Triangulating evidence in health sciences with Annotated Semantic Queries

IEU Monthly meeting 2022-04-06

Yi Liu

Outline

- High-level overview
  - Background
  - ASQ as a natural language interface
  - Demo
- Systematic analysis
  - Processing results
  - General findings
- Method details
  - Entity harmonization
  - Evidence retrieval
  - Ranking evidence

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

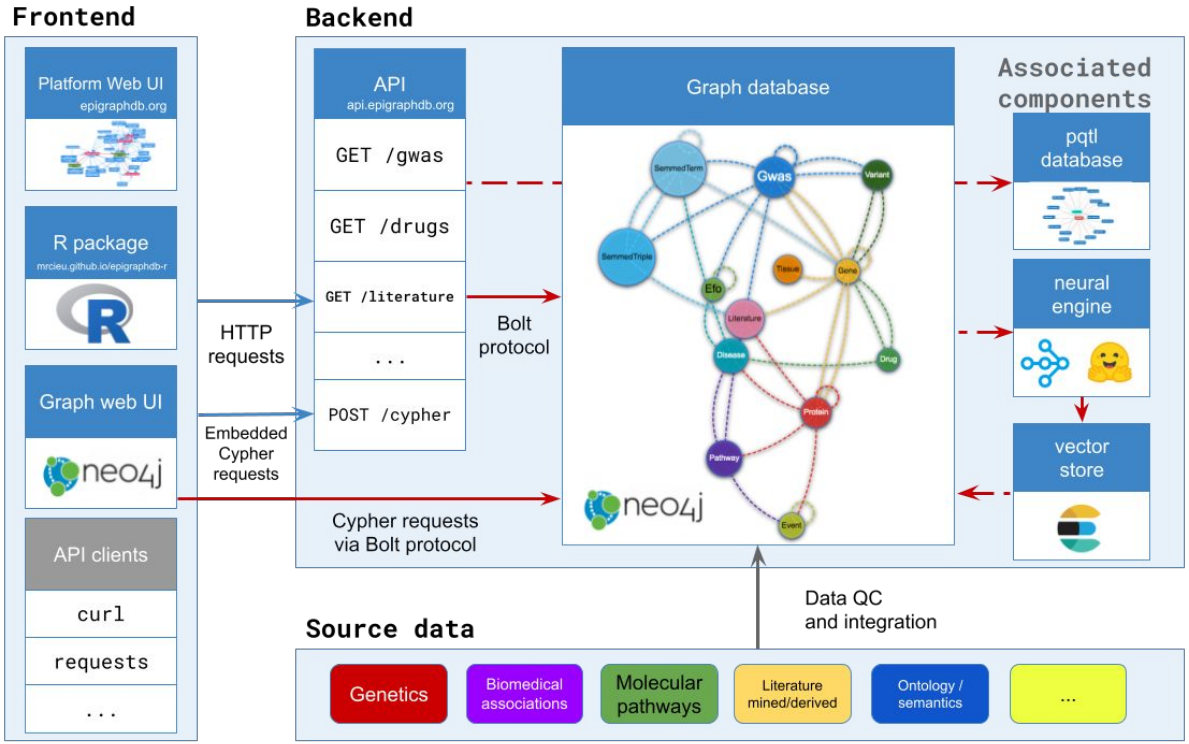
Integrating information from data sources representing different study designs has the potential to strengthen evidence in population health research. However, this concept of evidence "triangulation" presents a number of challenges for systematically identifying and integrating relevant information. 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.

Compiled at 2022-04-04 20:51.

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[4,5], 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 in-

Background: the EpiGraphDB data



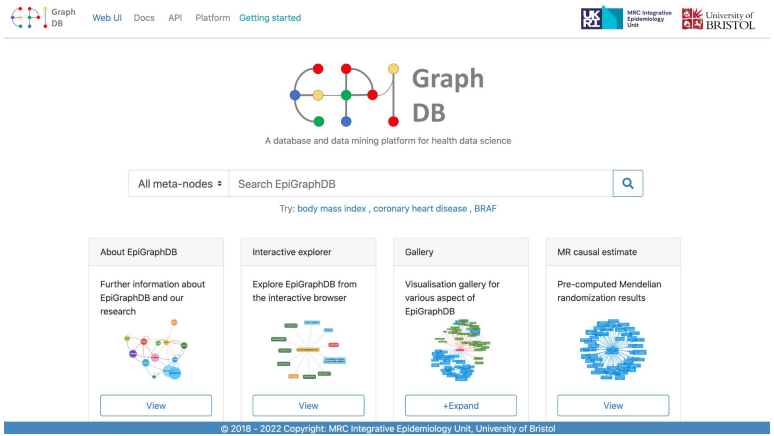
- GWAS traits and systematic MR
- Observational and genetic correlations
- Literature mined relationships
- Molecular pathways
- Protein-protein interactions
- Drug target relationships

Background: How to query EpiGraphDB

- Via the various topic-specific pages: https://epigraphdb.org

- Via programmatic access:
  - API: https://api.epigraphdb.org
  - R: https://mrcieu.github.io/epigraphdb-r
- Neo4j Cypher
  `MATCH (n:Gwas) RETURN n LIMIT 2`

- Docs: https://docs.epigraphdb.org
- Feedback welcome!



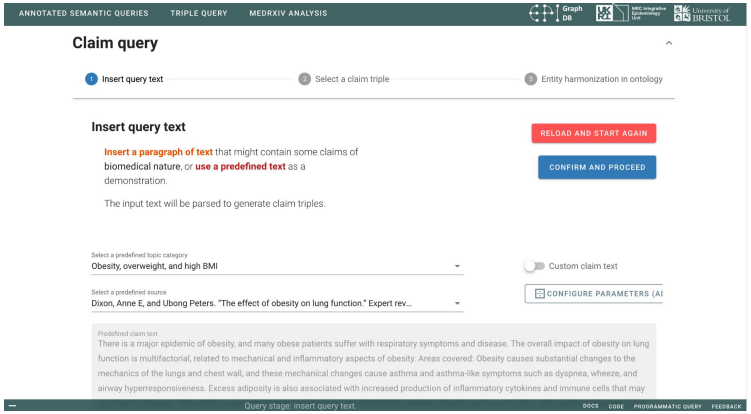
epigraphdb: Interface Package for the 'EpiGraphDB' Platform

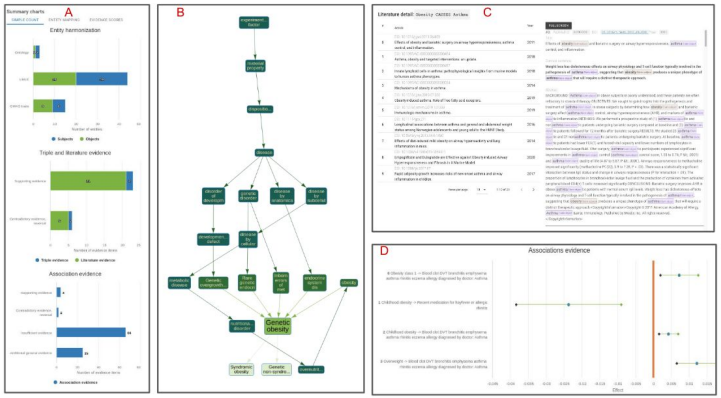
The interface package to access data from the 'EpiGraphDB' <https://epigraphdb.org> platform. It provides ea API <https://api.epigraphdb.org> and return the response data in the 'tibble' data frame format.

Version:	0.2.3
Imports:	magrittr, tibble, httr, glue, purrr, jsonlite
Suggests:	testthat, roxygen2, knitr, rmarkdown, spelling, devtools, usethis, pkgdown, styler, lintr, co
Published:	2022-01-14
Author:	Yi Liu [cre, aut], Valeriia Haberland [aut], Marina Vabistsevits [aut], Tom Gaunt [aut], M
Maintainer:	Yi Liu <yi6240.liu at bristol.ac.uk>
BugReports:	https://github.com/MRCIEU/epigraphdb-r/issues
License:	GPL-3
URL:	https://mrcieu.github.io/epigraphdb-r/
NeedsCompilation:	no
Language:	en-GB
Materials:	README NEWS
CRAN checks:	epigraphdb results

Asq, and you ~~shall~~ might get answers from EpiGraphDB

- https://asq.epigraphdb.org
- Improve the accessibility and introspectability of evidence triangulation
- Input a short piece of text involving scientific findings
  - e.g. Abstract
- ASQ extracts claims and retrieves EpiGraphDB evidence regarding a specific claim
- Assist expert knowledge



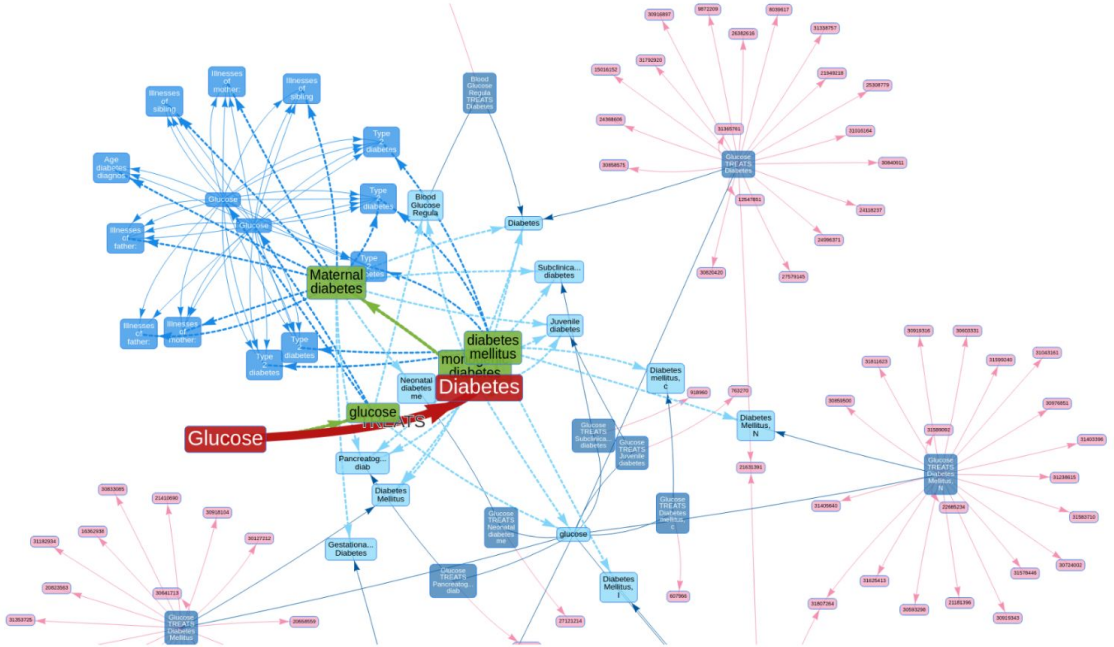


Claim triple: from free from text to structured entities

- Syntax: **(Subject) - [PREDICATE] - (Object)**
  - Glucose TREATS Diabetes
  - Obesity CAUSES Asthma

- UMLS Metathesaurus terms
- Semantic network relationships

- Retrieve entities from various EpiGraphDB taxonomies



From claim triple to triangulatable evidence

Triple and literature evidence

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

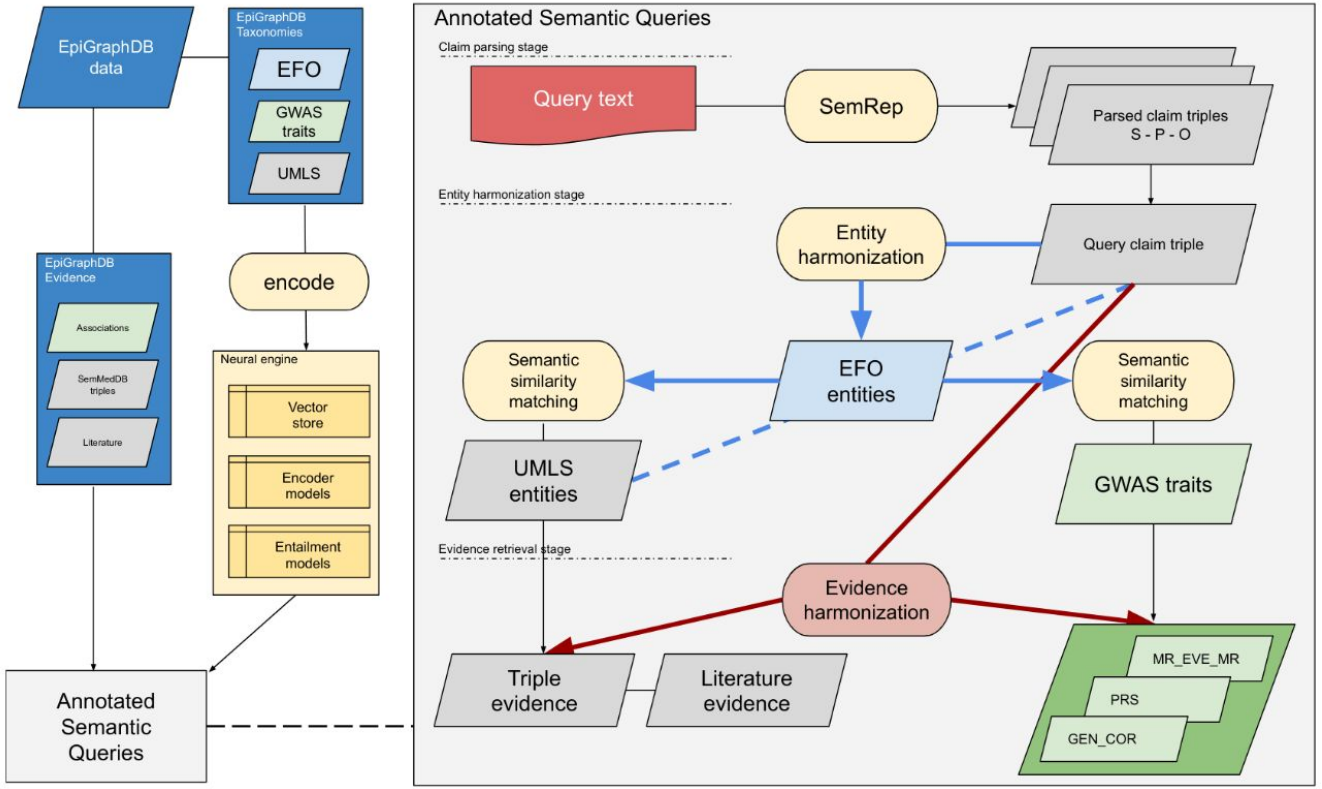
Association evidence

- 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

From claim triple to triangulatable evidence, contd

- **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

ASQ: architecture

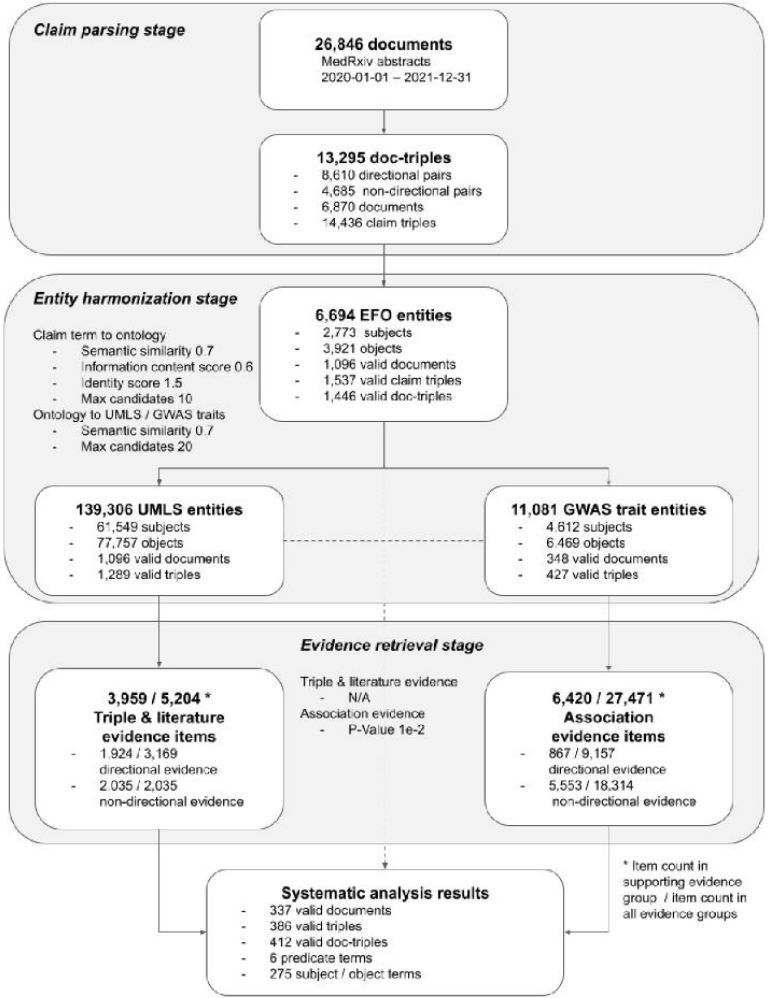


Demo

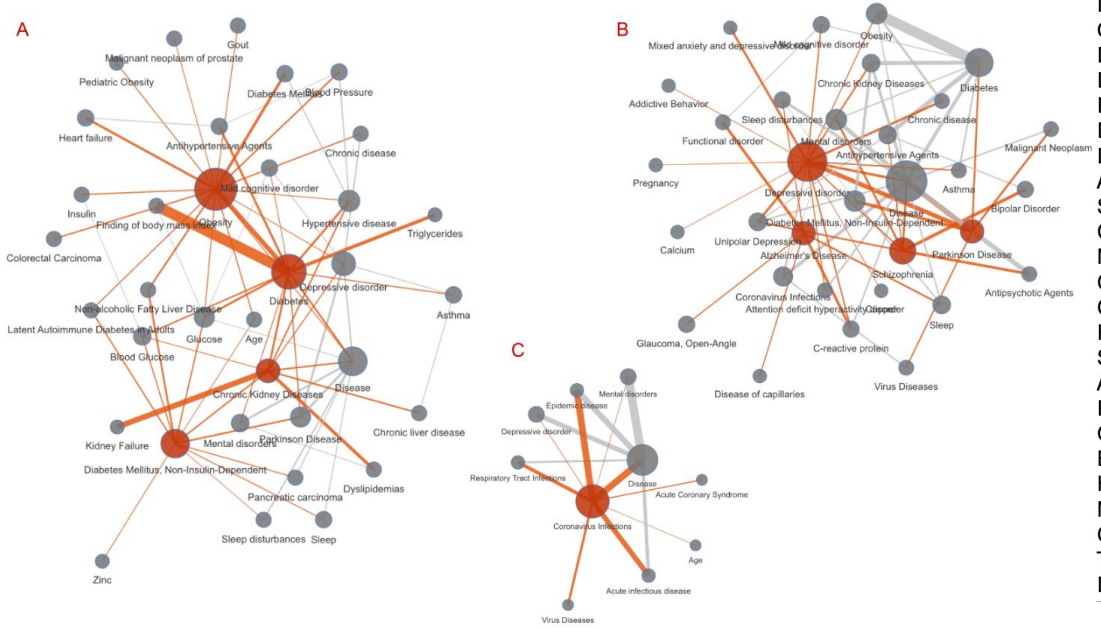
- Main entrypoint, query by text: https://asq.epigraphdb.org
- Query by triple: https://asq.epigraphdb.org/triple
- Systematic analysis results: https://asq.epigraphdb.org/medrxiv-analysis

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			Any	Init.
	T&L. + Assoc.	T&L.	Assoc.		
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

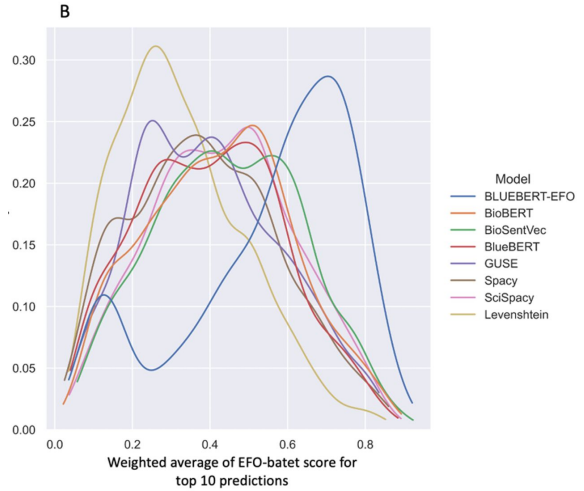
Method details: Entity harmonization

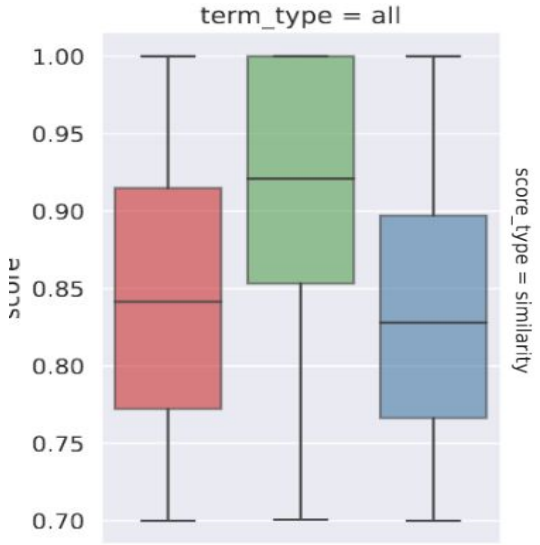
Entity representation

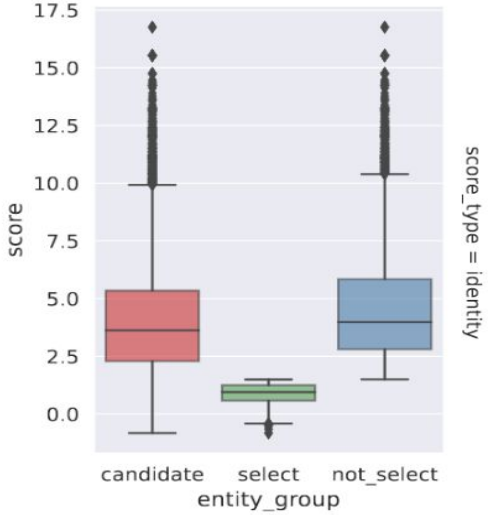
- text vector embeddings
- cosine similarity

Ontology as the anchor

- Identity score
- Information content score
  - “EFO”: 0
  - “disease”: 0.23
  - “metabolic disease”: 0.5
  - “obesity”: 0.8







Method details: Evidence types

	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

Methods details: Evidence ranking

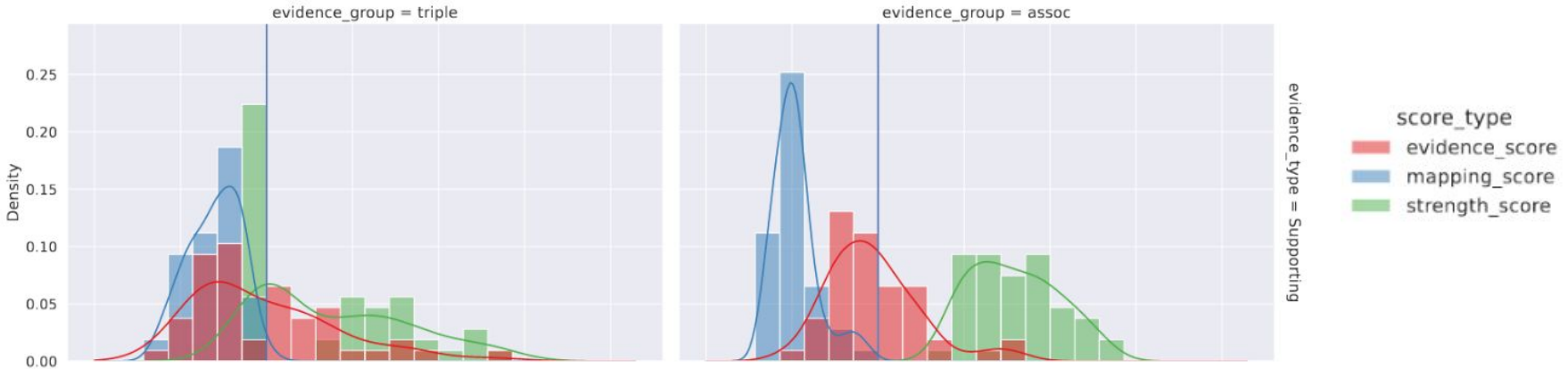
$$P_{\text{mapping}} = \prod_i \max_j \left(S_{\text{query}\rightarrow\text{EFO}_j} \times S_{\text{EFO}_j\rightarrow\text{evidence}} \right), i \in [\text{subject}, \text{object}]$$

$$P_{\text{T\&L.}} = 1 + log_{10} N_{\text{literature}}$$

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

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

$$E_{\text{Assoc.}} = P_{\text{mapping}} \times P_{\text{Assoc.}}$$



(Fin.) Entry points

- Main entrypoint, query by text: https://asq.epigraphdb.org
- Query by triple:
- Systematic analysis results: https://asq.epigraphdb.org/medrxiv-analysis
- Programmatic access:
  - API: https://asq-api.epigraphdb.org
  - Tutorial: forthcoming

Thank you
for listening.
Questions &
comments welcome.