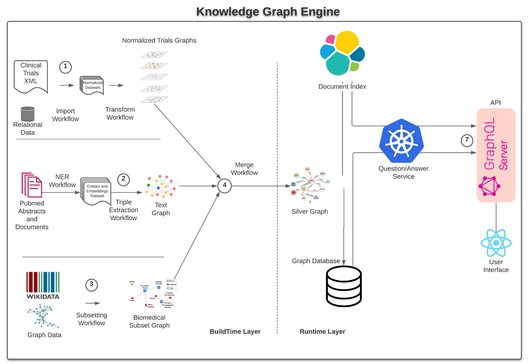


Figure 1. Schematic Diagram of Knowledge Graph Engine
Data is extracted from structured data, text, and graphs, and stored in a graph-like tabular format. The data is then merged by mapping references to an entity to a common identifier (usually the wikidata QID). The properties and relationships are mapped to predicates existing in wikidata. Once this mapping has been completed, the data is merged into a single graph. The documents (e.g. from pubmed) are stored in their own search index. The graph itself is loaded into a triple store.

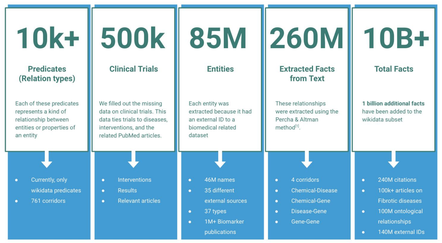


Figure 2. Summary of the knowledge graph contents
These counts are expected to increase as we add other datasets or increase what is extracted from current datasets.

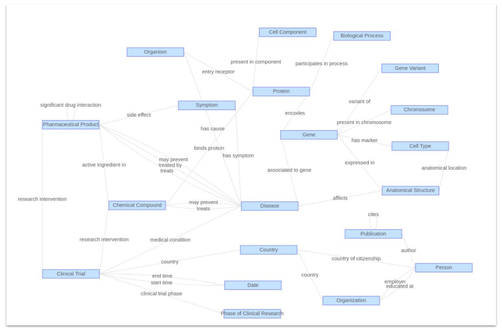
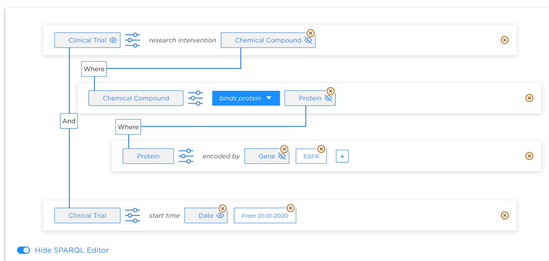


Figure 3. A sample of the schema of the knowledge graph
This visualization represents the entities and relationships of the graph subset currently accessible through the Visual Query Builder tool.

4.A. Query: Clinical trials since 01/2020 about compounds that bind to EGFR



4.B.

Figure 4. Examples of querying the graph
4.A. This is an example of the Visual Query Builder. Using this, someone can easily query data in the graph through a series of drop-downs. This means that the user does not have to learn a new query language to access the knowledge graph.
4.B. This is the SPARQL (graph query language) generated by the Visual Query Builder. Users familiar with SPARQL already, they can write their own queries.

This	This Label	Date 3	Chemical Compound 1	Chemical Compound 1 Label	Protein 1	Protein 1 Label	Gene 2	Gene 2 Label
http://www.wikidata.org/entity/Q102153531	The Rome Trial From Histology to Target: the Road to Personalize Target Therapy and Immunotherapy	2020-10-07T00:00:00.000Z	http://www.wikidata.org/entity/Q4220323	lapatinib	http://www.wikidata.org/entity/Q424401	Epidermal growth factor receptor	http://www.wikidata.org/entity/Q14865565	EGFR
http://www.wikidata.org/entity/Q102153531	The Rome Trial From Histology to Target: the Road to Personalize Target Therapy and Immunotherapy	2020-10-07T00:00:00.000Z	http://www.wikidata.org/entity/Q4220323	lapatinib	http://www.wikidata.org/entity/Q424401	epidermal growth factor receptor	http://www.wikidata.org/entity/Q14865565	EGFR
http://www.wikidata.org/entity/Q83794968	LAT for Oligoprogressive NSCLC Treated With First-line OSimertinib	2020-02-01T00:00:00.000Z	http://www.wikidata.org/entity/Q72506464	osimertinib	http://www.wikidata.org/entity/Q424401	Epidermal growth factor receptor	http://www.wikidata.org/entity/Q14865565	EGFR
http://www.wikidata.org/entity/Q83794968	LAT for Oligoprogressive NSCLC Treated With First-line OSimertinib	2020-02-01T00:00:00.000Z	http://www.wikidata.org/entity/Q72506464	osimertinib	http://www.wikidata.org/entity/Q424401	epidermal growth factor receptor	http://www.wikidata.org/entity/Q14865565	EGFR

Figure 5. Tabular results of the query in Figure 4.
The query returns both the names, and entity URIs in a tabular format.

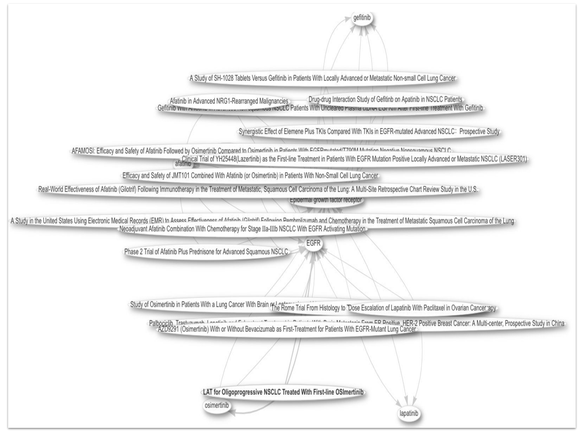


Figure 6. The graph visualization of the results in Figure 4.

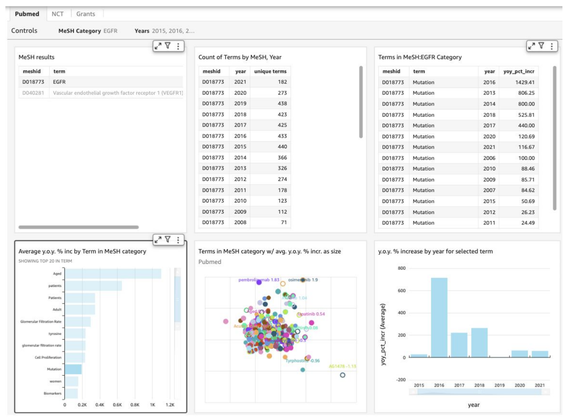
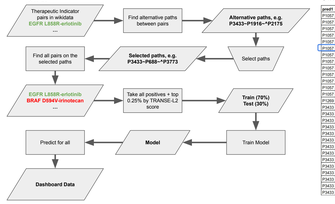


Figure 7. Trend Discovery analysis using MeSH terminology
Visualization of EGFR trend results in the Knowledge Graph across PubMed. "Mutation" term associated with EGFR and its change year over year for 2015-2020 is illustrated. Similar trends can be explored within Clinical Trials and NIH Grants information.

8.A.



8.B.

pred1	pred1name	pred2	pred2name	pred3	pred3name
P1057	chromosome	^P1057	chromosome	^P2175	medical condition treated
P1057	chromosome	^P1057	chromosome	P3971	increased expression in
P1057	chromosome	^P1057	chromosome	^P2175	medical condition treated
P1057	chromosome	^P1057	chromosome	P1916	gene substitution association with
P1057	chromosome	^P1057	chromosome	^P2175	medical condition treated
P1057	chromosome	^P1057	chromosome	P2293	genetic association
P1057	chromosome	^P1057	chromosome	^P2175	medical condition treated
P1057	chromosome	P31	instance of	^P462	color associated with
P1057	chromosome	P31	instance of	^P462	color associated with
P1057	chromosome	P703	found in taxon	^P703	found in taxon
P1057	chromosome	P703	found in taxon	^P31	instance of
P1057	chromosome	P703	found in taxon	^P3373	antagonist of
P1057	chromosome	P703	found in taxon	^P3776	inhibitor of
P1057	chromosome	P703	found in taxon	^P703	found in taxon
P1269	facet of	P31	instance of	^P31	instance of
P3433	biological variant of	^P527	has part(s)	^P2175	medical condition treated
P3433	biological variant of	P1916	gene substitution association with	^P2175	medical condition treated
P3433	biological variant of	P2293	genetic association	^P2175	medical condition treated
P3433	biological variant of	P2293	genetic association	^P279	subclass of
P3433	biological variant of	P2293	genetic association	P1889	different from
P3433	biological variant of	P2293	genetic association	P279	subclass of
P3433	biological variant of	P2293	genetic association	P921	possible treatment
P3433	biological variant of	P3373	expressed in	^P2175	medical condition treated
P3433	biological variant of	P3373	expressed in	P276	anatomical location
P3433	biological variant of	P688	encodes	^P128	physically interacts with
P3433	biological variant of	P688	encodes	^P3773	antagonist of
P3433	biological variant of	P688	encodes	P1889	different from
P3433	biological variant of	P703	found in taxon	^P128	physically interacts with
P3433	biological variant of	P703	found in taxon	^P703	found in taxon

8.C.

- Example Path #2:
  - Sequence variant -> in chromosome (P1057) -> Chromosome
  - Gene -> in chromosome (P1057) -> Chromosome
  - Disease -> genetic association (P2293) -> Gene
  - Drug -> medical condition treated (P2175) -> Disease
- Positive Pairs: 757
- Total Pairs: 650478

Figure 8. Gene Variant-to-Drug Link Prediction Task to Predict "Positive therapeutic Predictor" and "Negative Therapeutic Predictor" relationships in the Knowledge Graph: the *alternative* paths found, and used to generate the ranker training set
8.A. A table of all the paths found from sequence variant to drug. Note "^P###" means inverted relationship
8.B. A high level description of the process to train the ranker model.
8.C. Example of a single path with the counts of positive gene variant-to-drug pairs, as well as total pairs identified

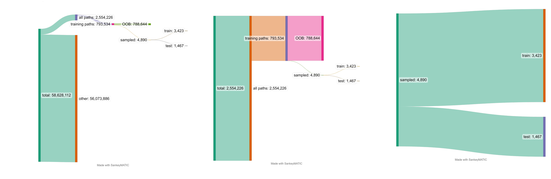


Figure 9. Constraining the data for training a ranker algorithm for predicting positive and negative therapeutic outcomes
There are 58 million possible sequence variant pairs. This is reduced by only selecting those pairs related by *alternative* paths, and then further sampling

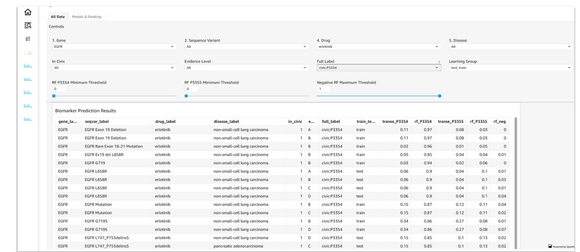


Figure 10. Examples of EGFR variants and Erlotinib drug relationship "P3354", denoting positive therapeutic prediction or predicting sensitivity/response to Erlotinib in non-small cell lung cancer.
Shown are examples or gene variants annotated in CIViC database as predictive of sensitivity/response to Erlotinib in NSCLC that were reserved in the test set during algorithm training have been predicted correctly by the ranker, and score very similarly to the examples on which the training was done.

Conclusions

We demonstrate how combining AI innovations in NLP and graph analytics, as well as novel approaches in aggregating and harmonizing disparate sources of biomedical knowledge can act as a novel and promising digital solution with potential to accelerate biomedical knowledge, answer queries, discover important trends or assist in generating new ideas, and how knowledge graphs can be used for various medical purposes such as clinical decision support and drug discovery.

References

1. Science Forum: Wikidata as a knowledge graph for the life sciences https://doi.org/10.7554/eLife.52614
2. Learning the Structure of Biomedical Relationships from Unstructured Text https://pubmed.ncbi.nlm.nih.gov/26219079/
3. Knowledge Graph Embedding for Link Prediction: A Comparative Analysis https://arxiv.org/abs/2002.00819
4. Finding melanoma drugs through a probabilistic knowledge graph https://peerj.com/articles/cs-106/
5. Translating embeddings for modeling multi-relational data https://dl.acm.org/doi/10.5555/2999792.2999923
6. https://civicdb.org/evidence/home

For additional information please contact: **antoaneta.vladimirova@roche.com**