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

## 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**

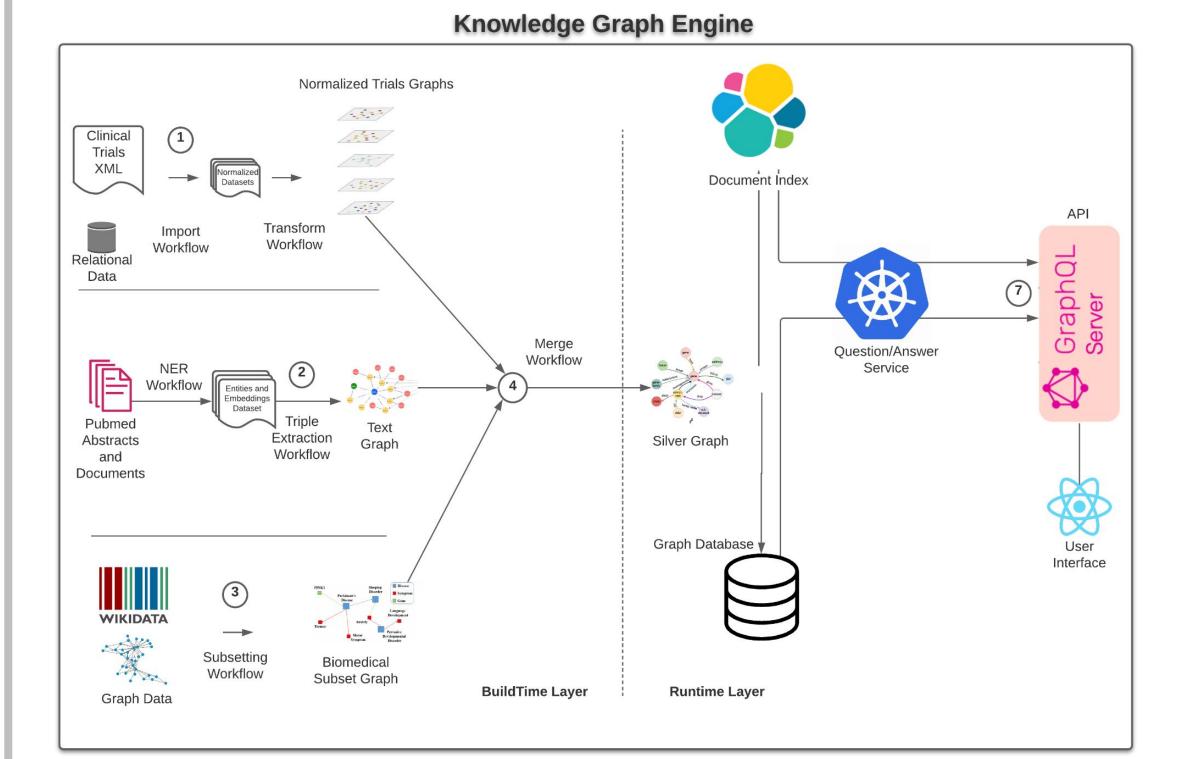
Knowledge graphs provide us the advantage to encode and leverage relationships in addition to concepts in the context of heterogeneous data. We leveraged graph and NLP AI techniques to build a domain-specific knowledge graph. We extracted the biomedically-relevant subset of wikidata, and augmented it 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<sup>5</sup> 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.

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

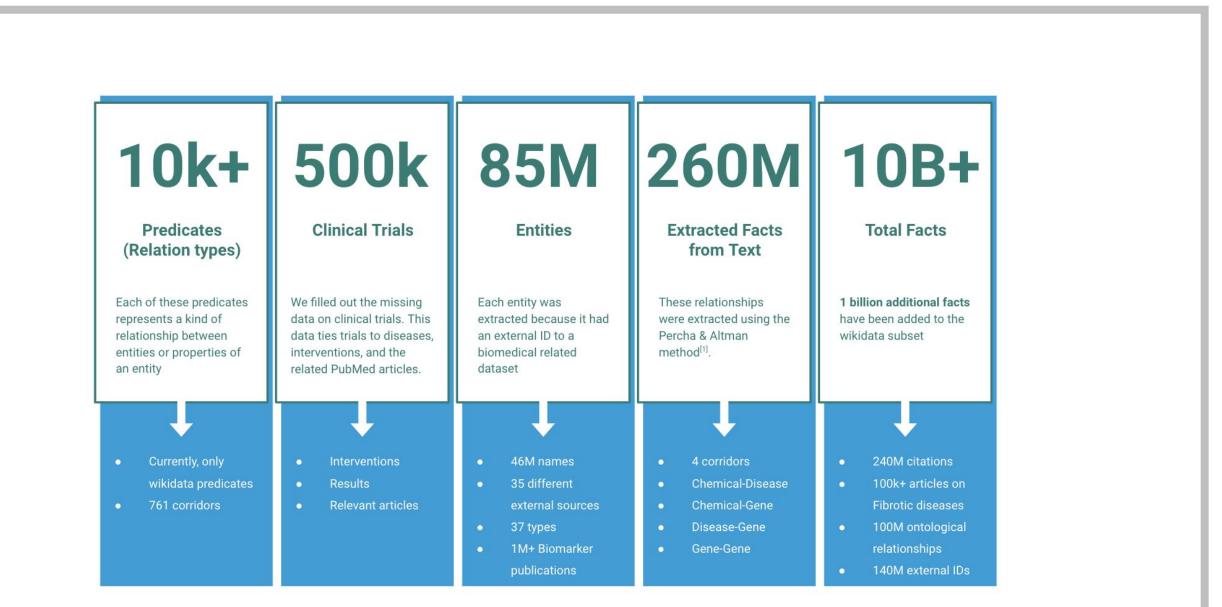
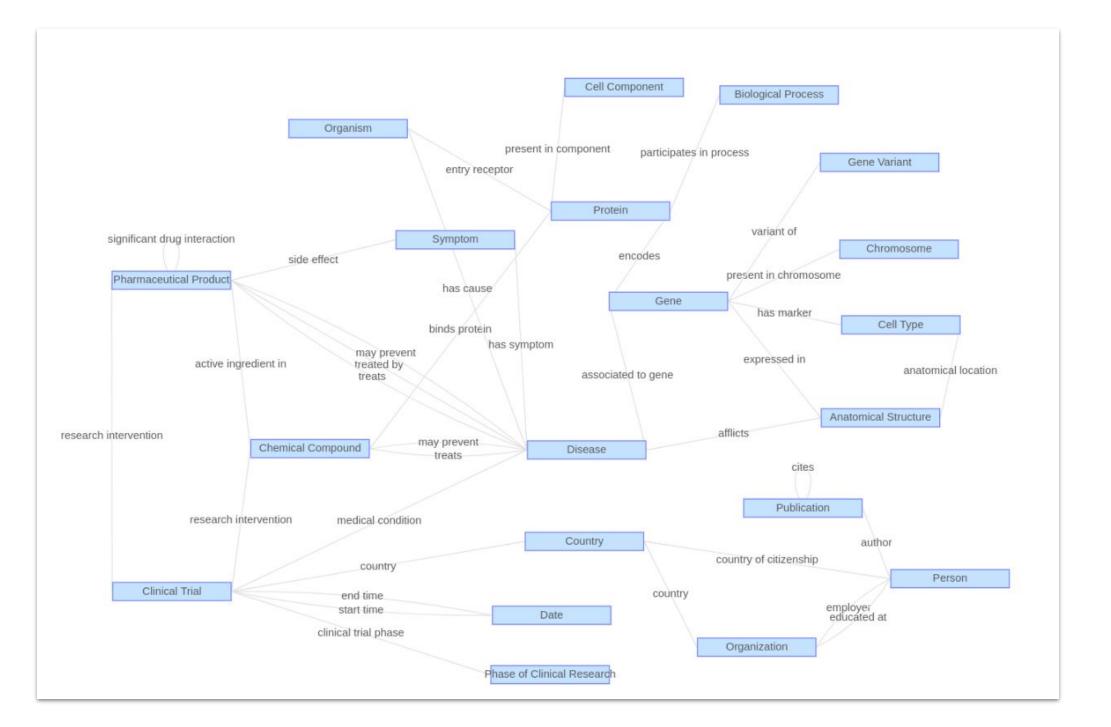
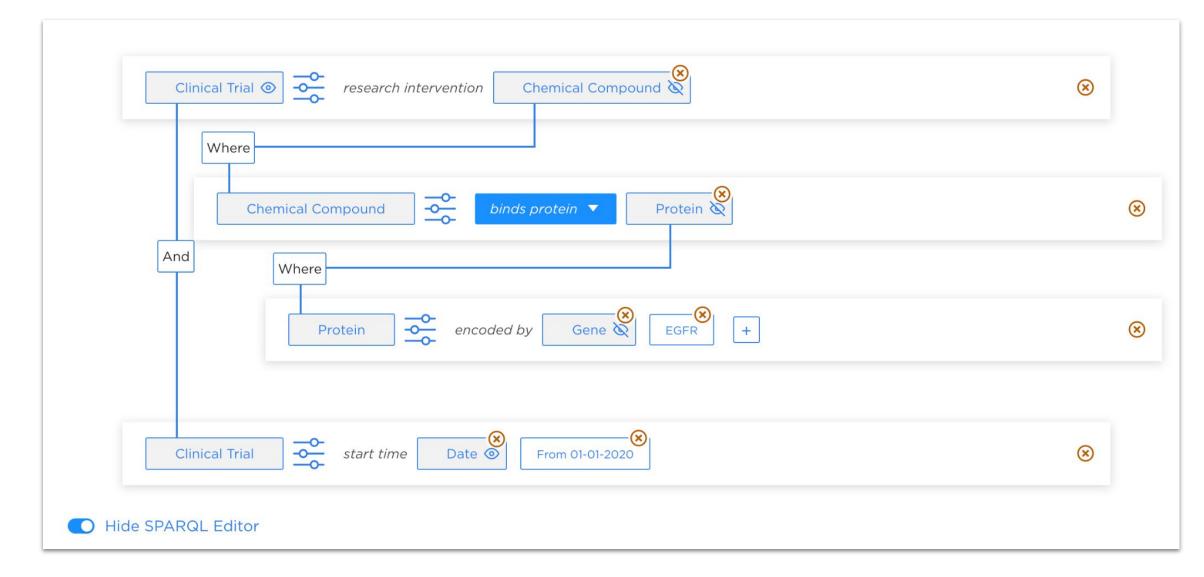


Figure 2. Summary of the knowledge graph contents These counts are expected to increase as we add other datasets or increase what is extracted from current datasets.



#### 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.

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



📋 Examples 🛛 Clear Query 🔗 Share Query 🐻 Save Query PREFIX xsd: <http://www.w3.org/2001/XMLSchema#> PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema# SELECT DISTINCT ?this ?this label ?Date 3 ?Chemical Compound 1 ?Chemical Compound 1 label ?Protein 1 ?Protein 1 label ?Gene 2 ?Gene 2 label WHERE ?this <http://www.wikidata.org/prop/direct/P31> <http://www.wikidata.org/entity/03061 ?this <http://www.wikidata.org/prop/direct/P4844> ?Chemical Compound ?Chemical Compound 1 <a href="http://www.wikidata.org/prop/direct/P31">http://www.wikidata.org/entity/011173</a> Chemical Compound 1 <a href="http://www.wikidata.org/prop/direct/P129">http://www.wikidata.org/prop/direct/P129</a> ?Protein ?Protein 1 <http://www.wikidata.org/prop/direct/P31> <http://www.wikidata.org/entity/Q8054? ?Protein 1 <http://www.wikidata.org/prop/direct/P702> ?Gene 2. VALUES ?Gene 2 <http://www.wikidata.org/entity/Q14865565> ?Gene\_2 <http://www.w3.org/2000/01/rdf-schema#label> ?Gene 2 label. ETT MED (TANGNAMOURC (TANG (DCare 2 label) Her!)

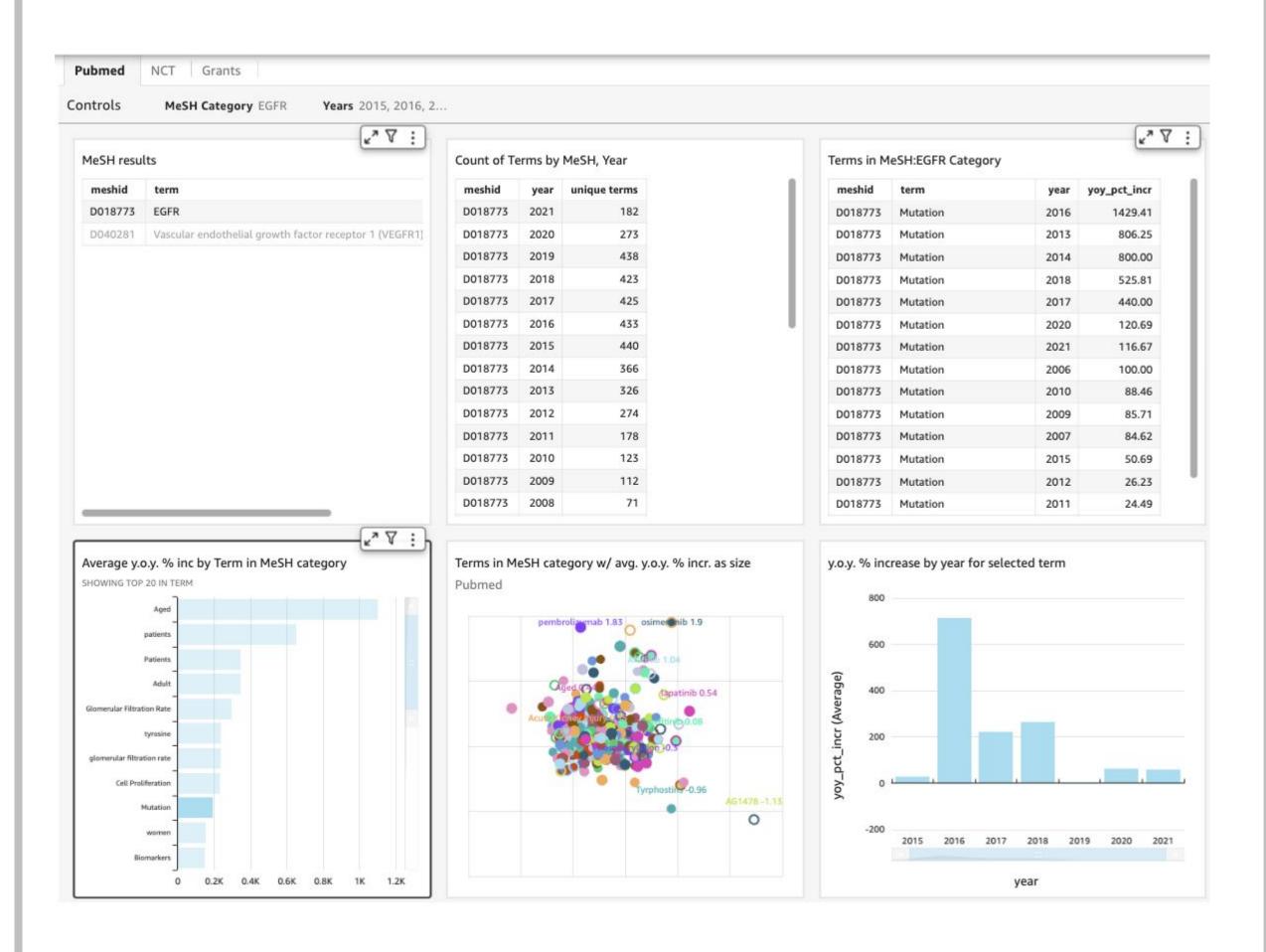
### Figure 4. Examples of querying the graph

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.

#### Gene 2 Labe listology to Target: t http://www.wikidata. Road to Personaliz /entity/Q10215353 Target Therapy and mmunotheran The Rome Trial Fro Histology to Target: th http://www.wikidata.org epidermal growth http://www.wikidata. http://www.wikidata.org http://www.wikidata.org Road to Personalize /entity/Q102153531 /entity/Q420323 07T00:00:00.000Z /entity/Q424401 Target Therapy and mmunotherapy LAT for http://www.wikidata.org Oligoprogressive http://www.wikidata.org http://www.wikidata.org Epidermal growth http://www.wikidata.org /entity/Q21506464 LAT for http://www.wikidata.org Oligoprogressive http://www.wikidata.org http://www.wikidata.org /entity/Q83794168 NSCLC Treated With 01T00:00:00.000Z /entity/Q21506464 /entity/Q424401 First-line OSImertin Figure 5. Tabular results of the query in Figure 4. The query returns both the names, and entity URIs in a tabular format. Study of SH-1028 Tablets Versus Gefitinib in Patients With Locally Advanced or Metastatic Non-small Cell Lung Cancer Afatinib in Advanced NRG1-Rearranged Malignancies Drug-drug Interaction Study of Gefitinb on Apatinib in NSCLC Patients WINDER Patients With Uncleared Mastria GUINA EDERNITATED FIRST-line Treatment With C nergistic Effect of Elemene Plus TKIs Compared With TKIs in EGFR-mutated Advanced NSCLC: Prospective Study FAMOSI: Efficacy and Safety of Afatinib Followed by Osimertinib Compared to Osimertinih in Patiente With ECEP mutated/T700M Mutation Negative Nonequamous NSC linical Trial of YH25448(Lazertinib) as the First-line Treatment in Patients With EGFR Mutation Positive Locally Advanced or Metastatic NSCLC (LASER301) fficacy and Safety of JMT101 Combined With Afatinib (or Osimertinib) in Patients With Non-Small Cell Lung Cancer Real-World Effectiveness of Afatinib (Gilotrif) Following Immunotherapy in the Treatment of Metastatic, Squamous Cell Carcinoma of the Lung: A Multi-Site Retrospective Chart Review Study in the U.S. A Study in the United States Using Electronic Medical Records (EMR) to Assess Effectiveness of Afatinib (Cilotrif) Following Pembrolizumab and Chemotherapy in the Treatment of Metastatic Squamous Cell Carcinoma of the Lung-Neoadjuvant Afatinib Combination With Chemotherapy for Stage IIa-IIIb NSCLC With EGFR Activating Mutation Phase 2 Trial of Afatinib Plus Prednisone for Advanced Squamous NSCLC Study of Osimertinib in Patients With a Lung Cancer With Brain or he Rome Trial From Histology to Dose Escalation of Lapatinib With Paclitaxel in Ovarian Cancer apy albociclib. Trastuzumab Longlinib and Education of the state of the st T for Oligoprogressive NSCLC Treated With First-line OSImertin

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.



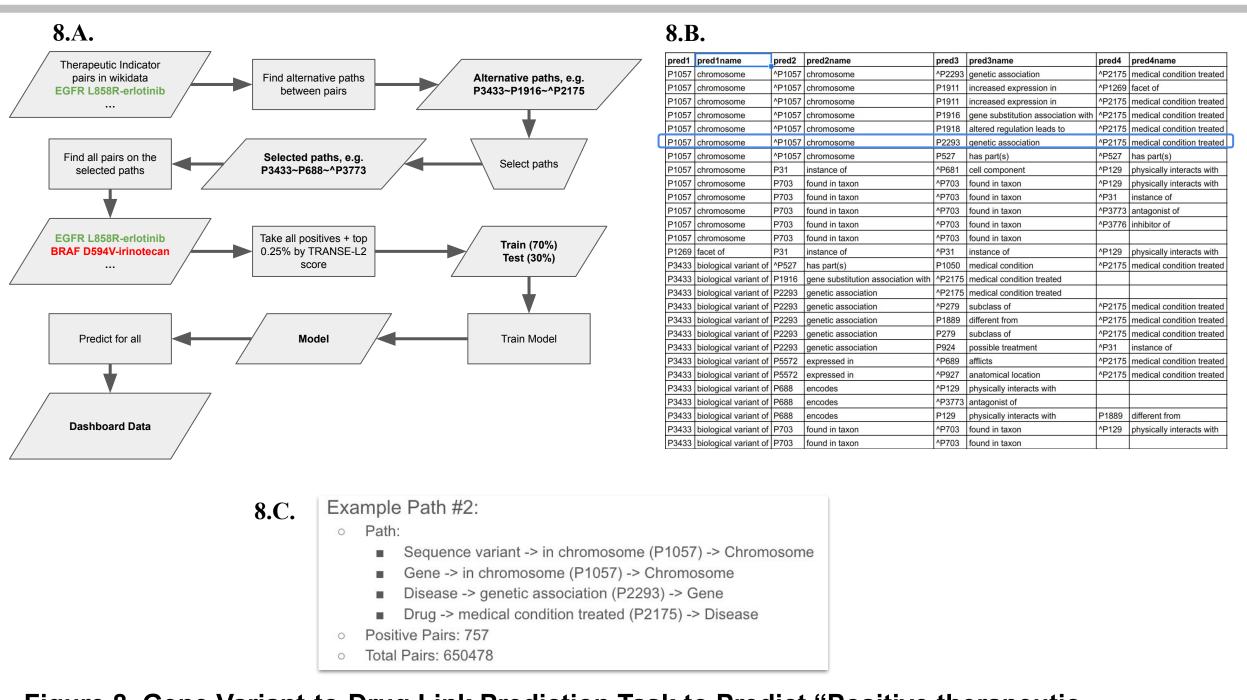
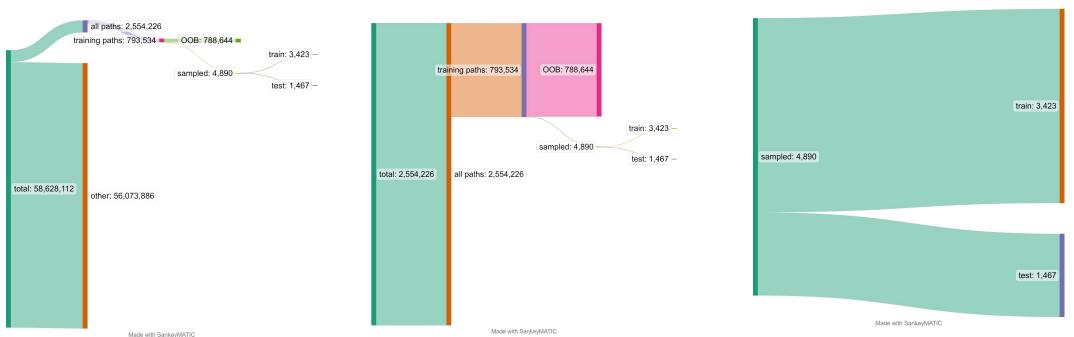


Figure 8. Gene Variant-to-Drug Link Prediction Task to Predict "Positive therapeutic Predictor" and "Negative Therapeutic Predictor" relationships in the Knowledge Graph: the *alternative* paths found, and used to generate the ranker training set **8.A.** A table of all the paths found from sequence 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

	Models & Ranking													
Controls														
1. Gene			2. Sequence Variant			4. Drug				5. Diseas	5. Disease			
EGFR			All			erlotinib				▼ All	All			
In Civic			Evidence Level			Full Label     :       civic:P3354     *					Learning Group test, train			
										<ul> <li>test, train</li> </ul>				
RE P3354	1inimum Threshold		RF P3355 Minimum Threshold			Negative RF Maxir	num Threshold							
			0			1								
•			•							•				
Biomar	er Prediction Results													
gene_l	seqvar_label	drug_label	disease_label	in_civic	e	full_label	train_te	transe_P3354	rf_P3354	transe_P3355	rf_P3355	rf_neg		
EGFR	EGFR Exon 19 Deletion	erlotinib	non-small-cell lung carcinoma	1	А	civic:P3354	train	0.11	0.97	0.08	0.03	0		
EGFR	EGFR Exon 19 Deletion	erlotinib	non-small-cell lung carcinoma	1	в	civic:P3354	train	0.11	0.97	0.08	0.03	0		
EGFR	EGFR Rare Exon 18-21 Mutation	erlotinib	non-small-cell lung carcinoma	1	В	civic:P3354	train	0.02	0.96	0.01	0.05	0		
EGFR	EGFR Ex19 del L858R	erlotinib	non-small-cell lung carcinoma	1	в	civic:P3354	train	0.05	0.95	0.04	0.04	0.01		
EGFR	EGFR G719	erlotinib	non-small-cell lung carcinoma	1	в	civic:P3354	train	0.03	0.94	0.02	0.06	0		
EGFR	EGFR L858R	erlotinib	non-small-cell lung carcinoma	1	А	civic:P3354	test	0.06	0.9	0.04	0.1	0.01		
EGFR	EGFR L858R	erlotinib	non-small-cell lung carcinoma	1	В	civic:P3354	test	0.06	0.9	0.04	0.1	0.03		
EGFR	EGFR L858R	erlotinib	non-small-cell lung carcinoma	1	с	civic:P3354	test	0.06	0.9	0.04	0.1	0.01		
EGFR	EGFR L858R	erlotinib	non-small-cell lung carcinoma	1	D	civic:P3354	test	0.06	0.9	0.04	0.1	0.04		
EGFR	EGFR Mutation	erlotinib	non-small-cell lung carcinoma	1	В	civic:P3354	train	0.15	0.87	0.12	0.11	0.04		
EGFR	EGFR Mutation	erlotinib	non-small-cell lung carcinoma	1	с	civic:P3354	train	0.15	0.87	0.12	0.11	0.02		
EGFR	EGFR G719S	erlotinib	non-small-cell lung carcinoma	1	в	civic:P3354	train	0.34	0.86	0.27	0.08	0.07		
EGFR	EGFR G719S	erlotinib	non-small-cell lung carcinoma	1	D	civic:P3354	train	0.34	0.86	0.27	0.08	0.07		
EGFR	EGFR L747_P753delinsS	erlotinib	non-small-cell lung carcinoma	1	D	civic:P3354	test	0.15	0.85	0.1	0.13	0.02		
EGFR	EGFR L747_P753delinsS	erlotinib	pancreatic adenocarcinoma	1	с	civic:P3354	test	0.15	0.85	0.1	0.13	0.02	Po	

#### 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 <u>https://doi.org/10.7554/eLife.52614</u> 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 <u>https://arxiv.org/abs/2002.00819</u>
- 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. <u>https://civicdb.org/evidence/home</u>

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