MRC Integrative Epidemiology Unit

University of BRISTOL

Annotated Semantic Queries, and beyond

Yi Liu, MRC Integrative Epidemiology Unit

JGI Health Data Research Network Worshop
10 April 2025

$(whoami)

Yi Liu

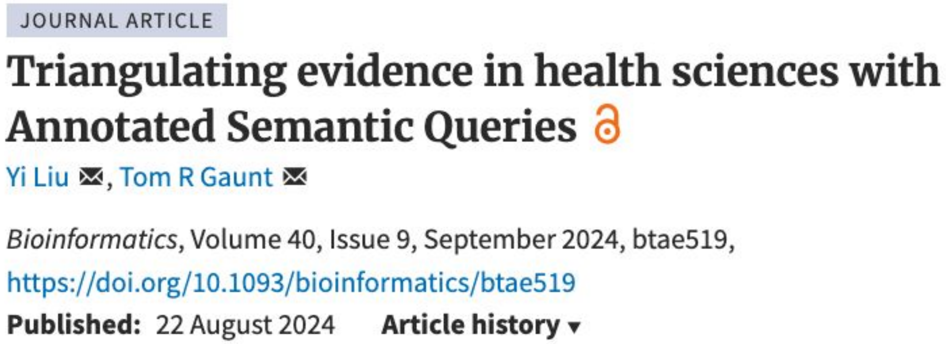
Research Fellow, MRC IEU Programme 3 (PI: Tom Gaunt) on data mining epidemiological relationships

I lead projects on the method development and application of computational approaches (data infrastructure, machine learning) for data mining.

- https://yiliu6240.github.io/
- https://mrcieu.github.io/

Today's talk

Our research work on the *Annotated Semantic Queries* (Liu and Gaunt, 2024) data platform for automating evidence triangulation.



JOURNAL ARTICLE

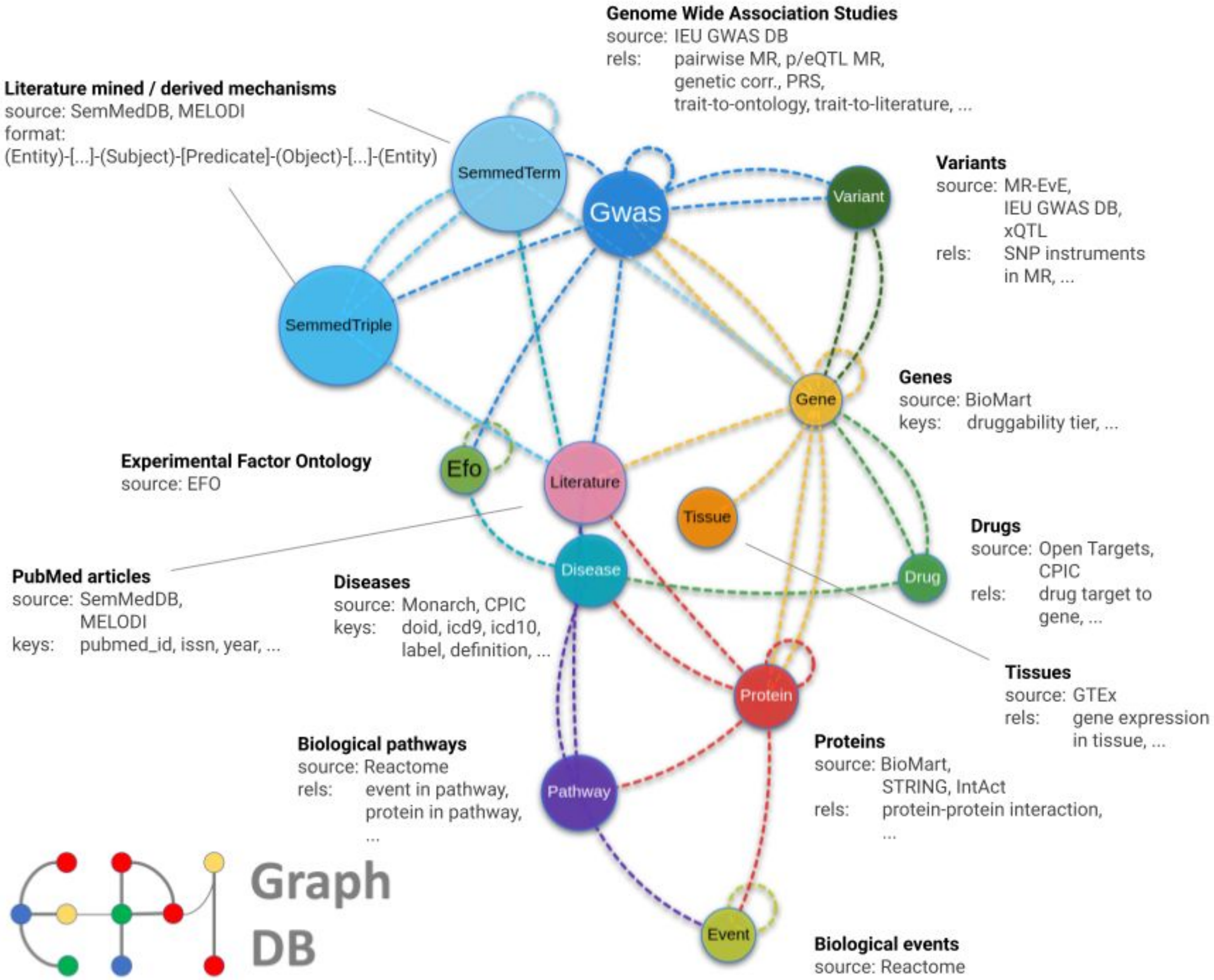
Triangulating evidence in health sciences with Annotated Semantic Queries

Yi Liu ✉, Tom R Gaunt ✉

Bioinformatics, Volume 40, Issue 9, September 2024, btae519, https://doi.org/10.1093/bioinformatics/btae519

Published: 22 August 2024 **Article history** ▾

The EpiGraphDB knowledge graph



Genome Wide Association Studies
source: IEU GWAS DB
rels: pairwise MR, p/eQTL MR, genetic corr., PRS, trait-to-ontology, trait-to-literature, ...

Literature mined / derived mechanisms
source: SemMedDB, MELODI
format:
(Entity)-[...]-(Subject)-[Predicate]-(Object)-[...]-(Entity)

Variants
source: MR-EvE, IEU GWAS DB, xQTL
rels: SNP instruments in MR, ...

Genes
source: BioMart
keys: druggability tier, ...

Experimental Factor Ontology
source: EFO

Drugs
source: Open Targets, CPIC
rels: drug target to gene, ...

PubMed articles
source: SemMedDB, MELODI
keys: pubmed_id, issn, year, ...

Diseases
source: Monarch, CPIC
keys: doid, icd9, icd10, label, definition, ...

Tissues
source: GTEx
rels: gene expression in tissue, ...

Biological pathways
source: Reactome
rels: event in pathway, protein in pathway, ...

Proteins
source: BioMart, STRING, IntAct
rels: protein-protein interaction, ...

Biological events
source: Reactome

Graph DB

Yi Liu, Benjamin Elsworth, Pau Erola, Valeriia Haberland, Gibran Hemani, Matt Lyon, Jie Zheng, Oliver Lloyd, Marina Vabistsevits, Tom R Gaunt, EpiGraphDB: a database and data mining platform for health data science, *Bioinformatics*, 2021.

- 58 citations (10 April 2025)
- Supported a few high-profile research works in IEU
- Foundation to our next works on NLP

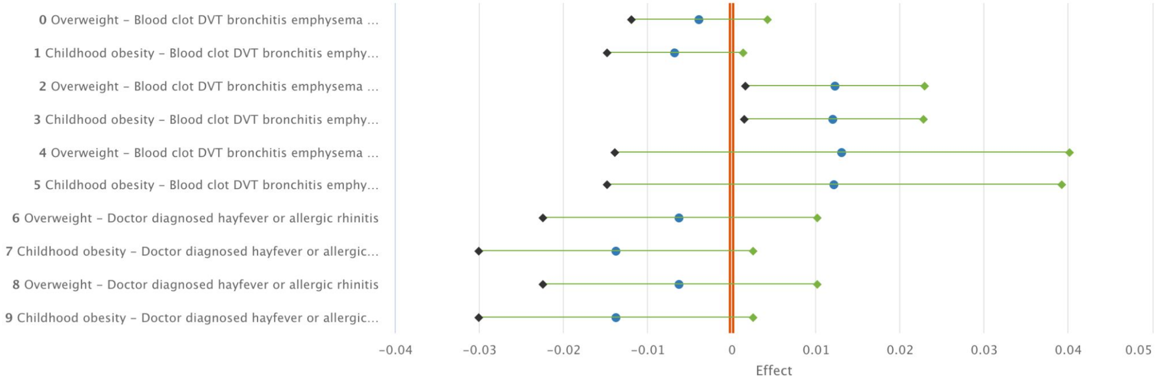
Curate and represent biomedical entities (as nodes) and epidemiological evidence (as edges) in a knowledge graph (KG) for data mining

Implemented using a Neo4j graph database and queried via Cypher

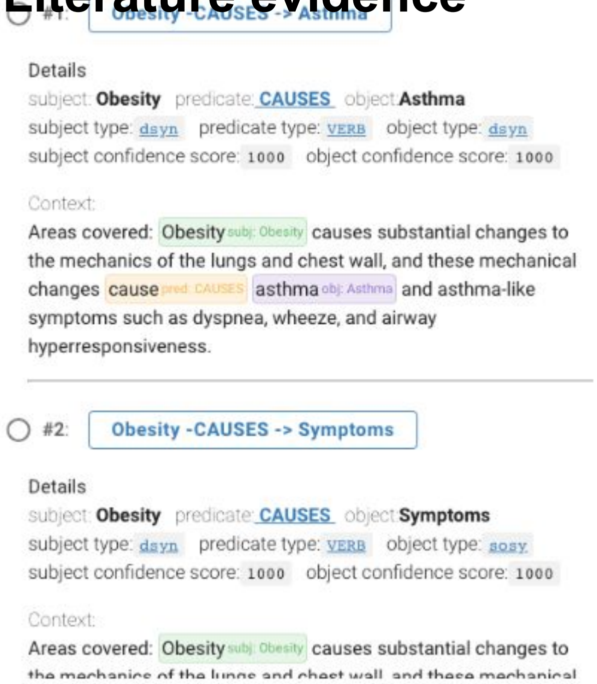
(Source) - [Predicate] -> (Target)

Knowledge, knowledge graph and data mining

Association evidence



Literature evidence



○ #1: Obesity -CAUSES -> Asthma

Details

subject: **Obesity** predicate: CAUSES object: **Asthma**

subject type: dsyn predicate type: VERB object type: dsyn

subject confidence score: 1000 object confidence score: 1000

Context:

Areas covered: 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.

○ #2: Obesity -CAUSES -> Symptoms

Details

subject: **Obesity** predicate: CAUSES object: **Symptoms**

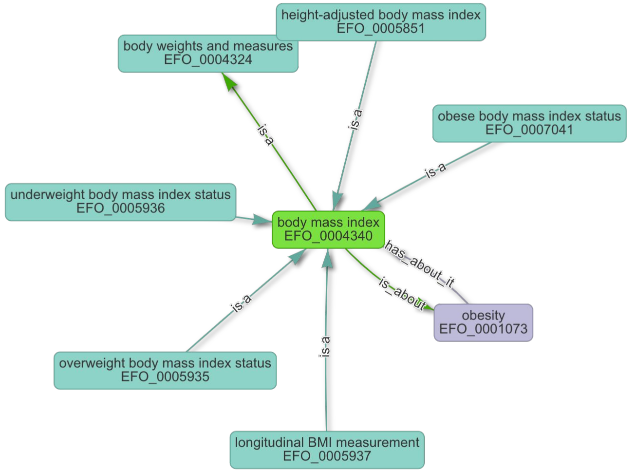
subject type: dsyn predicate type: VERB object type: sosy

subject confidence score: 1000 object confidence score: 1000

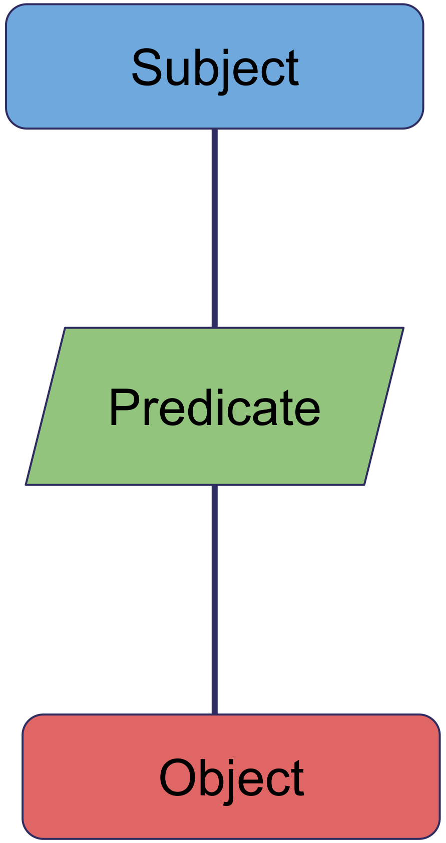
Context:

Areas covered: Obesity subj: Obesity causes substantial changes to the mechanics of the lungs and chest wall, and these mechanical

Ontology evidence


height-adjusted body mass index
EFO_0005851
body weights and measures
EFO_0004324
obese body mass index status
EFO_0007041
underweight body mass index status
EFO_0005936
body mass index
EFO_0004340
obesity
EFO_0001073
overweight body mass index status
EFO_0005935
longitudinal BMI measurement
EFO_0005937
is-a
is-a
is-a
is-a
is-a
has_about_it
is_about

From evidence to knowledge (**knowledge triple**)



Node properties

- phenotype group
- literature abstract

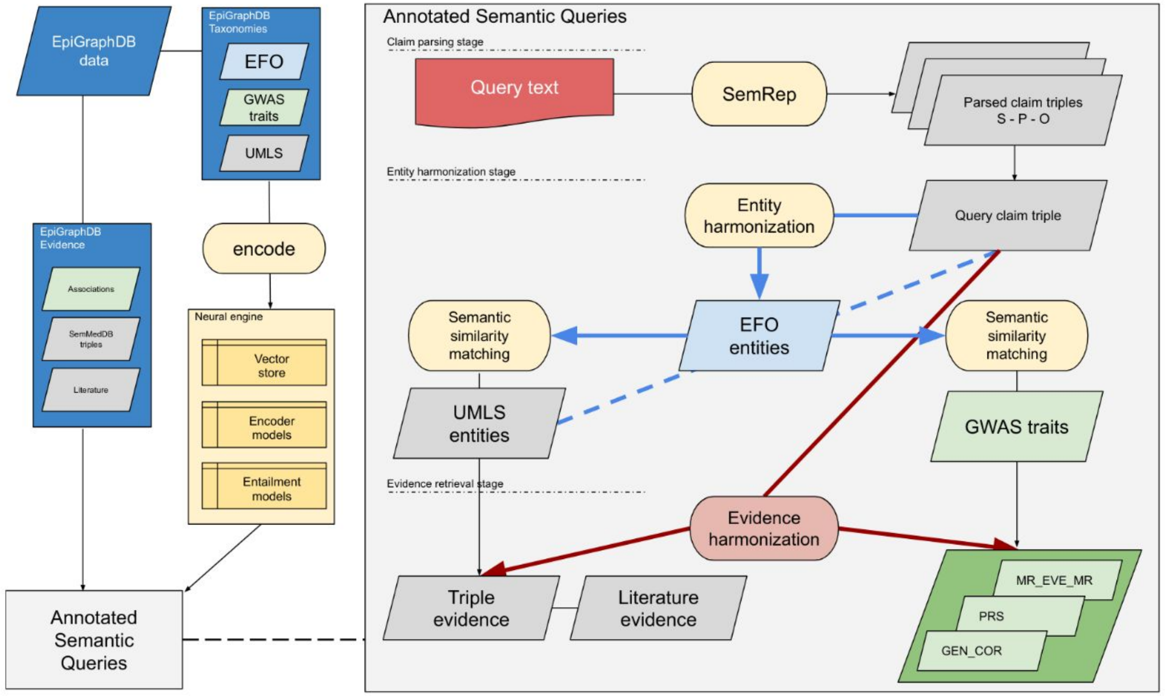
Edge properties

- Effect size
- Confidence score

Can we automate data mining in KG with natural language processing methods?

Annotated Semantic Queries

- User inputs a text involving scientific descriptions (e.g. paper abstract)
- Extract semantic triples from the text, and then for a triple of interest,
  - Perform evidence harmonization for candidate data in EpiGraphDB with the triple
  - Perform evidence prioritization with the triple

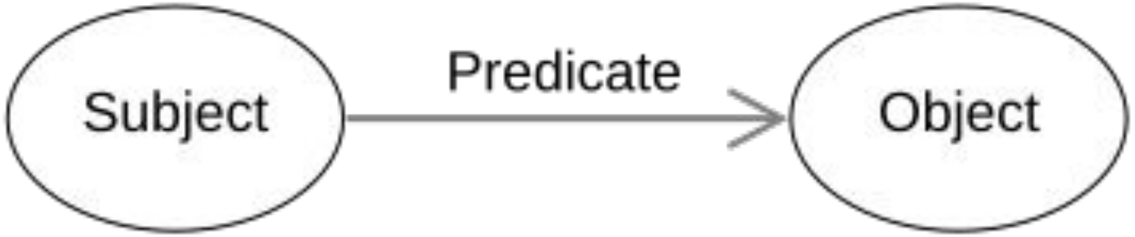


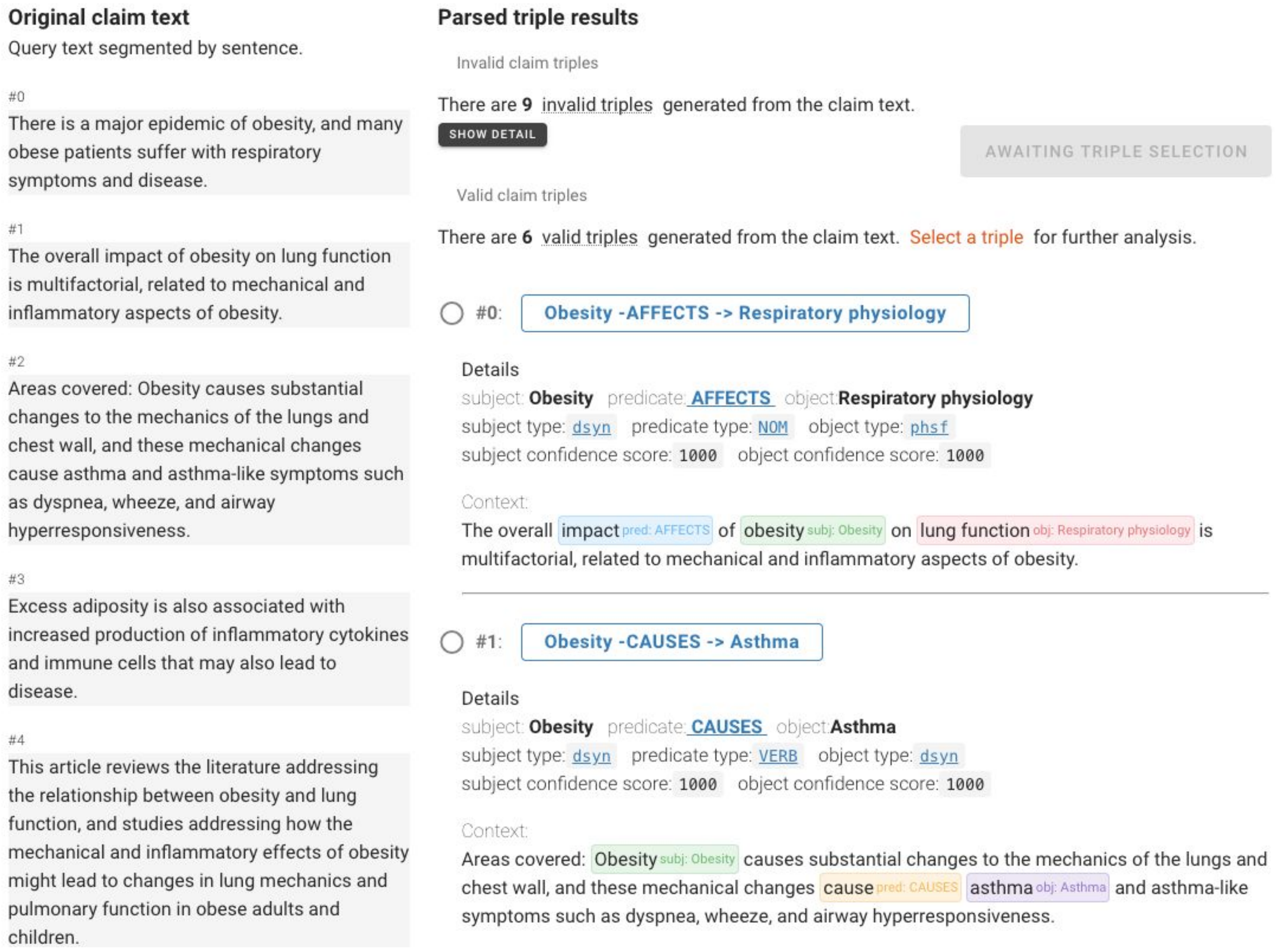
https://asq.epigraphdb.org/docs

Evidence triples

In ASQ we treat a piece of “evidence” as a semantic triple, involving

- The semantic triple
- The evidence source
    - Literature
    - Statistical associations
- The quantifiable information, on evidence strength





Original claim text

Query text segmented by sentence.

#0

There is a major epidemic of obesity, and many obese patients suffer with respiratory symptoms and disease.

#1

The overall impact of obesity on lung function is multifactorial, related to mechanical and inflammatory aspects of obesity.

#2

Areas covered: Obesity causes substantial changes to the mechanics of the lungs and chest wall, and these mechanical changes cause asthma and asthma-like symptoms such as dyspnea, wheeze, and airway hyperresponsiveness.

#3

Excess adiposity is also associated with increased production of inflammatory cytokines and immune cells that may also lead to disease.

#4

This article reviews the literature addressing the relationship between obesity and lung function, and studies addressing how the mechanical and inflammatory effects of obesity might lead to changes in lung mechanics and pulmonary function in obese adults and children.

Parsed triple results

Invalid claim triples

There are **9** invalid triples generated from the claim text.

SHOW DETAIL

AWAITING TRIPLE SELECTION

Valid claim triples

There are **6** valid triples generated from the claim text. Select a triple for further analysis.

○ #0: Obesity -AFFECTS -> Respiratory physiology

Details

subject: **Obesity** predicate: AFFECTS object: **Respiratory physiology**

subject type: `dsyn` predicate type: `NOM` object type: `phsf`

subject confidence score: `1000` object confidence score: `1000`

Context:

The overall impact pred: AFFECTS of obesity subj: Obesity on lung function obj: Respiratory physiology is multifactorial, related to mechanical and inflammatory aspects of obesity.

○ #1: Obesity -CAUSES -> Asthma

Details

subject: **Obesity** predicate: CAUSES object: **Asthma**

subject type: `dsyn` predicate type: `VERB` object type: `dsyn`

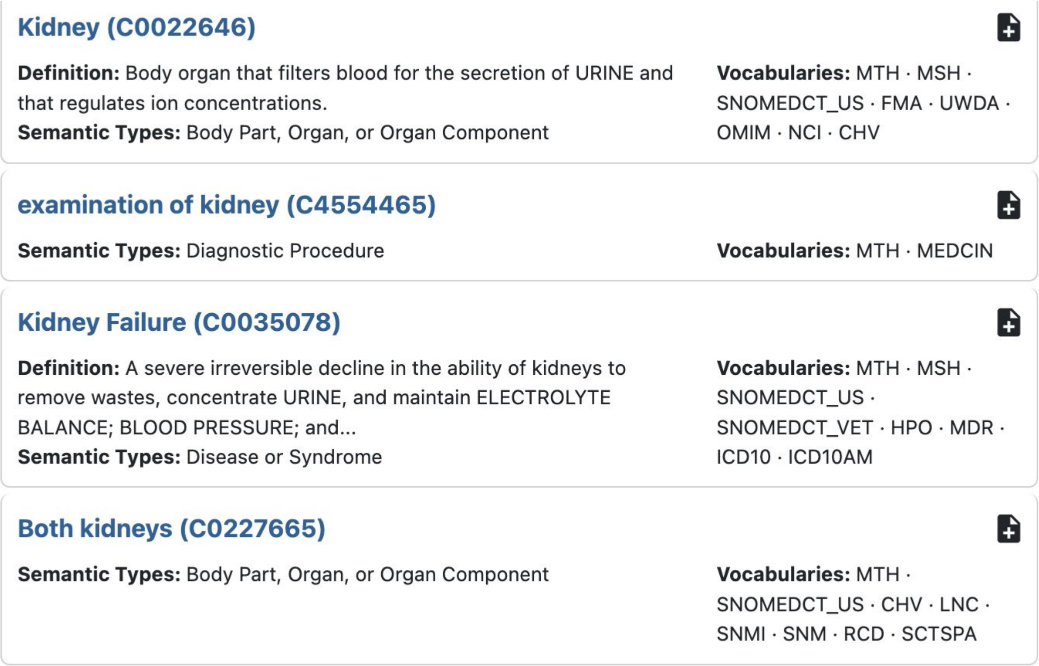
subject confidence score: `1000` object confidence score: `1000`

Context:

Areas covered: 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.

Triple extraction

- Perform by SemRep
- Entities (subjects, objects): UMLS Metathesaurus terms



Kidney (C0022646)

Definition: Body organ that filters blood for the secretion of URINE and that regulates ion concentrations.
Semantic Types: Body Part, Organ, or Organ Component

Vocabularies: MTH · MSH · SNOMEDCT_US · FMA · UWDA · OMIM · NCI · CHV

examination of kidney (C4554465)

Semantic Types: Diagnostic Procedure

Vocabularies: MTH · MEDCIN

Kidney Failure (C0035078)

Definition: A severe irreversible decline in the ability of kidneys to remove wastes, concentrate URINE, and maintain ELECTROLYTE BALANCE; BLOOD PRESSURE; and...
Semantic Types: Disease or Syndrome

Vocabularies: MTH · MSH · SNOMEDCT_US · SNOMEDCT_VET · HPO · MDR · ICD10 · ICD10AM

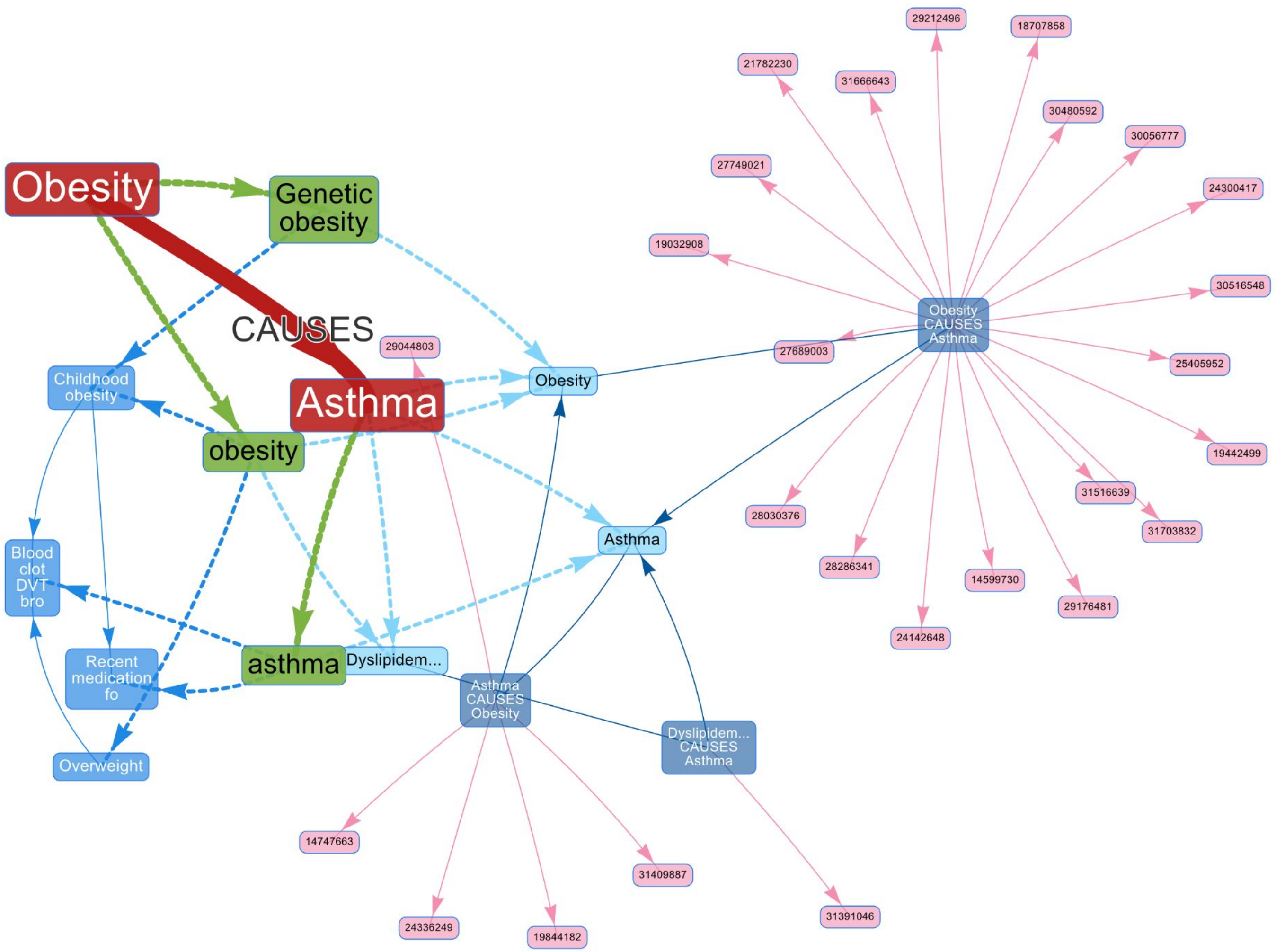
Both kidneys (C0227665)

Semantic Types: Body Part, Organ, or Organ Component

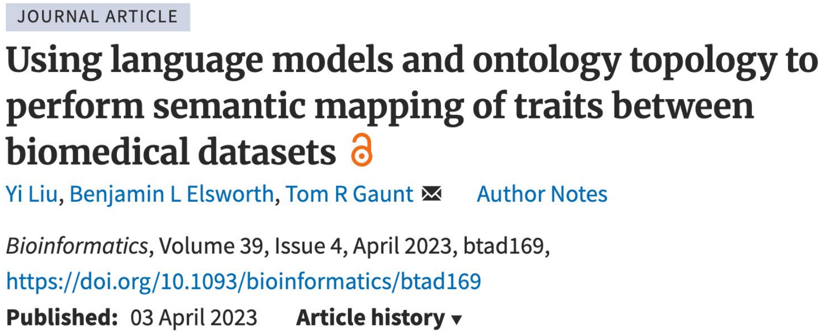
Vocabularies: MTH · SNOMEDCT_US · CHV · LNC · SNMI · SNM · RCD · SCTSPA

- Relationships: UMLS semantic relationships
- Directional relationships:
  - CAUSES, TREATS, PRODUCES, AFFECTS
- Nondirectional relationships:
  - INTERACTS_WITH, COEXISTS_WITH, ASSOCIATED_WITH

Evidence harmonization: entities



Mapping of entities done via semantic representation of the labels (Liu, Elsworth, Gaunt, 2023) via finetuned BlueBERT LLM



JOURNAL ARTICLE

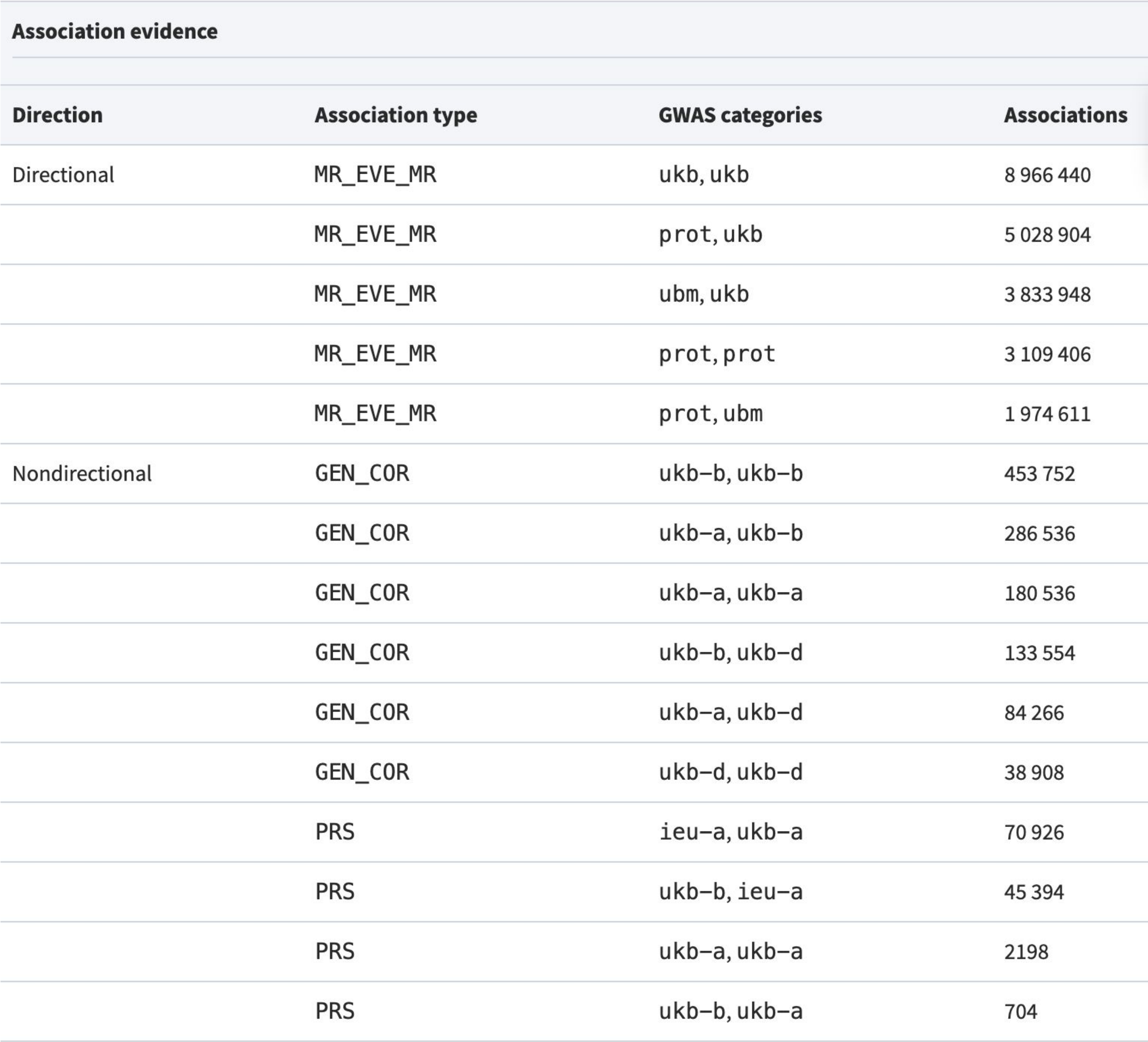
Using language models and ontology topology to perform semantic mapping of traits between biomedical datasets

Yi Liu, Benjamin L Elsworth, Tom R Gaunt ✉ Author Notes

Bioinformatics, Volume 39, Issue 4, April 2023, btad169, https://doi.org/10.1093/bioinformatics/btad169

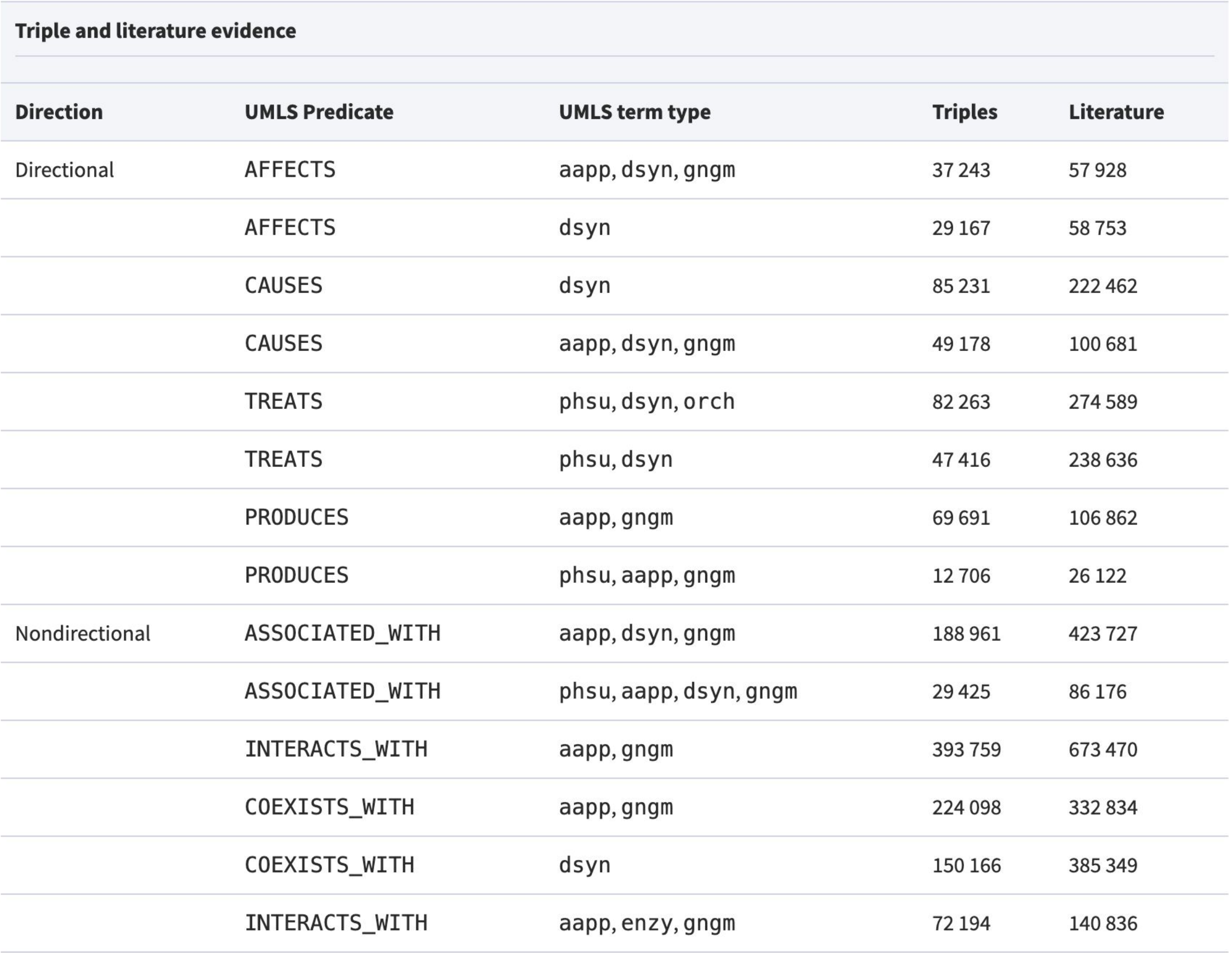
Published: 03 April 2023 **Article history** ▾

Harmonized EpiGraphDB evidence



Association evidence

Direction	Association type	GWAS categories	Associations
Directional	MR_EVE_MR	ukb, ukb	8 966 440
	MR_EVE_MR	prot, ukb	5 028 904
	MR_EVE_MR	ubm, ukb	3 833 948
	MR_EVE_MR	prot, prot	3 109 406
	MR_EVE_MR	prot, ubm	1 974 611
Nondirectional	GEN_COR	ukb–b, ukb–b	453 752
	GEN_COR	ukb–a, ukb–b	286 536
	GEN_COR	ukb–a, ukb–a	180 536
	GEN_COR	ukb–b, ukb–d	133 554
	GEN_COR	ukb–a, ukb–d	84 266
	GEN_COR	ukb–d, ukb–d	38 908
	PRS	ieu–a, ukb–a	70 926
	PRS	ukb–b, ieu–a	45 394
	PRS	ukb–a, ukb–a	2198
	PRS	ukb–b, ukb–a	704



Triple and literature evidence

Direction	UMLS Predicate	UMLS term type	Triples	Literature
Directional	AFFECTS	aapp, dsyn, gngm	37 243	57 928
	AFFECTS	dsyn	29 167	58 753
	CAUSES	dsyn	85 231	222 462
	CAUSES	aapp, dsyn, gngm	49 178	100 681
	TREATS	phsu, dsyn, orch	82 263	274 589
	TREATS	phsu, dsyn	47 416	238 636
	PRODUCES	aapp, gngm	69 691	106 862
	PRODUCES	phsu, aapp, gngm	12 706	26 122
Nondirectional	ASSOCIATED_WITH	aapp, dsyn, gngm	188 961	423 727
	ASSOCIATED_WITH	phsu, aapp, dsyn, gngm	29 425	86 176
	INTERACTS_WITH	aapp, gngm	393 759	673 470
	COEXISTS_WITH	aapp, gngm	224 098	332 834
	COEXISTS_WITH	dsyn	150 166	385 349
	INTERACTS_WITH	aapp, enzy, gngm	72 194	140 836

Evidence prioritization

Strength of a piece of evidence w.r.t the question of interest consists of two components:

- The relevancy of the evidence
  - Semantic affinity of the subjects and objects
- The strength of the evidence itself
  - literature: number of lit triples occurred in the literature
  - assoc: normalised effect size

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

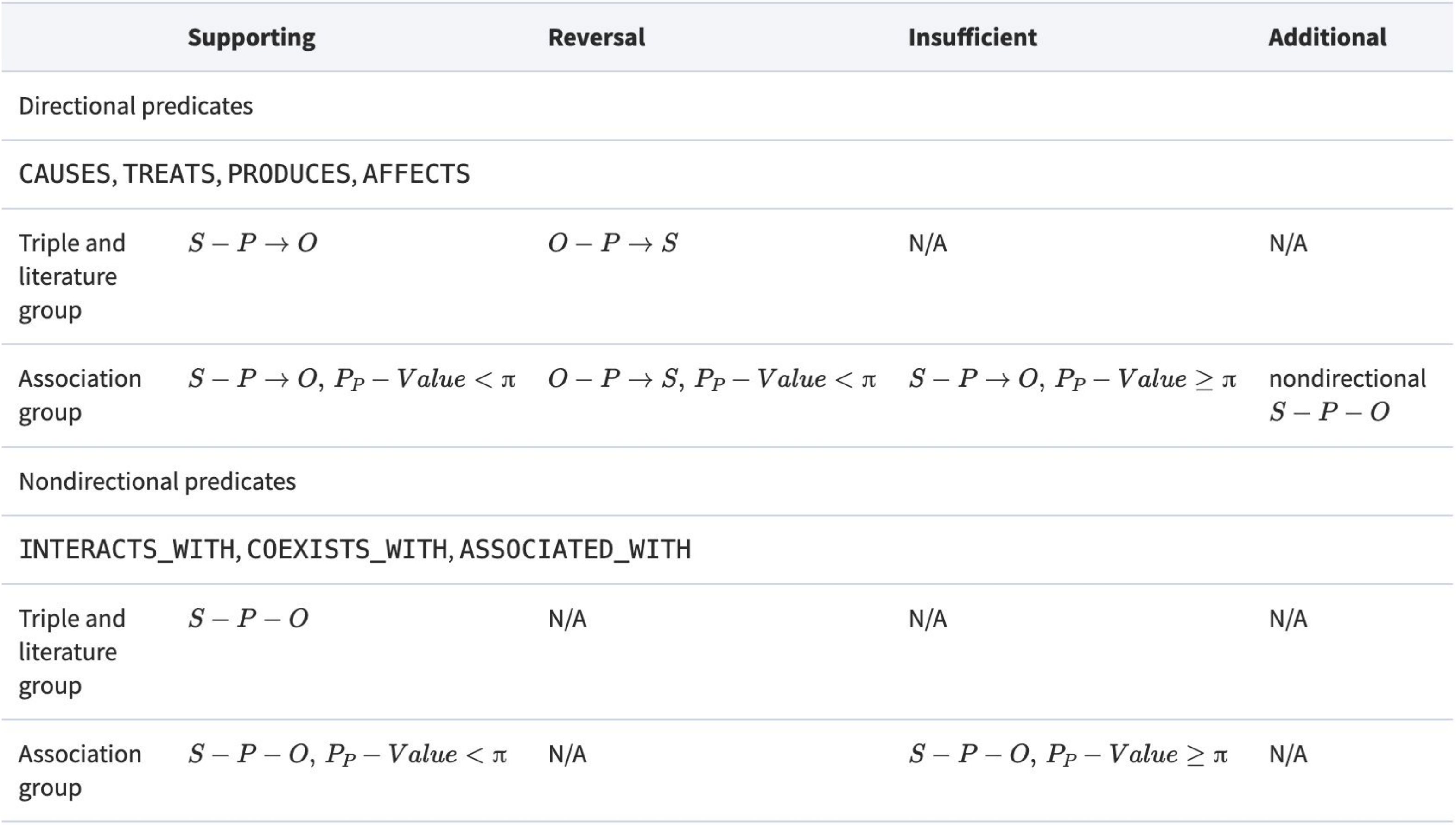
$$P_{\text{T\&L.}} = 1 + log_{10} N_{\text{literature}}$$

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

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

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

Evidence harmonization: relationships



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

ASQ: Fact checking scientific claims

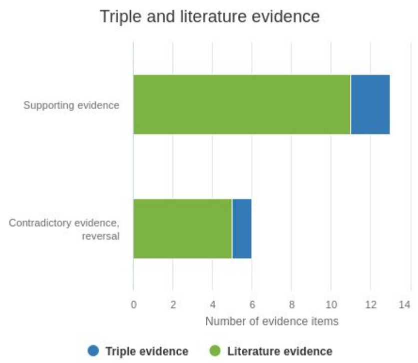
Scientific text: “Obesity subj: Obesity causes substantial changes to the mechanics of the lungs and chest wall, and these mechanical changes cause pred: CAUSES asthma obj: Asthma and asthma-like symptoms such as dyspnea, wheeze, and airway hyperresponsiveness”

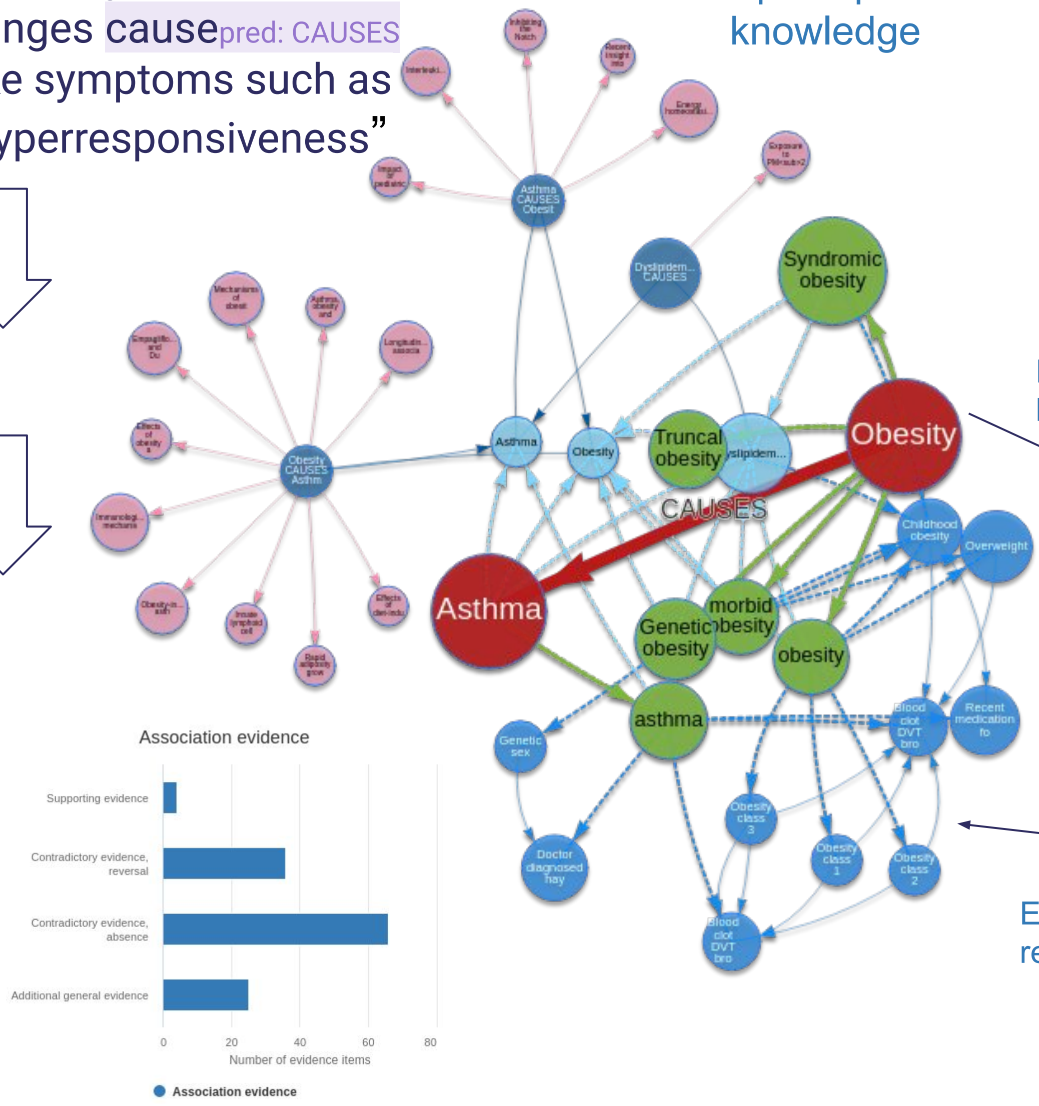
claim triple

Obesity CAUSES Asthma

EpiGraphDB

- Biomedical entities
- Supporting / contradictory evidence

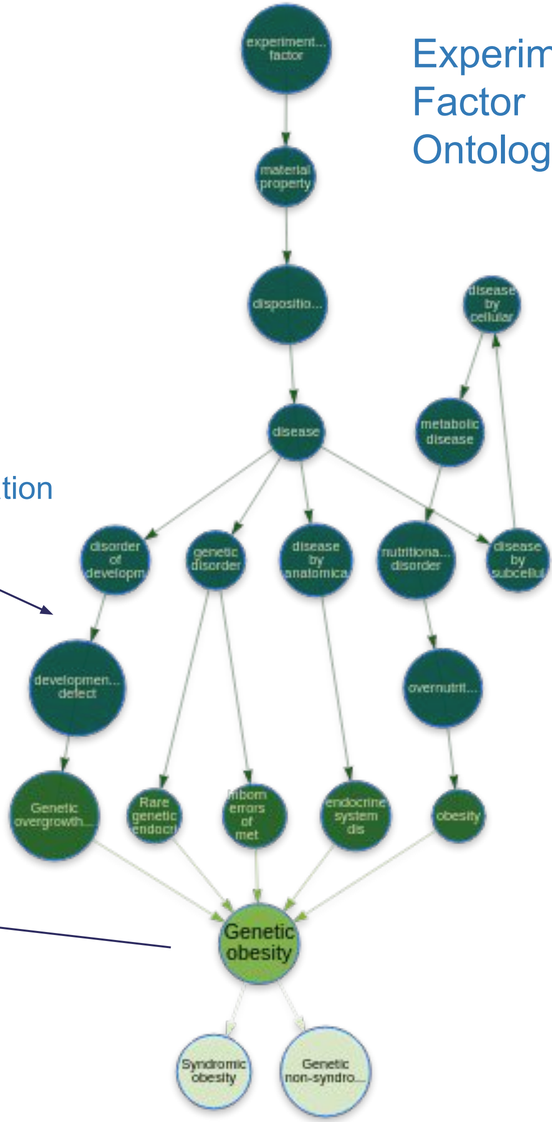




EpiGraphDB knowledge

Entity harmonization

Evidence retrieval



Experimental Factor Ontology

elasticsearch

Next step research

- Here I described the ASQ work as a query interface to EpiGraphDB data
- There are lots of areas that can be extended from this proof of concept, e.g. further integration with more data types and their harmonisation
- In addition, the LLMs have substantially evolved since our work
- So, what’s next

Next step:

A Roche-UoB partnership project on Assertion Recognition in Biomedical Literature

Roche funded PhD studentship (2025 - 2029), supervised by

- Yi Liu, Tom Gaunt
- Zahraa Abdallah
- Michael Tillich, Martin Baron

- Improve the recognition of assertions in biomedical texts
- Develop approaches to integrate and triangulation assertions with other evidence
- Develop approaches to prioritise the assertions and linked publications according to user requirements

Next step:

A PhD project supported by UoB PGR scholarship (2025-2029) on Applying LLM and NLP approaches to automate processes in evidence synthesis and triangulation

Supervised by

- Yi Liu, Zhaozhen Xu, Julian Higgins
- Edwin Simpson

- Based on a set of inclusion / exclusion criteria, apply LLM and RAG methods to filter and screen candidate studies
- Adapt methods to a wide array of extraction objectives
  - For specific research questions, e.g. the PICO framework
  - For specific study designs, e.g. randomized control trials, Mendelian randomization studies
- Data extraction from tables and supplementary material, based on the content in the main text

Acknowledgements

Tom Gaunt

Benjamin Elsworth

Pau Erola

Gibran Hemani

(and colleagues in EpiGraphDB research team)

Julian Higgins

Zhaozhen Xu

IEU and BRMS

Zahraa Abdallah

Edwin Simpson

Shelby Temple

University of Bristol

Michael Tillich

Martin Baron

Nadja Schaefer

Roche

MRC Integrative Epidemiology Unit

University of BRISTOL

MRC Integrative Epidemiology Unit

University of BRISTOL

Thank you

MRC Integrative Epidemiology Unit

𝕏 @MRC_IEU

bristol.ac.uk/integrative-epidemiology