EpiGraphDB-ASQ as a natural language interface to biomedical knowledge graph

RSE workshop in data & AI workshop, 16 Febuary 2023

Yi Liu
MRC Integrative Epidemiology Unit, University of Bristol

whoami

Senior research associate in health data science,
MRC Integrative Epidemiology Unit, University of Bristol

Data mining epidemiological relationship programme
(Programme Lead: Professor Tom Gaunt)

- Lead EpiGraphDB working group
- Architect and lead developer on EpiGraphDB platform and components
- Data mining and knowledge discovery with knowledge graph and machine learning methods

Outline

- Previous research projects
  - The EpiGraphDB knowledge graph
  - The BlueBERT-EFO model
- ASQ as a natural language interface
  - Entity harmonization
  - Evidence groups
  - Evidence prioritization

Triangulating evidence in health sciences with Annotated Semantic Queries

Yi Liu[1,*] and Tom R Gaunt[1,2,*]

[1]MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK
[2]NIHR Bristol Biomedical Research Centre, University of Bristol, Bristol, UK
[*]corresponding authors

ABSTRACT

Integrating information from data sources representing different study designs has the potential to strengthen evidence in population health research. However, this concept of evidence "triangulation" presents a number of challenges for systematically identifying and integrating relevant information. We present ASQ (Annotated Semantic Queries), a natural language query interface to the integrated biomedical entities and epidemiological evidence in EpiGraphDB, which enables users to extract "claims" from a piece of unstructured text, and then investigate the evidence that could either support, contradict the claims, or offer additional information to the query. This approach has the potential to support the rapid review of pre-prints, grant applications, conference abstracts and articles submitted for peer review. ASQ implements strategies to harmonize biomedical entities in different taxonomies and evidence from different sources, to facilitate evidence triangulation and interpretation. ASQ is openly available at `https://asq.epigraphdb.org`.

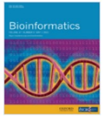
1 Introduction

Researchers in health sciences are encouraged to seek multiple strands of complementary evidence to minimise the risk of bias creating false positives. This has been referred to as the *triangulation*[1] of evidence, which may combine results from different study designs with different sources of bias, including from established findings in the literature. Platforms which offer a portal to integrated heterogeneous data such as Open Targets[2] and EpiGraphDB[3] are highly valuable sources which have the potential to support evidence triangulation by integrating evidence with relevant information from a range of dedicated data providers, including biomedical ontologies[45], genetic associations[6] and literature-derived evidence[7]. One of the main objectives for the web interface of such integrated data platforms is to present users with focused information from various integrated sources in order to facilitate the fast navigation and discovery

Liu, Gaunt., 2022 MedRxiv

Previous work #1: the EpiGraphDB platform

Liu, et al., Gaunt., 2021 Bioinformatics



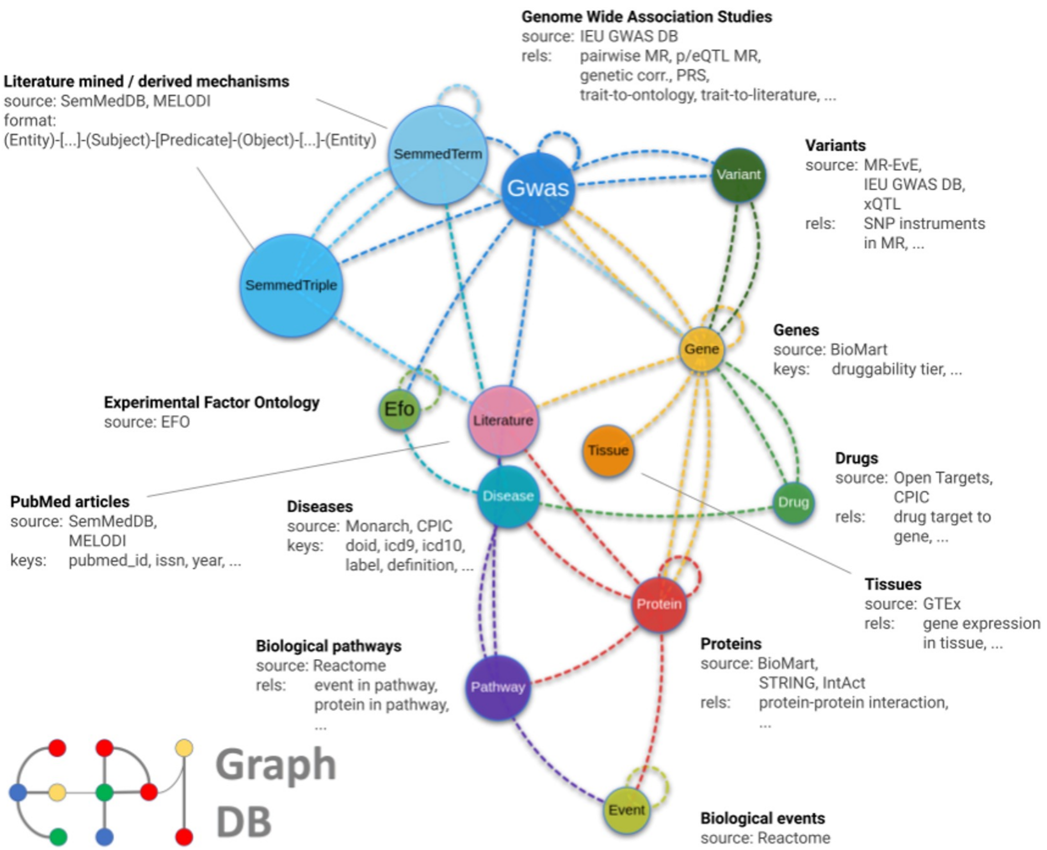
Volume 37, Issue 9
1 May 2021

EpiGraphDB: a database and data mining platform for health data science

Yi Liu, Benjamin Elsworth, Pau Erola, Valeriia Haberland, Gibran Hemani, Matt Lyon, Jie Zheng, Oliver Lloyd, Marina Vabistsevits, Tom R Gaunt   Author Notes

Bioinformatics, Volume 37, Issue 9, 1 May 2021, Pages 1304–1311,
https://doi.org/10.1093/bioinformatics/btaa961
Published: 24 November 2020   **Article history** ▾

EpiGraphDB



EpiGraphDB v1.0

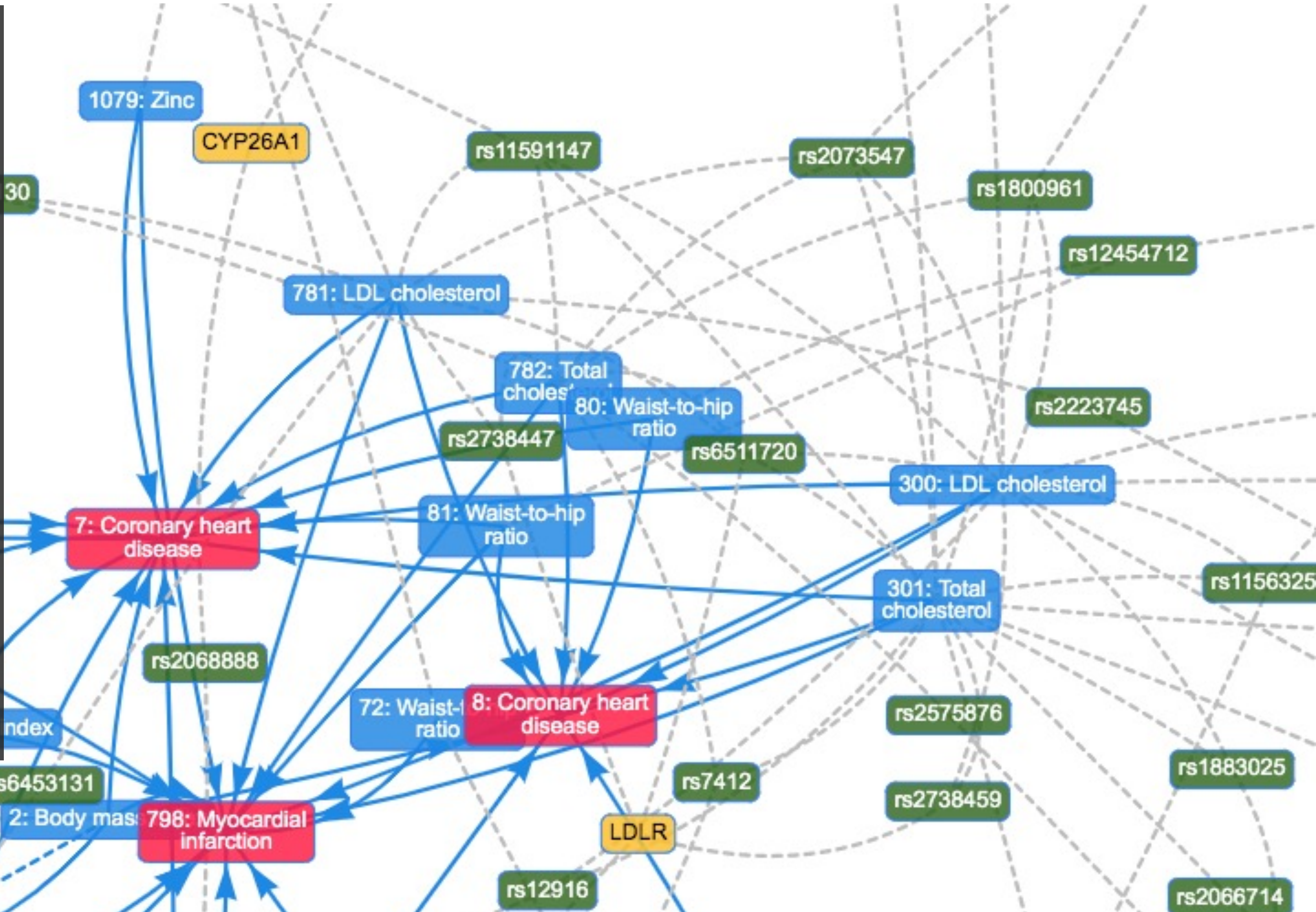
- # nodes: 9,995,580
- # edges: 204,943,810
- # node types: 12
- # edge types: 38

Integrated epidemiological evidence
http://docs.epigraphdb.org

- Causal relationships
- Association relationships
- Molecular pathways
- Literature mined / derived evidence
- Others

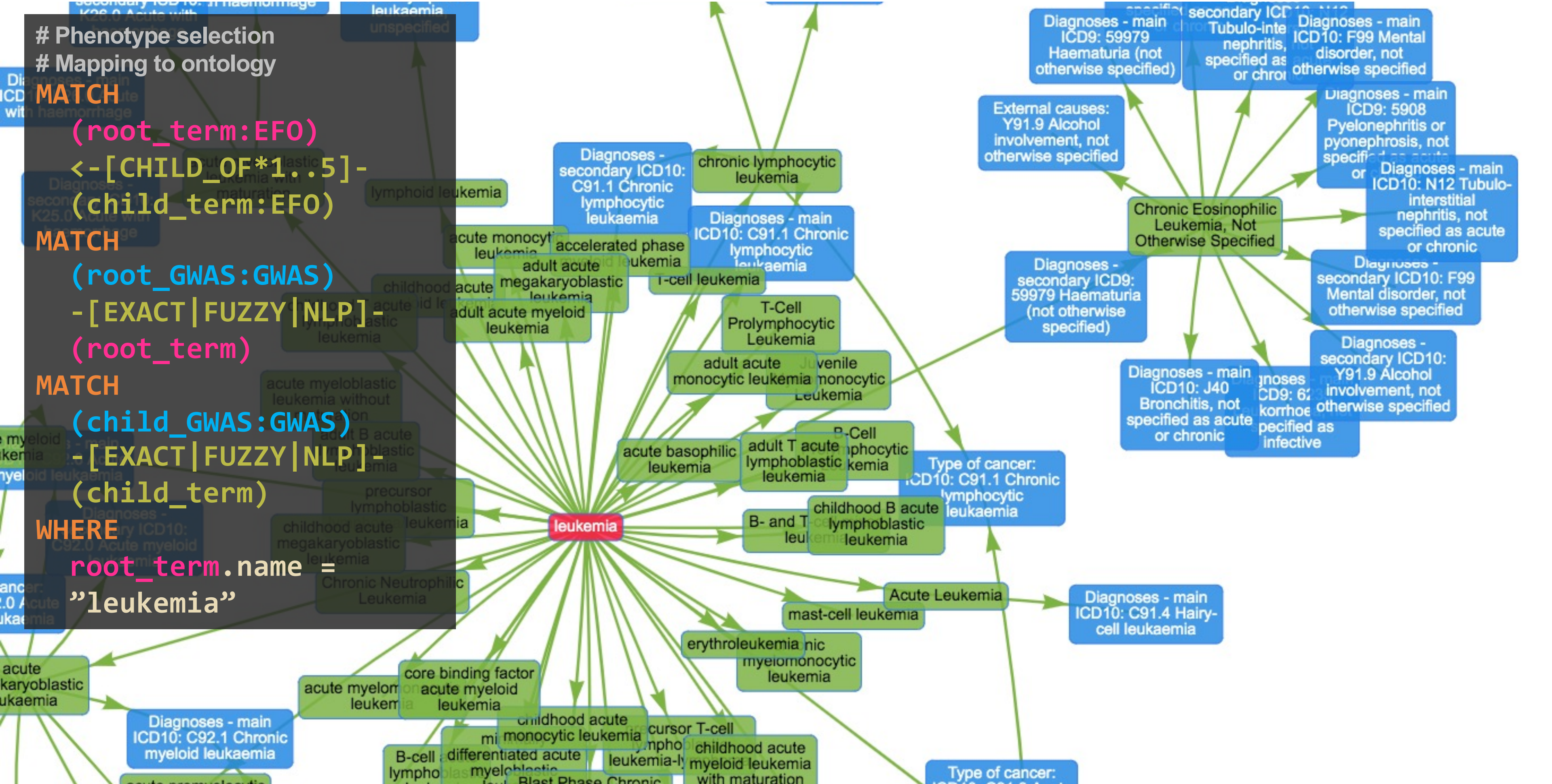
EpiGraphDB query example

```
# Exploration of causal pathways of cardiovascular risk factors
# And identification of drug targets
MATCH
  (disease:Trait)
  <-[MR]-
  (risk_factor:Trait)
  -[ASSOC]-(SNP)
  -[ASSOC]-(Gene)
  -[ASSOC]-(Drug)
WHERE
  disease.category = "cardiovascular"
AND
  MR.method = "IVW"
```



EpiGraphDB query example

```
# Phenotype selection
# Mapping to ontology
MATCH
  (root_term:EFO)
  <-[CHILD_OF*1..5]-
  (child_term:EFO)
MATCH
  (root_GWAS:GWAS)
  -[EXACT|FUZZY|NLP]-
  (root_term)
MATCH
  (child_GWAS:GWAS)
  -[EXACT|FUZZY|NLP]-
  (child_term)
WHERE
  root_term.name =
  "leukemia"
```

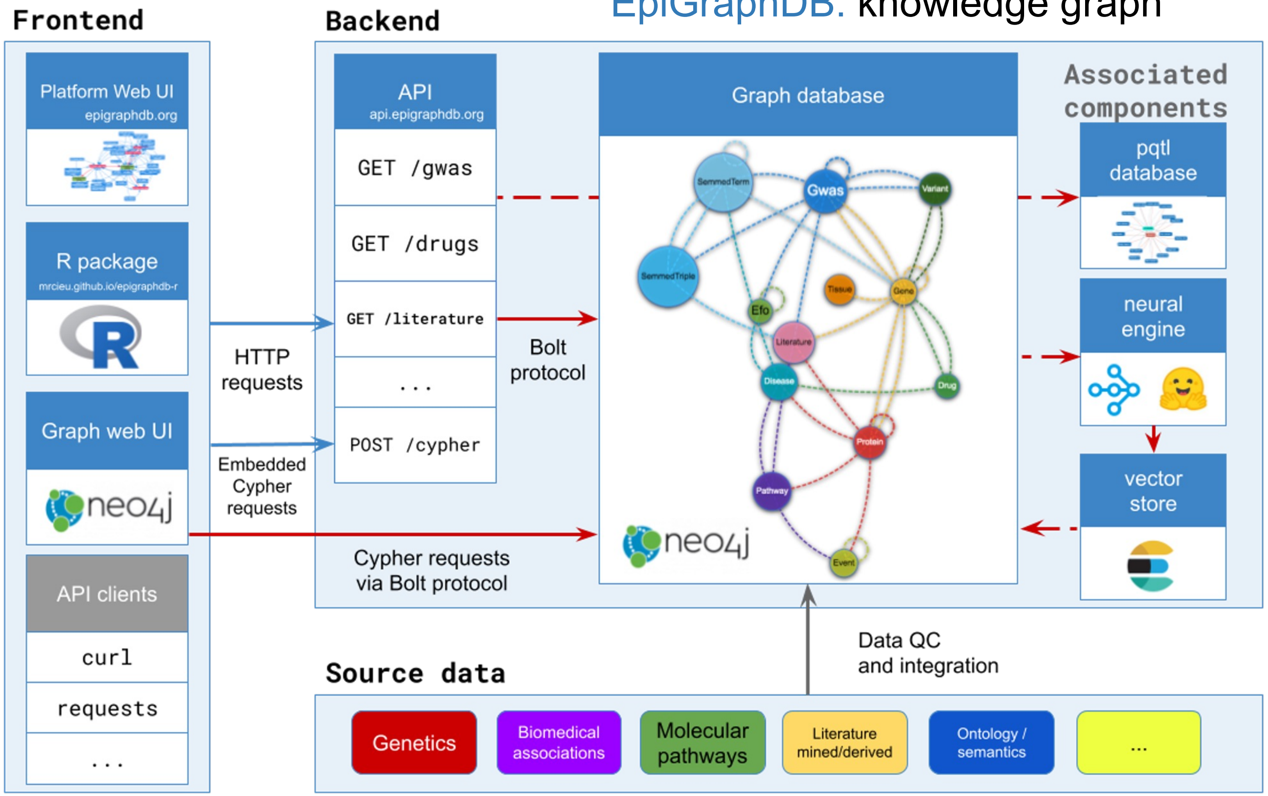


EpiGraphDB and IEU health data science

MRC Integrative Epidemiology Unit

University of BRISTOL

EpiGraphDB: knowledge graph



OpenGWAS: GWAS consortium

ieu open gwas project

MELODI Presto: Literature enrichment

MELODI

TwoSampleMR, MR-EvE, xQTL-MR, etc.: Biostatistical analytics

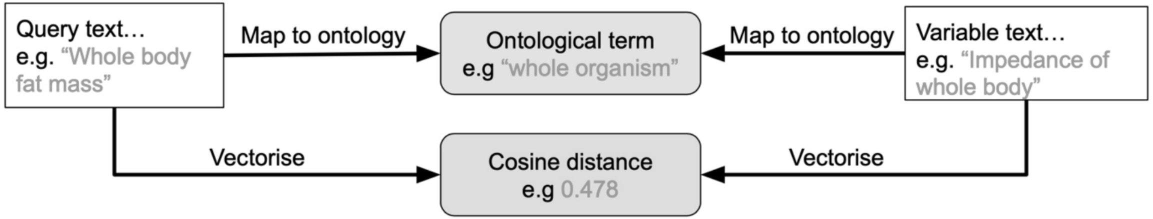
https://epigraphdb.org

https://github.com/mrcieu/epigraphdb

Previous work #2: Trait Mapping

Liu, Elsworth, Gaunt, 2022 BioRxiv,
Using language model and ontology topology to perform semantic mapping of traits between biomedical datasets

Previous #2: Trait mapping

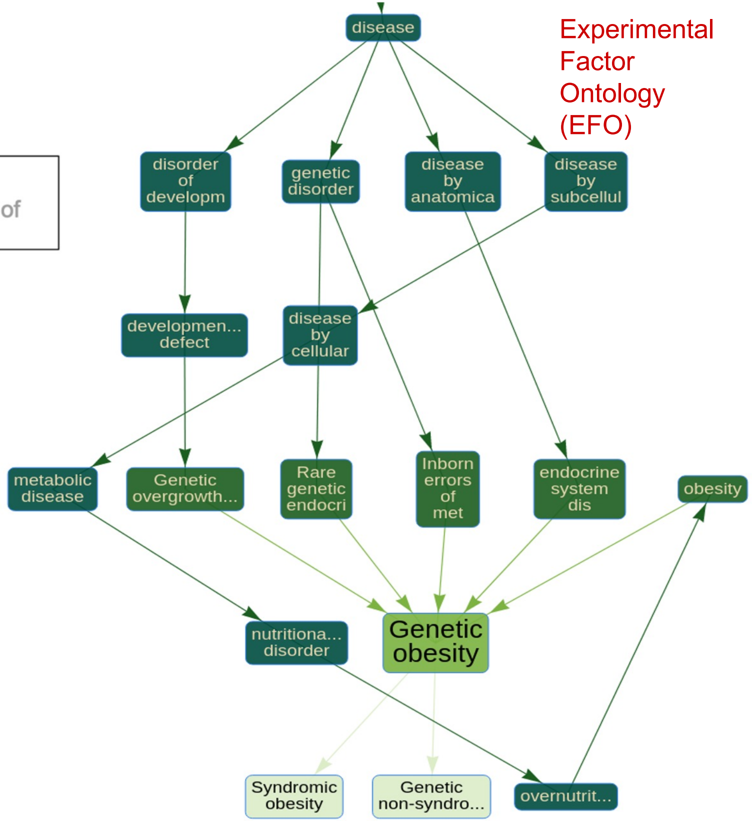


An ontology distance predictor based on the BERT language model

BlueBERT-EFO :=

BlueBERT (Peng, Yan, Lu, 2019)

+ Experimental Factor Ontology



ML training

- Training data:
  Experimental Factor Ontology
    - EFO as a graph
    - Pairwise distance of ontology terms
        - Shortest distance between two nodes
        - Self distance of a node of its synonyms
- Finetuning a transformer language model with a sequence classification task

cancer-related condition

Search EFO

http://purl.obolibrary.org/obo/MONDO_0045054 Copy

A disorder either associated with an increased risk for malignant transformation (e.g., intraepithelial neoplasia, leukoplakia, dysplastic nevus, myelodysplastic s
develops as a result of the presence of an existing malignant neoplasm (e.g., paraneoplastic syndrome). [NCIT : C8278]

Synonyms: `problem/condition, cancer-related` `cancer-related condition` `cancer related problem/condition` `cancer-related problem or condition` `problem/condition, cancer related` `oncologic complications`

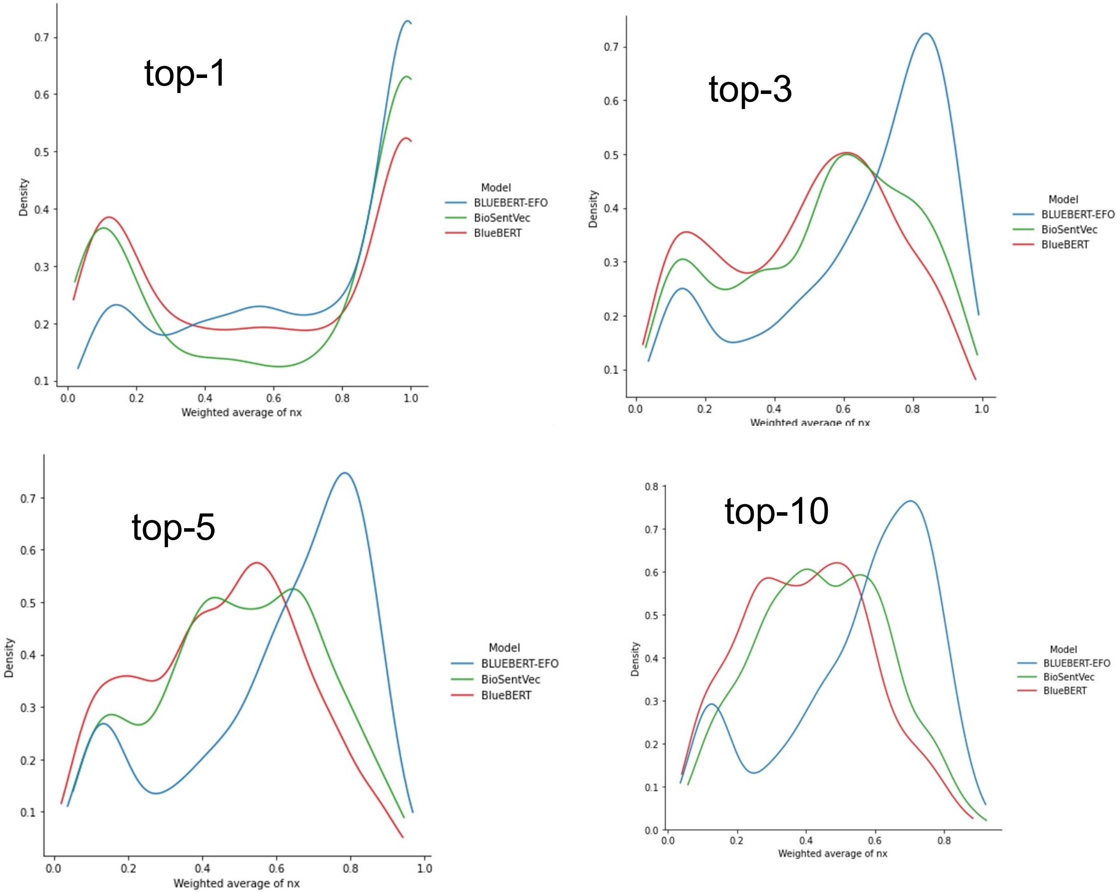
	trait	efo_term	pred	target	diff
0	Malignant mesothelioma	mesothelioma	1.000238	1.0	0.000238
1	Metabolite levels	metabolite measurement	0.999371	1.0	0.000629
2	Tyrosine levels	tyrosine measurement	1.000991	1.0	0.000991
3	Butyrylcholinesterase levels	butyrylcholinesterase measurement	0.998593	1.0	0.001407
4	Hypertension (SNP x SNP interaction)	hypertension	1.001603	1.0	0.001603
5	Optic disc area	optic disc area measurement	0.998011	1.0	0.001989
6	Esophageal cancer (alcohol interaction)	esophageal carcinoma	1.003343	1.0	0.003343
7	Obsessive-compulsive disorder or autism spectr...	obsessive-compulsive disorder, autism spectrum...	0.995708	1.0	0.004292
8	Vestibular neuritis	vestibular neuronitis	0.995653	1.0	0.004347
9	Nonalcoholic fatty liver disease	non-alcoholic fatty liver disease	1.004780	1.0	0.004780
10	Pit-and-Fissure caries	pit and fissure surface dental caries	0.995210	1.0	0.004790
11	White matter lesion progression	white matter lesion progression measurement	1.004839	1.0	0.004839
12	Prostate cancer (SNP x SNP interaction)	prostate carcinoma	0.993228	1.0	0.006772
13	Large artery stroke (TOAST classification)	large artery stroke	0.993190	1.0	0.006810
14	Pulse pressure (dietary potassium intake inter...	pulse pressure measurement, dietary potassium ...	0.992917	1.0	0.007083
15	Schizophrenia or cigarettes per day (pleiotropy)	schizophrenia, cigarettes per day measurement	1.008028	1.0	0.008028
16	Hypersomnia (HLA-DQB1*06:02 negative)	hypersomnia	0.991550	1.0	0.008450
17	Prostate cancer (early onset)	prostate carcinoma	0.991139	1.0	0.008861
18	Hemoglobin concentration	hemoglobin measurement	1.012157	1.0	0.012157
19	Niacinamide levels	niacinamide measurement	0.987777	1.0	0.012223

Semantic phenotype harmonization

Ontology classifier (BlueBERT-EFO) based on Transformer language model greatly improved relevancy of candidate retrieval

Trait-to-Trait relationships predicted by BlueBERT-EFO resemble their corresponding representation in the ontology

Serves as the basis of entity harmonization in evidence triangulation



EpiGraphDB-ASQ (Annotated Semantic Queries, ASQ)

Liu, Gaunt, 2022 MedRxiv,
Triangulating evidence in health sciences with Annotated Semantic Queries

ASQ: Fact checking scientific claims

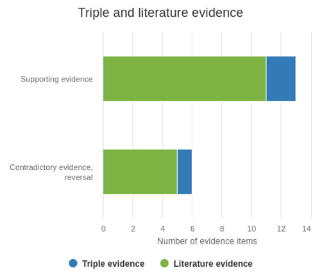
Scientific text: "Obesity subj: Obesity causes substantial changes to the mechanics of the lungs and chest wall, and these mechanical changes cause pred: CAUSES asthma obj: Asthma and asthma-like symptoms such as dyspnea, wheeze, and airway hyperresponsiveness"

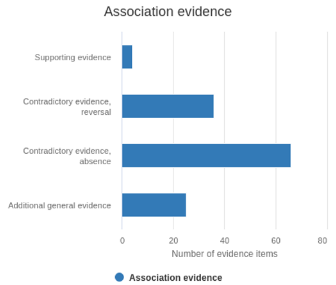
claim triple

Obesity CAUSES Asthma

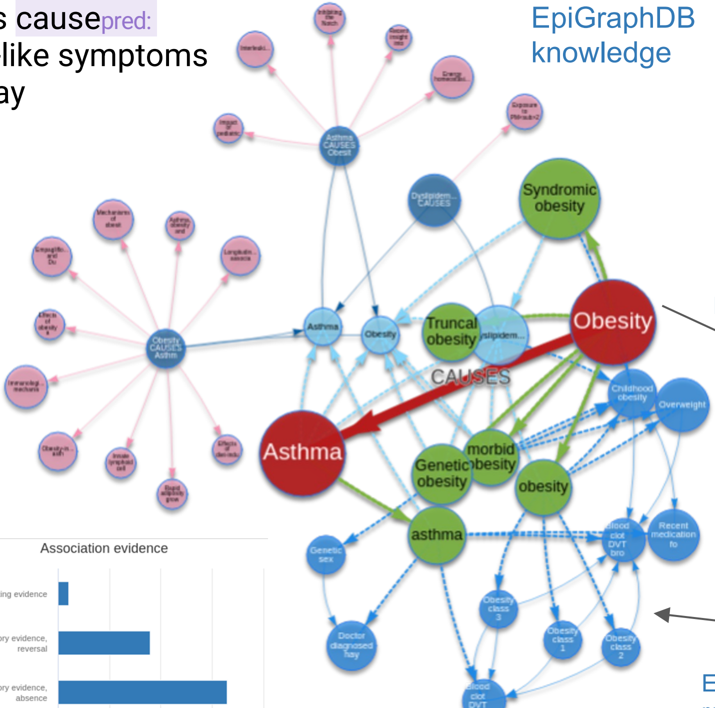
EpiGraphDB

- Biomedical entities
- Supporting / contradictory evidence





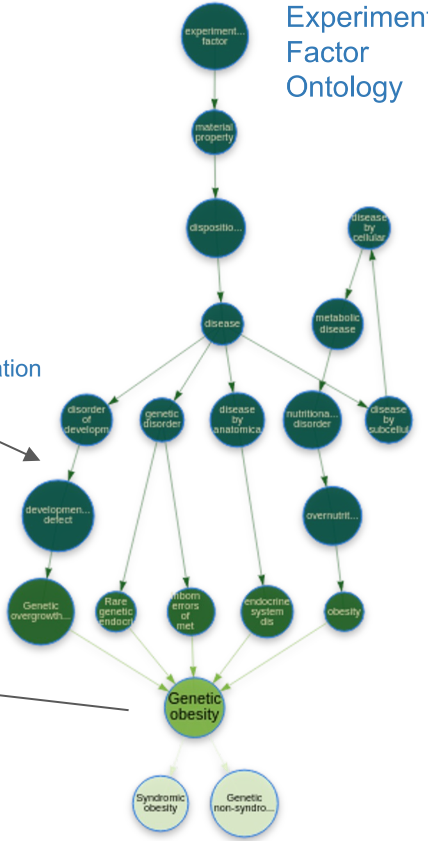
Evidence triangulation with EpiGraphDB knowledge



Entity harmonization

Evidence retrieval

Experimental Factor Ontology

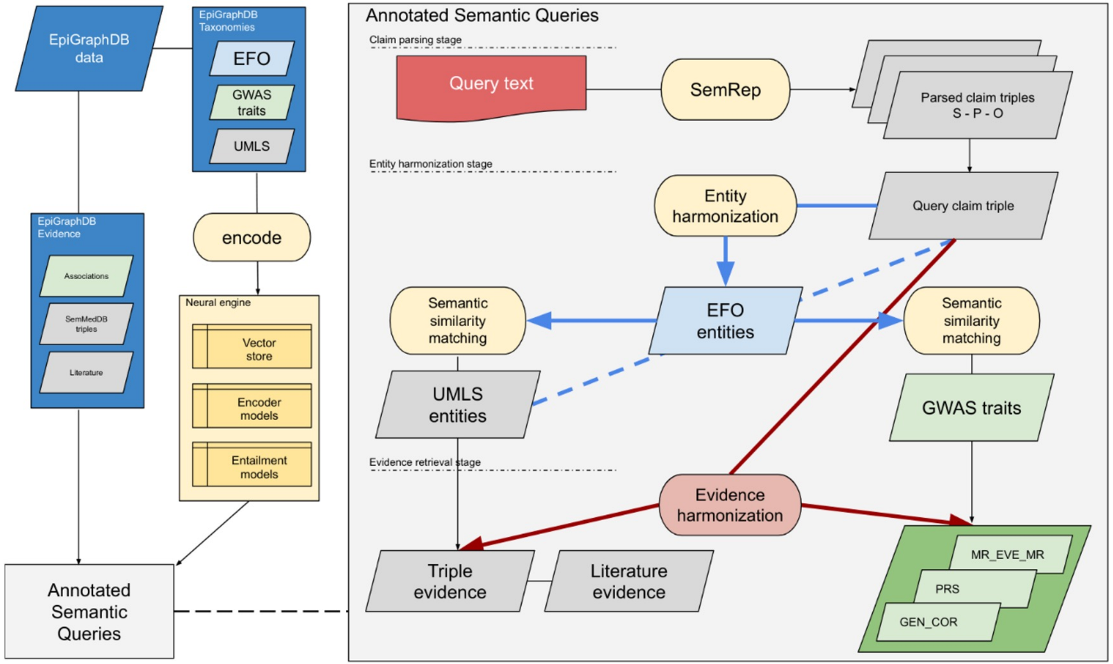


elasticsearch

Annotated Semantic Queries (ASQ)

"Fact checking" biomedical claims

- EpiGraphDB curated knowledge
- Entity harmonization
- Evidence harmonization
- Evidence groups w.r.t. type of the evidence (e.g. literature, statistics)
- Evidence groups w.r.t. the claim (supporting, contradictory, etc.)



From free text to structured entities

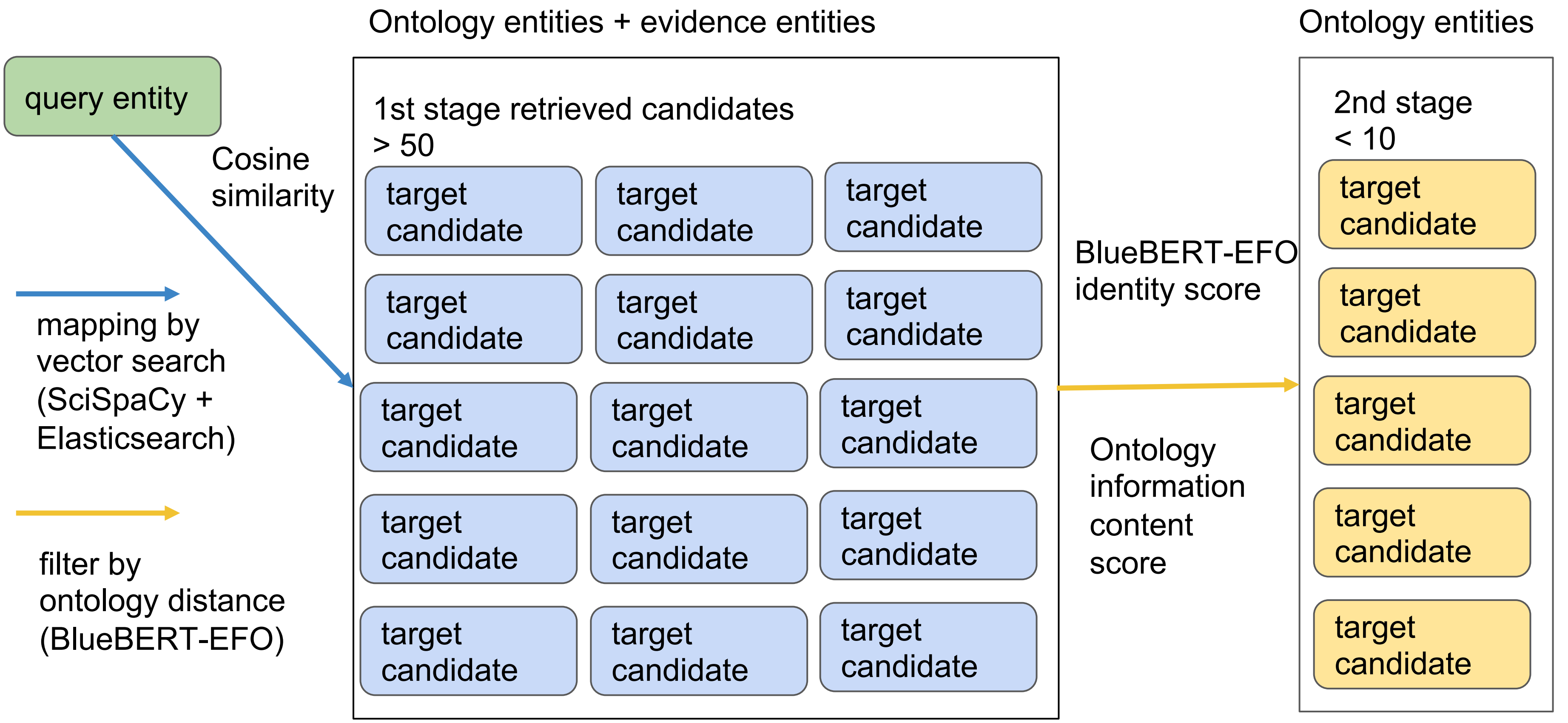
- Named entity recognition: SemRep (Kilicoglu et al., 2020 BMC Bioinformatics)
- Syntax: **(Subject) - [PREDICATE] - (Object)**
  - Glucose TREATS Diabetes
  - Obesity CAUSES Asthma
- Subjects / objects:
  UMLS Metathesaurus terms
- Predicate:
  UMLS Semantic network relationships

Scientific text: "Obesity subj: Obesity causes substantial changes to the mechanics of the lungs and chest wall, and these mechanical changes cause pred: CAUSES asthma obj: Asthma and asthma-like symptoms such as dyspnea, wheeze, and airway hyperresponsiveness"

claim triple

Obesity CAUSES Asthma

Entity harmonization



Evidence entities and groups

From claim triple to triangulatable evidence

Triple and literature evidence group

- Semantic SemMedDB triples derived from literature
- Source literature
- EpiGraphDB entities:
    - (LiteratureTerm)
    - (LiteratureTriple)
    - (Literature)

Association evidence group

- Systematic statistical analysis results
- EpiGraphDB entities
    - (Gwas) (OpenGWAS)
    - [MR_EVE_MR] (Hemani et al)
    - [PRS] (Richardson et al)
    - [GEN_COR] (Neale Lab)
- Common properties: beta, se, p-val

Evidence types *w.r.t* the claim

- **Supporting evidence**: *sufficiently* supports the claim
- **Reversal evidence**: *sufficiently* contradicts the claim from reversal direction
- **Insufficient evidence**: scope of evidence identification
- **Additional evidence**: additional information for expert knowledge

	Supporting	Reversal	Insufficient	Additional
Directional predicates `CAUSES`, `TREATS`, `PRODUCES`, `AFFECTS`				
Triple and literature	$S - P \rightarrow O$	$O - P \rightarrow S$	N/A	N/A
Association	$S - P \rightarrow O$, $P_P - Value < \pi$	$O - P \rightarrow S$, $P_P - Value < \pi$	$S - P \rightarrow O$, $P_P - Value \geq \pi$	non-directional $S - P - O$
Non-directional predicates `INTERACTS_WITH`, `COEXISTS_WITH`, `ASSOCIATED_WITH`				
Triple and literature	$S - P - O$	N/A	N/A	N/A
Association	$S - P - O$, $P_P - Value < \pi$	N/A	$S - P - O$, $P_P - Value \geq \pi$	N/A

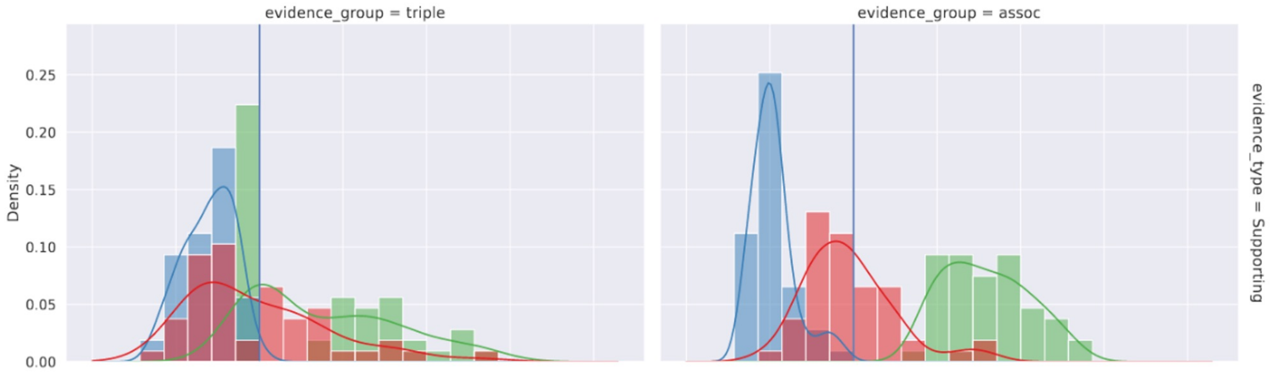
Evidence strength and prioritization

$$P_{\text{mapping}} = \prod_i \max_j \left(S_{\text{query} \to \text{EFO}_j} \times S_{\text{EFO}_j \to \text{evidence}} \right),$$

$$i \in [\text{subject}, \text{object}]$$

$$P_{\text{T\&L.}} = 1 + log_{10} N_{\text{literature}} \qquad P_{\text{Assoc.}} = \max\left(0, 1 + \log_{10}\left|\frac{\beta}{\sigma}\right|\right)$$

$$E_{\text{T\&L.}} = P_{\text{mapping}} \times P_{\text{T\&L.}} \qquad E_{\text{Assoc.}} = P_{\text{mapping}} \times P_{\text{Assoc.}}$$



Strength of an individual evidence to the claim

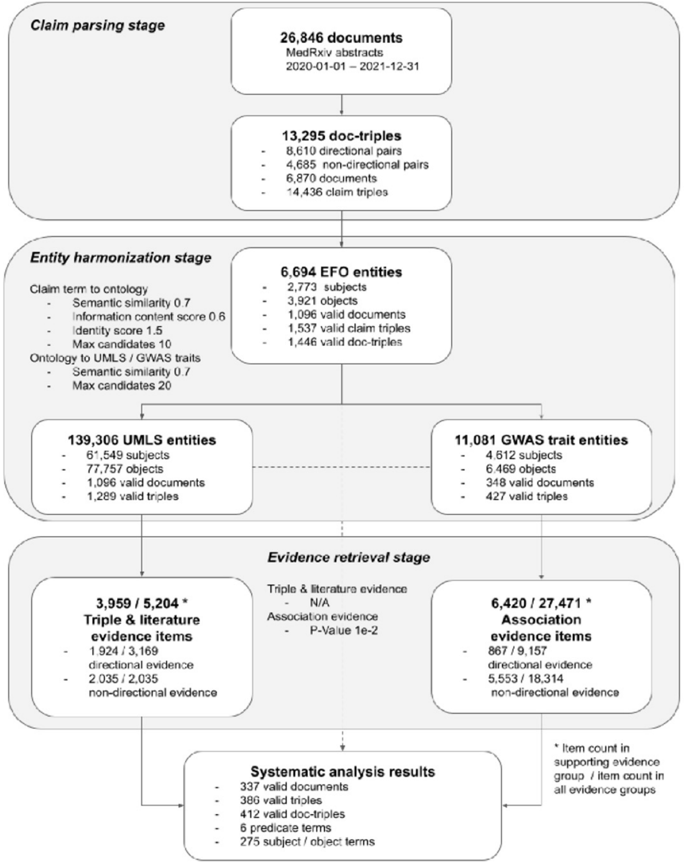
- Semantic similarity of the evidence entities to claim entities
- Strength of the evidence per se

Aggregated into the strength of an evidence group to compare with other evidence groups

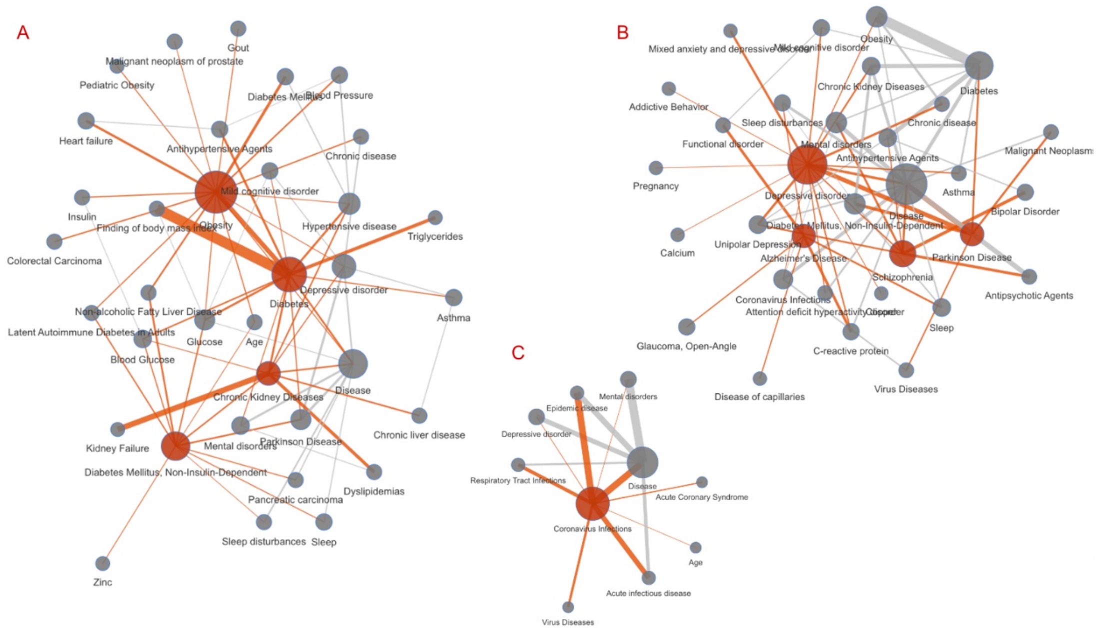
Metrics should NOT replace in depth investigations

Systematic analysis

- We parsed abstracts of medRxiv submissions from 2020 - 2021
- Automated using the batch-processing capability of ASQ
- Available https://asq.epigraphdb.org/medrxiv-analysis



Systematic analysis, contd



Claim term	Supporting T&L. + Assoc.	T&L.	Assoc.	Any	Init.
Disease	41	74	44	77	715
Obesity	20	25	25	30	125
Diabetes	17	19	18	20	87
Depressive disorder	14	20	16	26	100
Parkinson Disease	13	13	13	13	111
Diabetes Mellitus, Non-Insulin-Dependent	10	12	12	15	84
Alzheimer's Disease	8	10	8	10	111
Schizophrenia	8	11	8	11	32
C-reactive protein	7	7	9	10	24
Malignant Neoplasms	7	8	15	19	100
Chronic Kidney Diseases	6	9	6	9	35
Chronic disease	5	6	5	6	44
Fatigue	5	5	6	6	25
Sleep	5	5	6	6	21
Atrial Fibrillation	5	6	6	9	57
Pain	4	4	6	6	30
Glucose	4	5	4	6	20
Blood Glucose	4	5	4	5	15
Hypertensive disease	4	12	4	13	90
Mental disorders	4	8	4	10	42
Cardioembolic stroke	3	3	3	3	14
Testosterone	3	5	3	6	21
Diabetes Mellitus	3	4	5	6	25

Thank you for listening. Questions & comments welcome.

DMER programme and EpiGraphDB working group

- Tom Gaunt
- Benjamin Elsworth
- Pau Erola
- Valeriia Haberland
- Jie Zheng
- Marina Vabistsevits
- Oliver Lloyd

- DMER programme https://biocompute.org.uk
- EpiGraphDB platform https://epigraphdb.org
- EpiGraphDB-ASQ https://asq.epigraphdb.org