UKB-KG: Knowledge Graph for Integrating and Enhancing Biomedical Insights from the UK Biobank

Dr. Hongtu Zhu
Kenan Distinguished Professor
University of North Carolina at Chapel Hill
URL: www.med.unc.edu/bigs2/

My collaborators: Jieping Ye, Jie Wang, Zhe Wang, Yue Shen, and Kaixian Yu

Content

1. Motivation
2. Methods
3. Results
4. Applications

Content

1. Motivation
2. Methods
3. Results
4. Applications

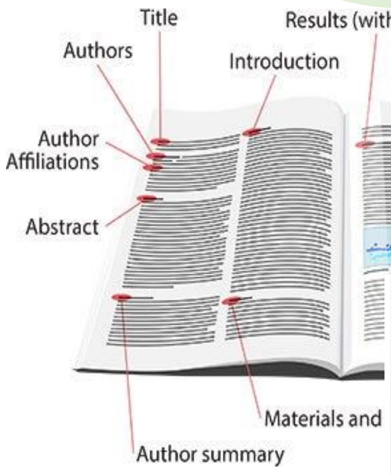
BIOMEDICAL DATA RESOURCES

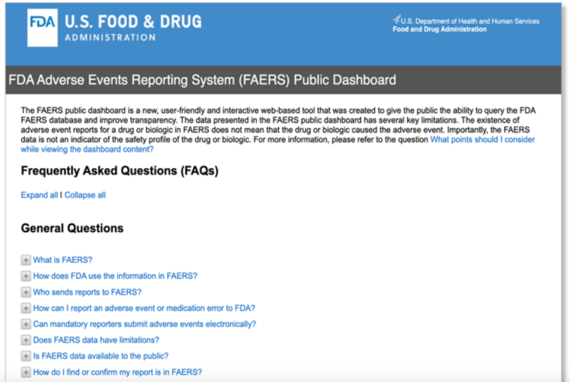
Literature

- **Peer-reviewed articles, preprints, case reports**
- **Sources: PubMed, bioRxiv, medRxiv**

Ontologies

- **Standard vocabularies for data harmonization**
- **Examples: UMLS, SNOMED CT, ICD-10, MeSH**




FDA U.S. FOOD & DRUG ADMINISTRATION
U.S. Department of Health and Human Services Food and Drug Administration

FDA Adverse Events Reporting System (FAERS) Public Dashboard

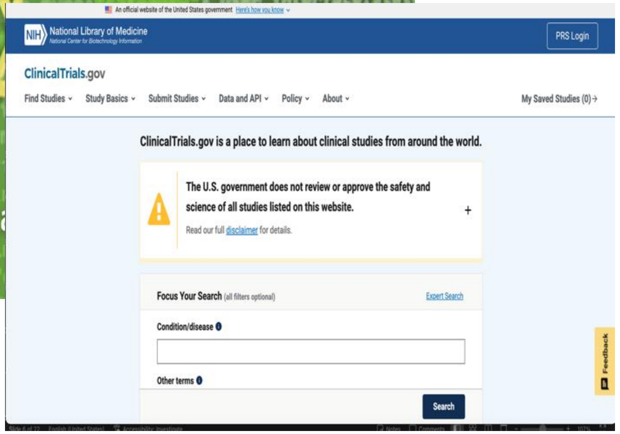
The FAERS public dashboard is a new, user-friendly and interactive web-based tool that was created to give the public the ability to query the FDA FAERS database and improve transparency. The data presented in the FAERS public dashboard has several key limitations. The existence of adverse event reports for a drug or biologic in FAERS does not mean that the drug or biologic caused the adverse event. Importantly, the FAERS data is not an indicator of the safety profile of the drug or biologic. For more information, please refer to the question What points should I consider while viewing the dashboard content?

Frequently Asked Questions (FAQs)

Expand all I Collapse all

General Questions

- What is FAERS?
- How does FDA use the information in FAERS?
- Who sends reports to FAERS?
- How can I report an adverse event or medication error to FDA?
- Can mandatory reporters submit adverse events electronically?
- Does FAERS data have limitations?
- Is FAERS data available to the public?
- How do I find or confirm my report is in FAERS?


An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
PRS Login
ClinicalTrials.gov
Find Studies
Study Basics
Submit Studies
Data and API
Policy
About
My Saved Studies (0)
Expert Search
Search
Feedback

ClinicalTrials.gov is a place to learn about clinical studies from around the world.

The U.S. government does not review or approve the safety and science of all studies listed on this website.

Read our full disclaimer for details.

Focus Your Search (all filters optional)

Condition/disease

Other terms

Drug Information

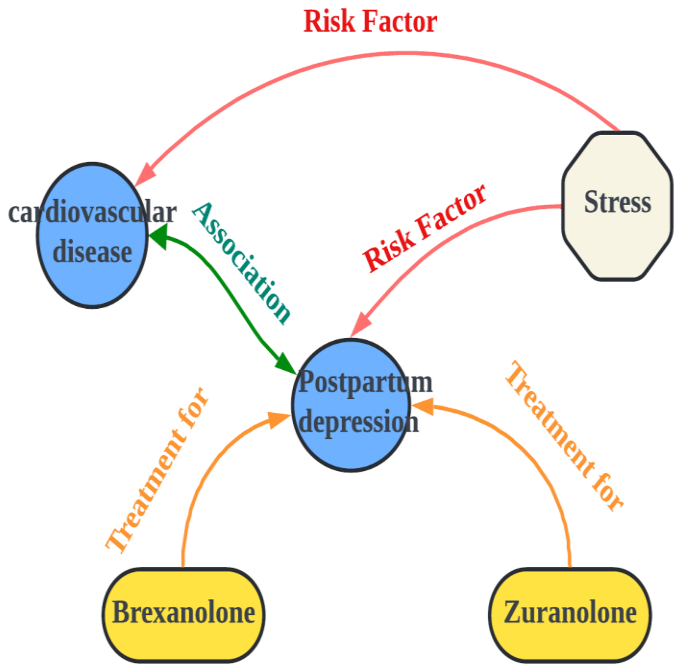
- **Prescription and utilization records**
- **Pharmacogenomic annotations (gene–drug interactions)**
- **Drug databases and adverse-event reports (e.g., FAERS)**

Clinical Trials & Registries

- **Interventional protocols and real-world outcomes**
- **Examples: ClinicalTrials.gov, SEER cancer registry**

BIOMEDICAL KNOWLEDGE GRAPH

- A **heterogeneous directed graph** comprising various types of entities (nodes , e.g.,proteins, diseases) connected by relationships (edges , e.g., "causes," "treats").
- Each triple represents a fact in the real world



Triples:

(Cardiovascular Disease, Association, Postpartum Depression)

(Stress, Risk Factor, Cardiovascular Disease)

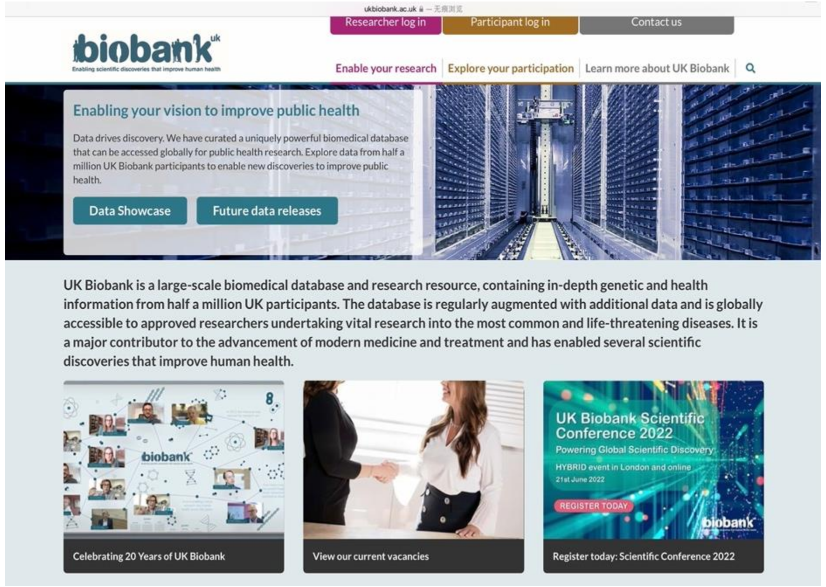
(Stress, Risk Factor, Postpartum Depression)

(Brexanolone, Treatment for, Postpartum Depression)

(Zuranolone, Treatment for, Postpartum Depression)

Any Issues?

THE UK BIOBANK STUDY


ukbiobank.ac.uk
Researcher log in
Participant log in
Contact us
biobank uk
Enabling scientific discoveries that improve human health
Enable your research
Explore your participation
Learn more about UK Biobank

Enabling your vision to improve public health

Data drives discovery. We have curated a uniquely powerful biomedical database that can be accessed globally for public health research. Explore data from half a million UK Biobank participants to enable new discoveries to improve public health.

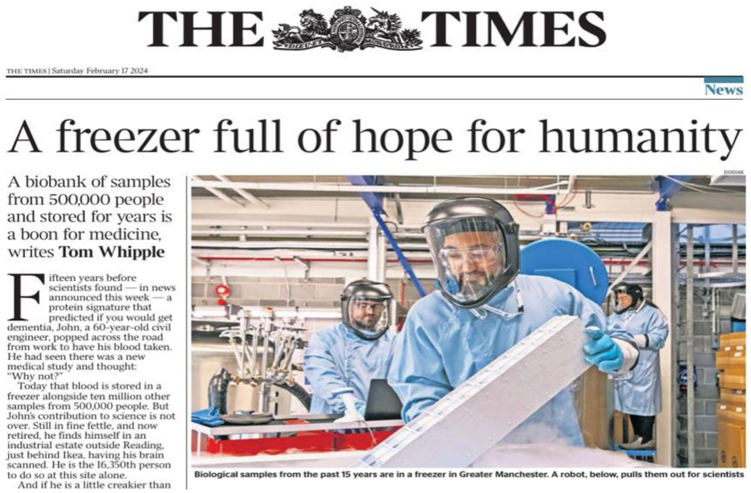
UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. The database is regularly augmented with additional data and is globally accessible to approved researchers undertaking vital research into the most common and life-threatening diseases. It is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health.

UK Biobank Scientific Conference 2022
Powering Global Scientific Discovery
HYBRID event in London and online
21st June 2022
REGISTER TODAY
biobank

Celebrating 20 Years of UK Biobank

View our current vacancies

Register today: Scientific Conference 2022


THE TIMES

THE TIMES | Saturday February 17 2024

News

A freezer full of hope for humanity

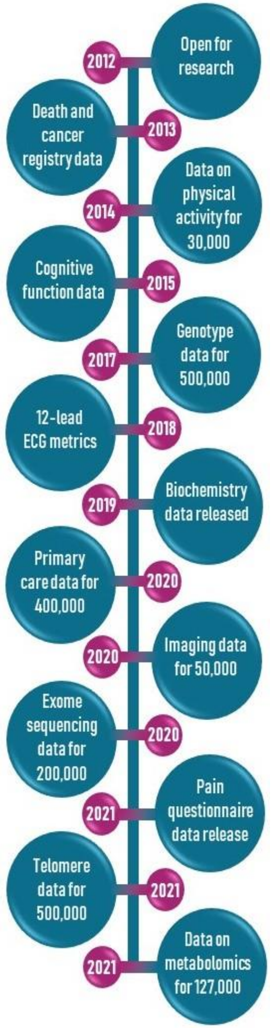
A biobank of samples from 500,000 people and stored for years is a boon for medicine, writes **Tom Whipple**

Fifteen years before scientists found — in news announced this week — a protein signature that predicted if you would get dementia, John, a 60-year-old civil engineer, popped across the road from work to have his blood taken. He had seen there was a new medical study and thought: "Why not?"

Today that blood is stored in a freezer alongside ten million other samples from 500,000 people. But John's contribution to science is not over. Still in fine fettle, and now retired, he finds himself in an industrial estate outside Reading, just behind Ikea, having his brain scanned. He is the 16,350th person to do so at this site alone.

And if he is a little creakier than

Biological samples from the past 15 years are in a freezer in Greater Manchester. A robot, below, pulls them out for scientists


2012 Open for research

2013 Death and cancer registry data

2014 Data on physical activity for 30,000

2015 Cognitive function data

2017 Genotype data for 500,000

2018 12-lead ECG metrics

2019 Biochemistry data released

2020 Primary care data for 400,000

2020 Imaging data for 50,000

2020 Exome sequencing data for 200,000

2021 Pain questionnaire data release

2021 Telomere data for 500,000

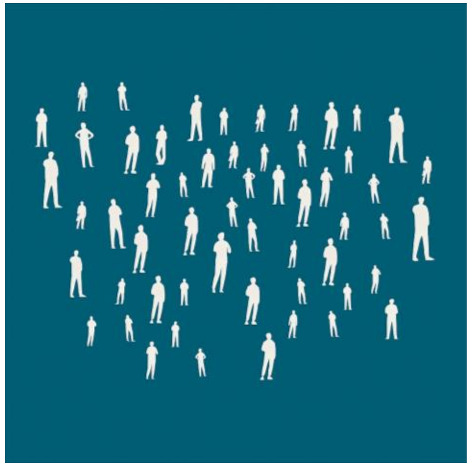
2021 Data on metabolomics for 127,000

- Imaging (n = 100 000): Brain, heart & full-body MRI; DEXA; carotid ultrasound
- Genetics: Genome-wide genotyping, WES & WGS (all participants), Proteomics
- EHR Linkages: Mortality, cancer registry, hospital admissions & primary-care data
- Biomarkers: 30+ serum assays at baseline + first repeat
- Activity: 7-day wrist accelerometry (all) with seasonal follow-up (subset)
- Surveys: Diet, lifestyle, occupation, pain, cognition & mental/digestive health
- Samples: Blood & urine (all); saliva (n = 100 000)

UKB-KG: BACKGROUND

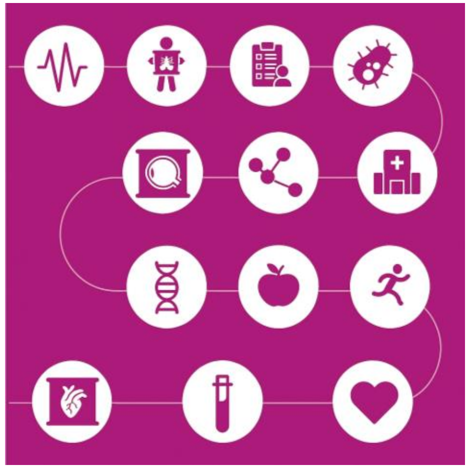
UK Biobank (UKB) - The world's most important health research database

Enabling scientific discoveries that improve human health



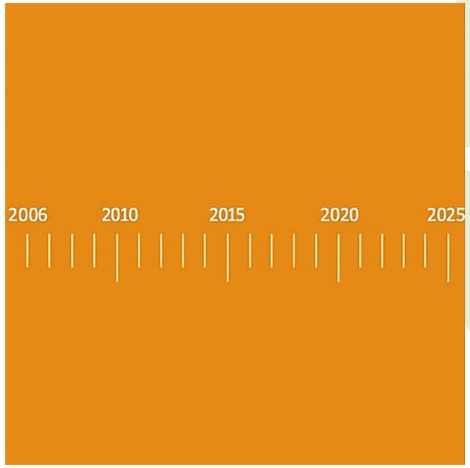
Large

502,467 men and women aged 40-69 with diverse ethnic backgrounds agreed to make all their health-related data available for research.



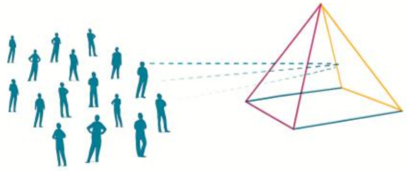
Deep

Extensive dataset includes phenotypic, genomic, lifestyle, environmental and biological factors relevant to subsequent development of different diseases.



Long

Sufficient duration of follow-up for large numbers of many different health outcomes to have occurred since assessment of exposures.



Accessible

All of the data are readily available for any type of health-related research by academic and commercial researchers worldwide.

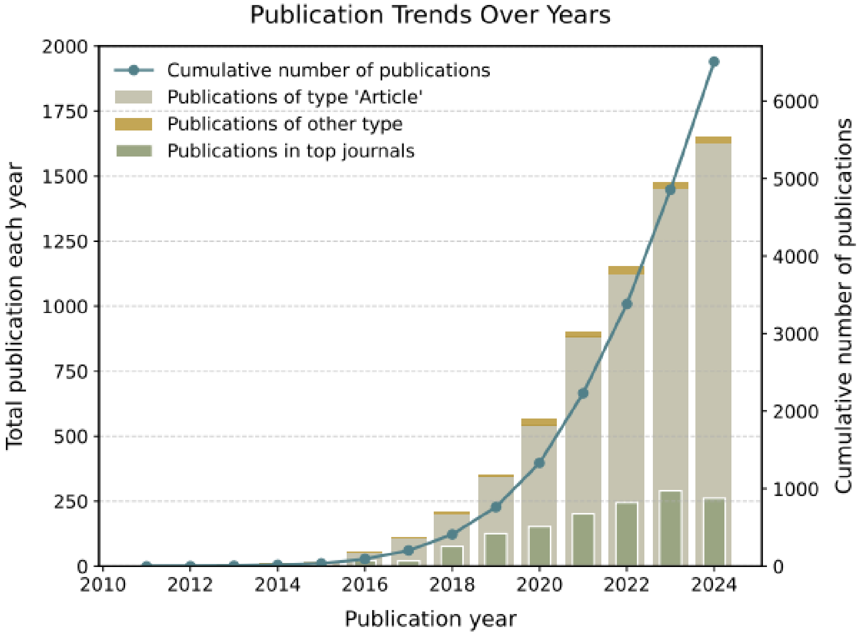
UKB has a significant research impact, supporting numerous high-quality studies on disease progression, risk factor identification, personalized medicine, disease prediction and prevention, and large-scale genetic research.

UKB-KG: BACKGROUND

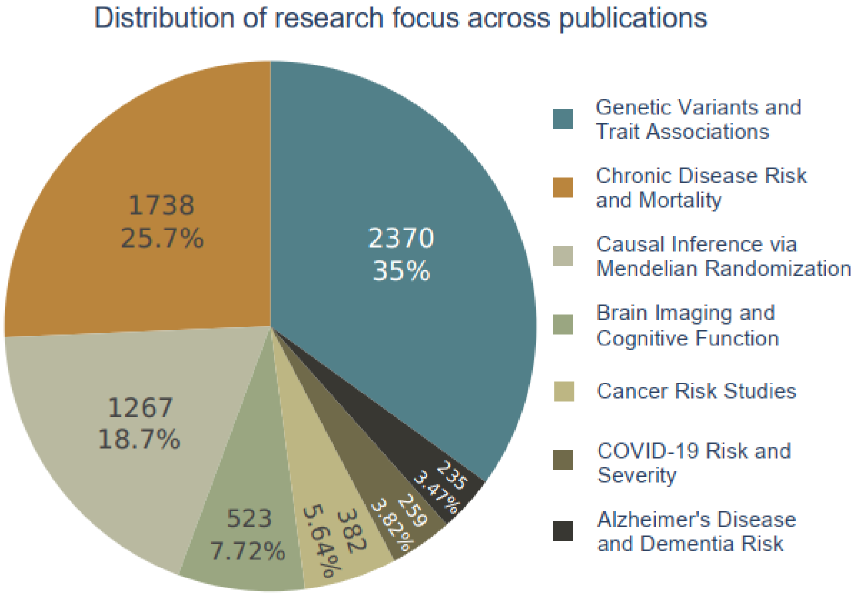
UK Biobank (UKB) - The world's most important health research database

Enabling scientific discoveries that improve human health

However, the dispersion of UKB-related research across numerous publications poses challenges for efficiently synthesizing and integrating findings.



(a) Publication trends over years



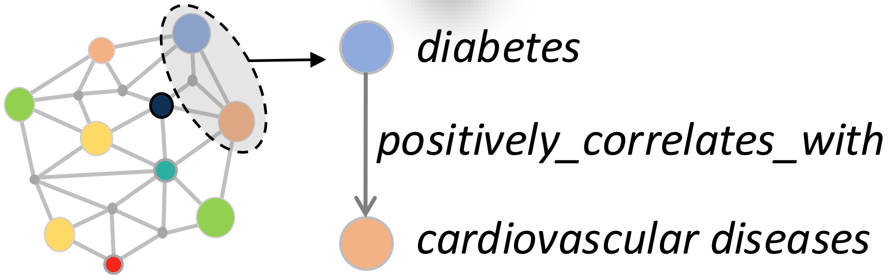
(b) Thematic distribution of publications

A medical knowledge graph offers an effective solution to this challenge by structurally organizing and integrating scattered UKB findings, especially with the recent support of large language models.

UKB-KG: BACKGROUND

Medical Knowledge Graph - a powerful tool for organizing and utilizing vast, heterogeneous biomedical data

Unstructured literature (such as free text and tables)



diabetes

positively_correlates_with

cardiovascular diseases

Structured knowledge units:
triples *[head, relation, tail]*

Large Language Models

- Strong contextual and semantic understanding
- Low human supervision requirement
- Scalability across large corpora

Large language models are revolutionizing MKG construction, significantly enhanced the extraction and organization of medical knowledge from vast literature

UKB-KG: BACKGROUND

Medical Knowledge Graph - a powerful tool for organizing and utilizing vast, heterogeneous biomedical data

Existing MKGs face significant challenges.
A major limitation is the omission of **contextual features**.

☐ **Source information**
essential for assessing the credibility and relevance of extracted relationships

☐ **Baseline characteristics**
like gender, age, ethnicity, and lifestyle, influencing disease susceptibility, symptoms, and treatment outcomes.

[severe chest pain, is_a_precursor_to, heart attack]

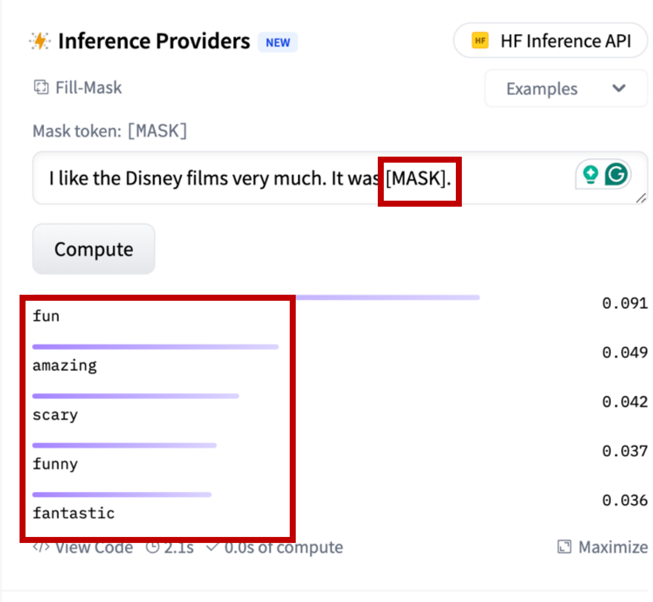
↓

- May primarily apply to men
- Women often exhibit atypical symptoms like fatigue

UKB-KG, a specialized and reliable MKG designed to support diverse UKB-related biomedical studies, addressing challenges like data reliability and missing contextual features

Language Models >>Large LMs (LLMs)

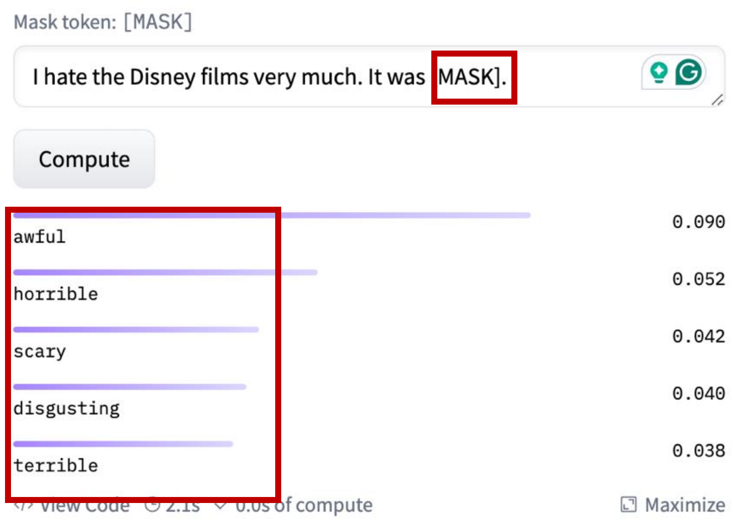
- **Definition:** A language model aims to predict the probability of the occurrence of a token or a sequence of tokens.
- The probability prediction of a language model is closely related to **context** and **corpus** information.
- Language model is **not a new technical** concept specially for LLMs, but has evolved with the advance of artificial intelligence over the decades.
- **Definition:** LLMs have billions of parameters, trained on massive corpora.



Mask token: [MASK]

I like the Disney films very much. It was [MASK].

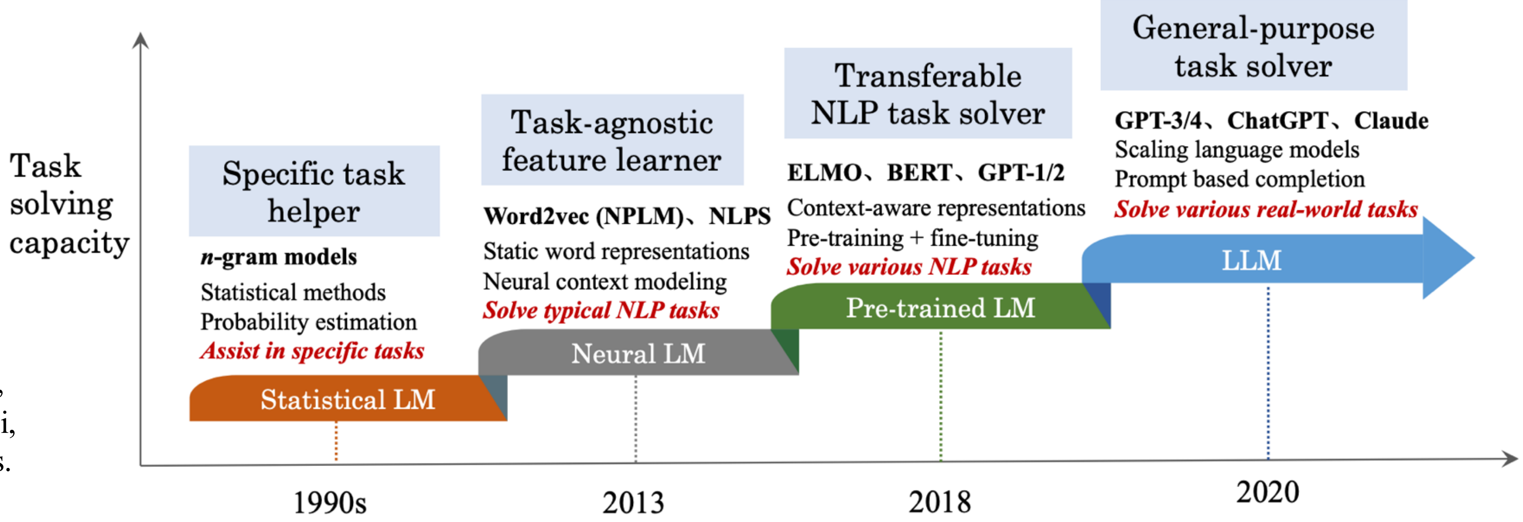
fun	0.091
amazing	0.049
scary	0.042
funny	0.037
fantastic	0.036



Mask token: [MASK]

I hate the Disney films very much. It was [MASK].

awful	0.090
horrible	0.052
scary	0.042
disgusting	0.040
terrible	0.038



Task solving capacity

Specific task helper
***n*-gram models**
Statistical methods
Probability estimation
Assist in specific tasks
Statistical LM
1990s

Task-agnostic feature learner
Word2vec (NPLM)、NLPS
Static word representations
Neural context modeling
Solve typical NLP tasks
Neural LM
2013

Transferable NLP task solver
ELMO、BERT、GPT-1/2
Context-aware representations
Pre-training + fine-tuning
Solve various NLP tasks
Pre-trained LM
2018

General-purpose task solver
GPT-3/4、ChatGPT、Claude
Scaling language models
Prompt based completion
Solve various real-world tasks
LLM
2020

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., . . . Wen, J. (2023). A Survey of Large Language Models. *ArXiv*. https://arxiv.org/abs/2303.18223

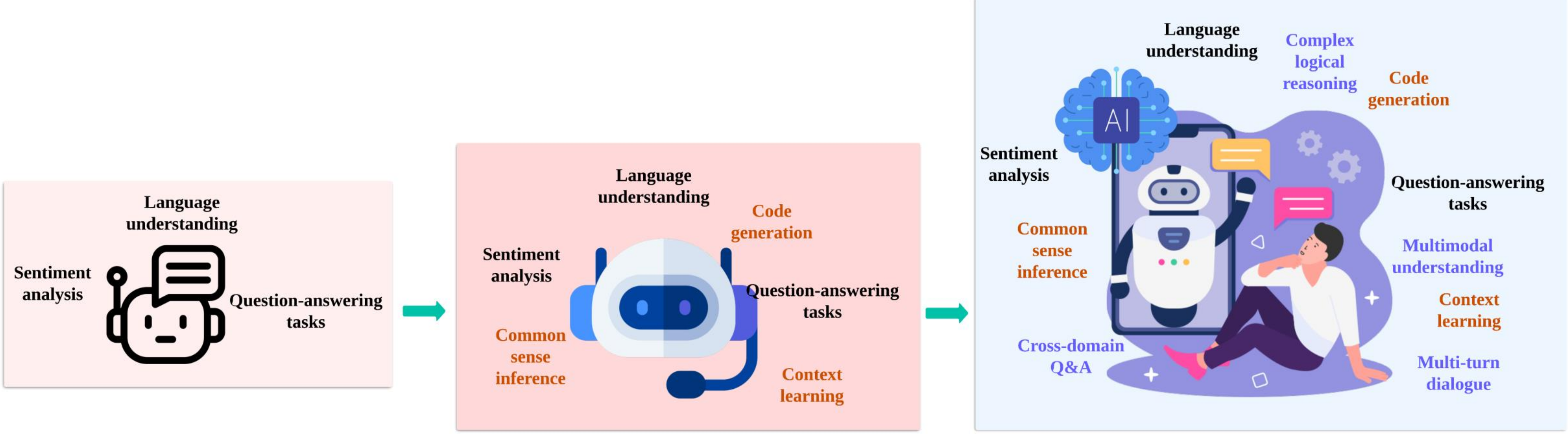
Evolution brings new abilities

Key Point: With the continuous iteration and updates of LLMs, the range of problems they can solve has become increasingly rich, demonstrating some new abilities.

▶ **Growing Complexity:** Each new LLM version shows improvements in reasoning, creativity, and context handling.

▶ **Emergent Properties:** Larger, more diverse training corpora lead to surprising capabilities (e.g., zero-shot translation, chain-of-thought prompting).

▶ **Broadening Applications:** Beyond text generation, LLMs now assist in code completion, legal drafting, and even scientific problem-solving.



Language understanding
Sentiment analysis
Question-answering tasks

20 billion parameters

Language understanding
Code generation
Sentiment analysis
Question-answering tasks
Common sense inference
Context learning

130 billion parameters

Language understanding
Complex logical reasoning
Code generation
Sentiment analysis
Question-answering tasks
Common sense inference
Multimodal understanding
Context learning
Cross-domain Q&A
Multi-turn dialogue

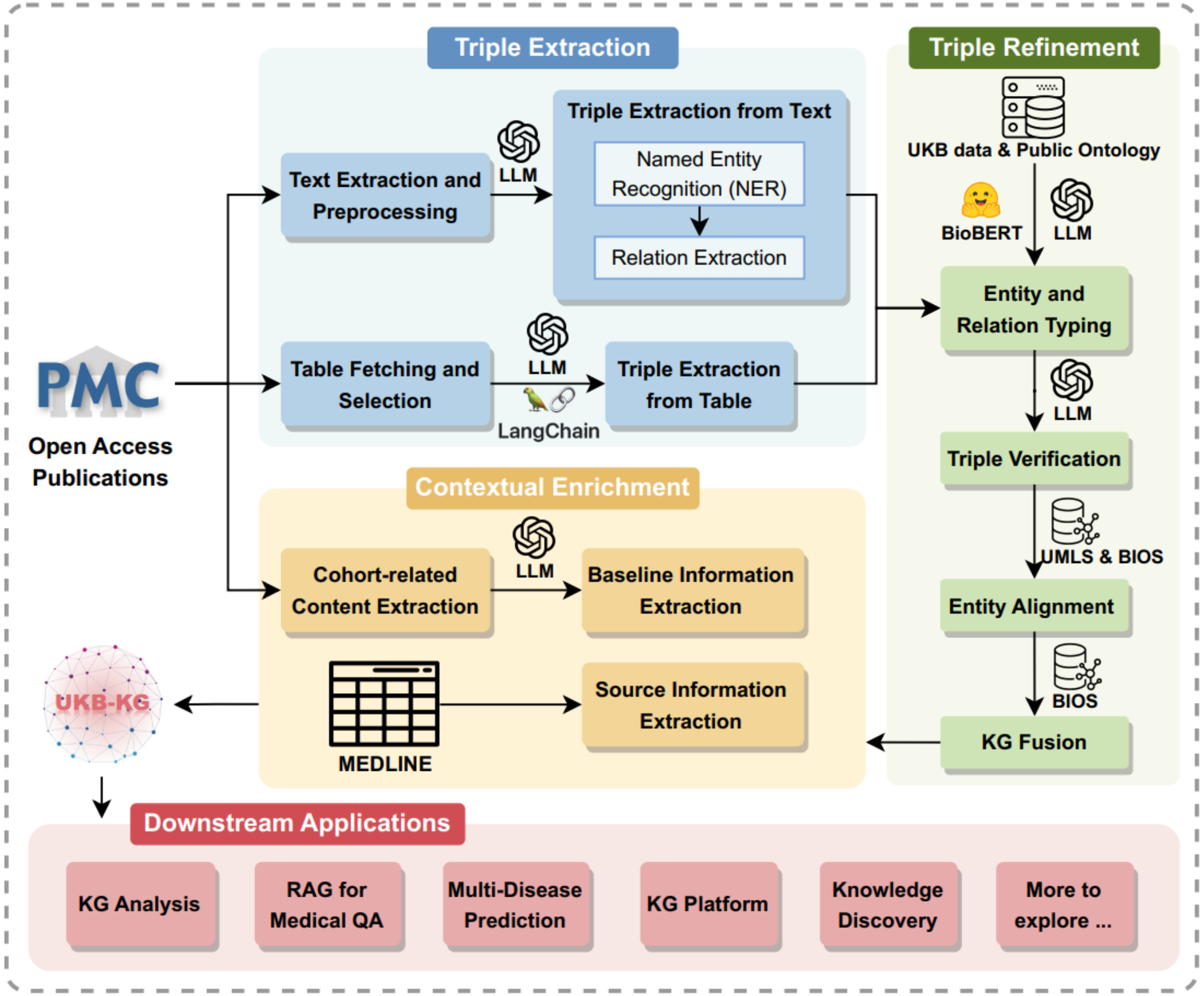
1750 billion parameters

Content

1. Motivation
2. **Methods**
3. Results
4. Applications

UKB-KG: METHODS

Pipeline



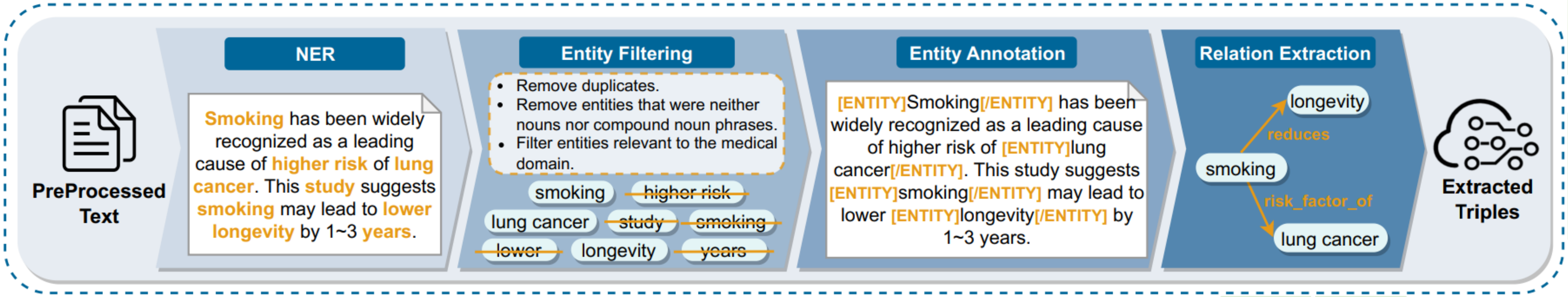
Source: PMC open access publications

- **Module1 – Triple Extraction**
- **Module2 – Triple Refinement**
- **Module3 – Contextual Enrichment**

UKB-KG: can support many downstream applications

UKB-KG: METHODS

Pipeline - Module1: Triple Extraction



PreProcessed Text

NER

Smoking has been widely recognized as a leading cause of higher risk of lung cancer. This study suggests smoking may lead to lower longevity by 1~3 years.

Entity Filtering

- Remove duplicates.
- Remove entities that were neither nouns nor compound noun phrases.
- Filter entities relevant to the medical domain.

smoking ~~higher risk~~ lung cancer ~~study~~ ~~smoking~~ ~~lower~~ longevity ~~years~~

Entity Annotation

[ENTITY]Smoking[/ENTITY] has been widely recognized as a leading cause of higher risk of [ENTITY]lung cancer[/ENTITY]. This study suggests [ENTITY]smoking[/ENTITY] may lead to lower [ENTITY]longevity[/ENTITY] by 1~3 years.

Relation Extraction

Extracted Triples

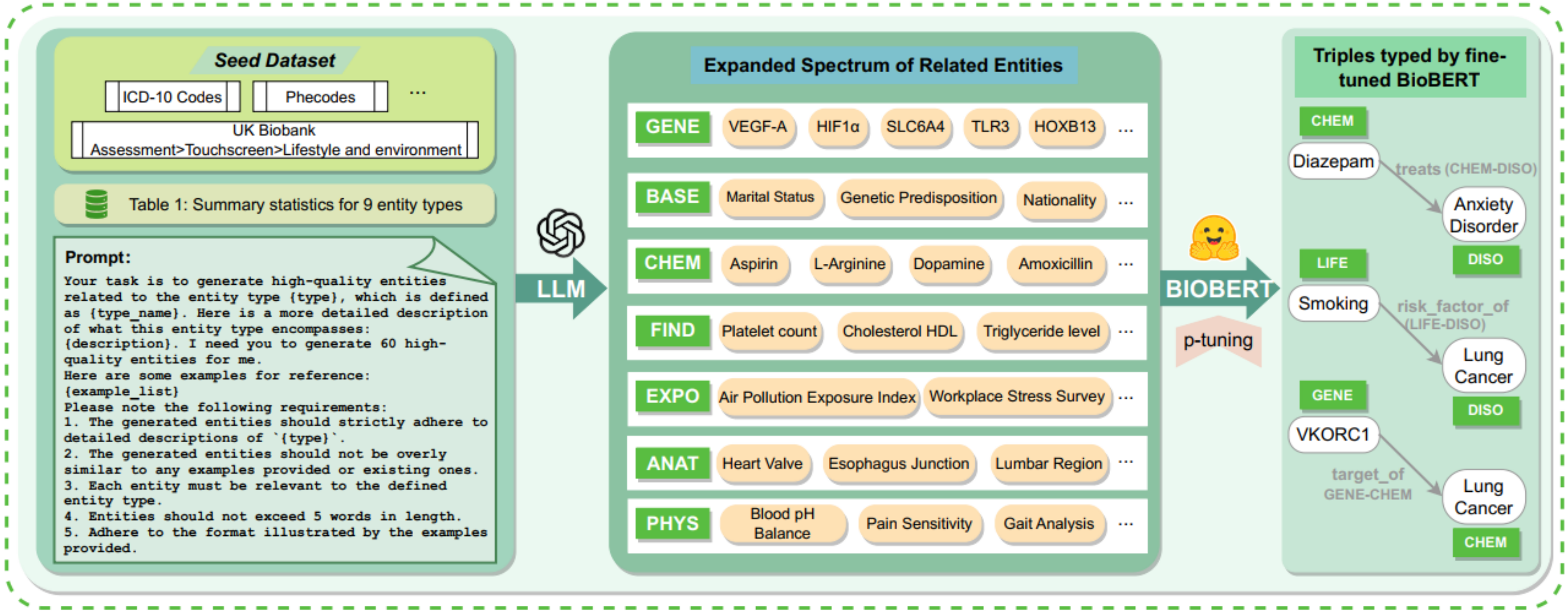
- **Named Entity Recognition (NER)**
  Extract meaningful medical entities.
- **Entity Filtering**
  Remove duplicates and entities unrelated to medicine.
- **Relation Extraction**
  Annotate entities in the corpus using *[ENTITY]* and *[/ENTITY]* tags.
  Leverage the In-Context Learning capability of LLM to extract relations.

UKB-KG: METHODS

Pipeline – Module2: Triple Refinement

- **Triple Typing:** Classify entities into 10 categories using a BioBERT-based model trained on LLM-generated data.
- **Triple Verifier:** Apply LLM self-verification to reduce hallucinations.
- **Entity Alignment:** Standardize terms by mapping entities to UMLS and BIOS concepts.
- **Graph Fusion:** Improve graph connectivity by integrating relevant relationships by integrating relevant relationships from BIOS

The process of Triple Typing:

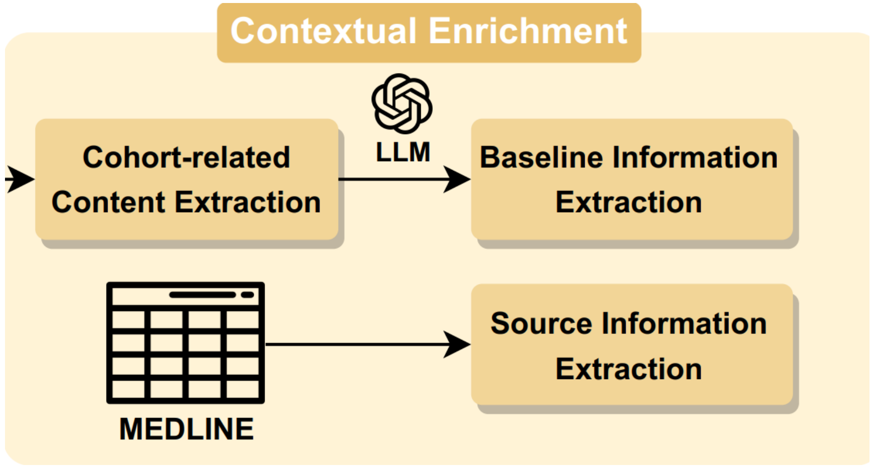


Prompt:

```
Your task is to generate high-quality entities
related to the entity type {type}, which is defined
as {type_name}. Here is a more detailed description
of what this entity type encompasses:
{description}. I need you to generate 60 high-
quality entities for me.
Here are some examples for reference:
{example_list}
Please note the following requirements:
1. The generated entities should strictly adhere to
detailed descriptions of `{type}`.
2. The generated entities should not be overly
similar to any examples provided or existing ones.
3. Each entity must be relevant to the defined
entity type.
4. Entities should not exceed 5 words in length.
5. Adhere to the format illustrated by the examples
provided.
```

UKB-KG: METHODS

Pipeline – Module3: Contextual Enrichment


Contextual Enrichment
Cohort-related Content Extraction
LLM
Baseline Information Extraction
MEDLINE
Source Information Extraction

Add contextual attributes to each triple, including:

- **Source Information** (e.g., author, journal)
  Most information was obtained from the MEDLINE Data Element (Field)
- **Baseline Information** (e.g., gender ratio, age distribution, race)
  Obtained by parsing baseline tables, or extracted from the description text.

Content

1. Motivation
2. Methods
3. Results
4. Applications

UKB-KG: RESULTS

UKB-KG - a high quality medical knowledge graph constructed using LLMs

- Constructed from 6,858 UKB-related publications, with key insights extracted from:
  - unstructured texts in abstracts, results, and conclusions
  - structured tables, particularly genetic tables.
- Comprises 64,734 unique nodes and 137,328 relationships.
- Entities are classified into 10 types, like GENE(genes), DISO(diseases) and LIFE(lifestyle).

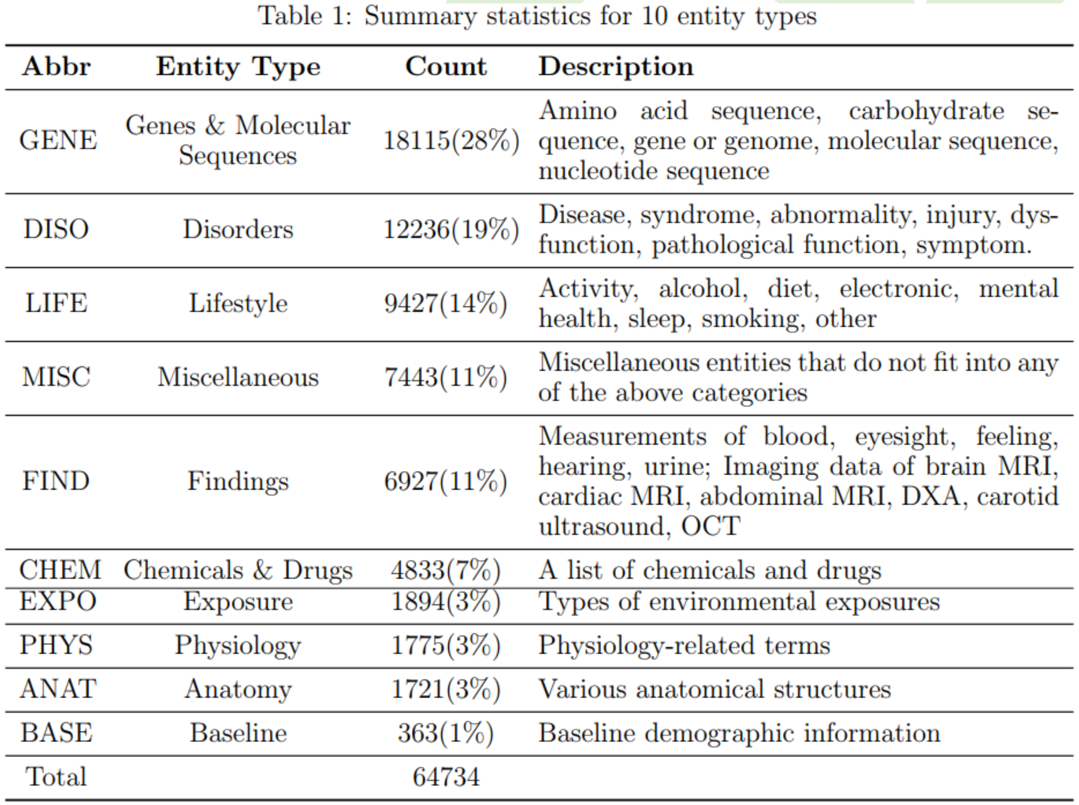


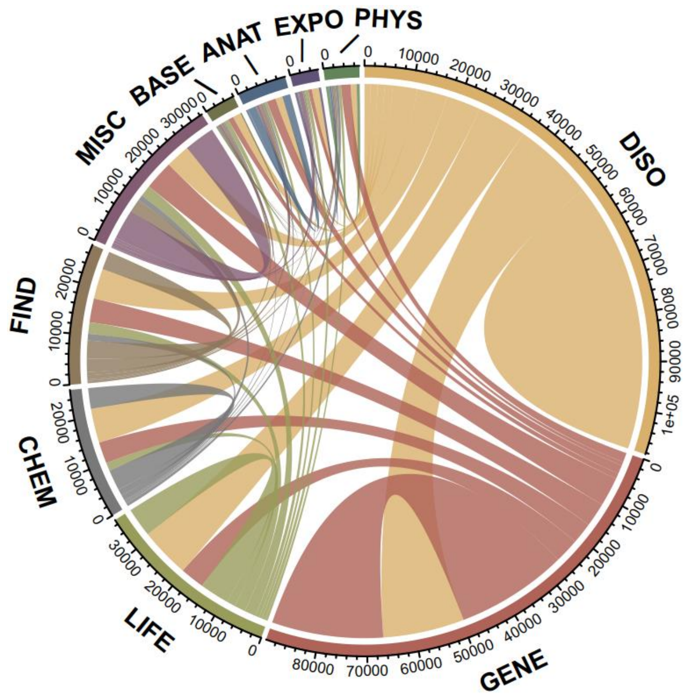
Table 1: Summary statistics for 10 entity types

Abbr	Entity Type	Count	Description
GENE	Genes & Molecular Sequences	18115(28%)	Amino acid sequence, carbohydrate sequence, gene or genome, molecular sequence, nucleotide sequence
DISO	Disorders	12236(19%)	Disease, syndrome, abnormality, injury, dysfunction, pathological function, symptom.
LIFE	Lifestyle	9427(14%)	Activity, alcohol, diet, electronic, mental health, sleep, smoking, other
MISC	Miscellaneous	7443(11%)	Miscellaneous entities that do not fit into any of the above categories
FIND	Findings	6927(11%)	Measurements of blood, eyesight, feeling, hearing, urine; Imaging data of brain MRI, cardiac MRI, abdominal MRI, DXA, carotid ultrasound, OCT
CHEM	Chemicals & Drugs	4833(7%)	A list of chemicals and drugs
EXPO	Exposure	1894(3%)	Types of environmental exposures
PHYS	Physiology	1775(3%)	Physiology-related terms
ANAT	Anatomy	1721(3%)	Various anatomical structures
BASE	Baseline	363(1%)	Baseline demographic information
Total		64734	

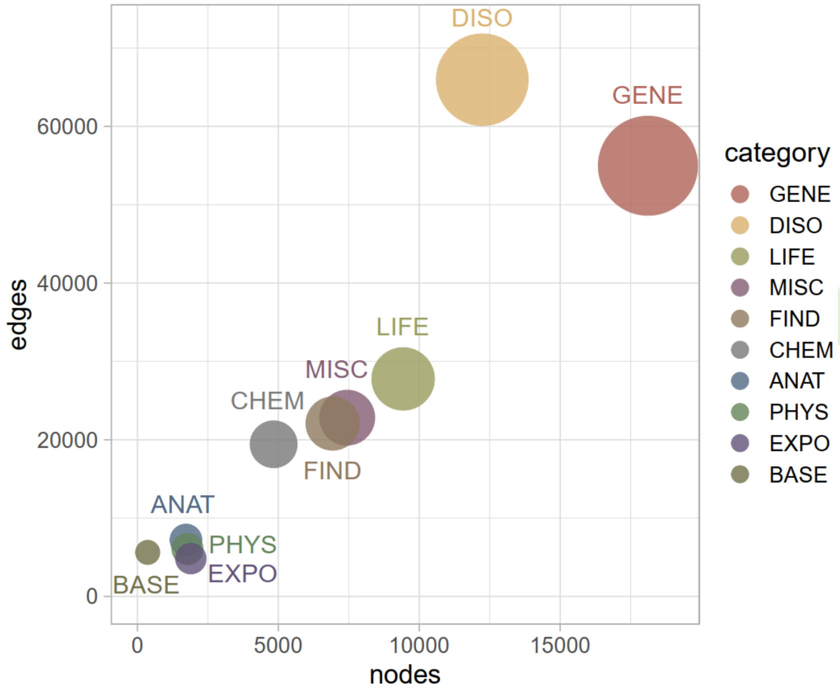
UKB-KG: RESULTS

UKB-KG - a high quality medical knowledge graph constructed using LLMs

- The flow of relationships between different types of nodes:



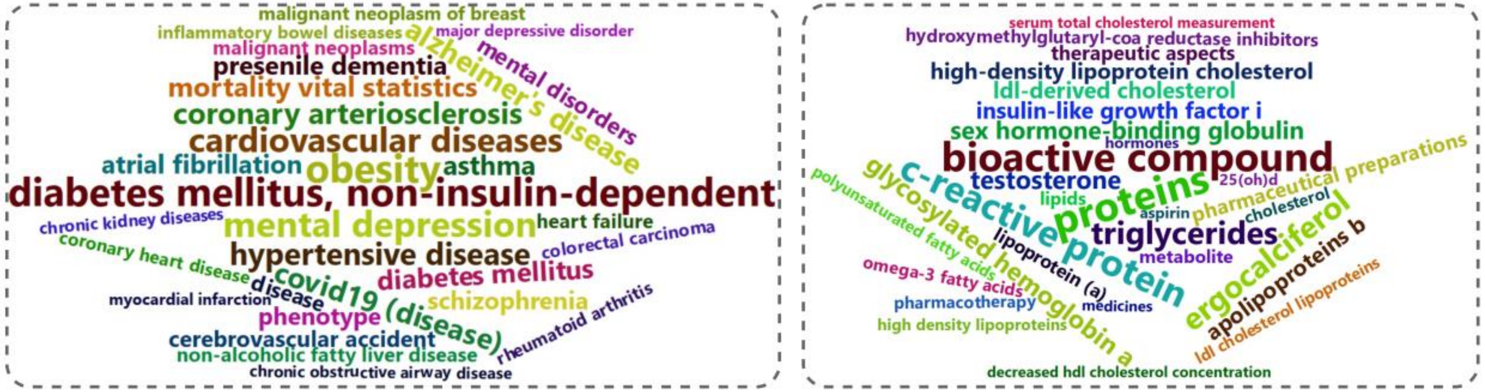
- The distribution of the number of nodes and edges for each type:



UKB-KG: RESULTS

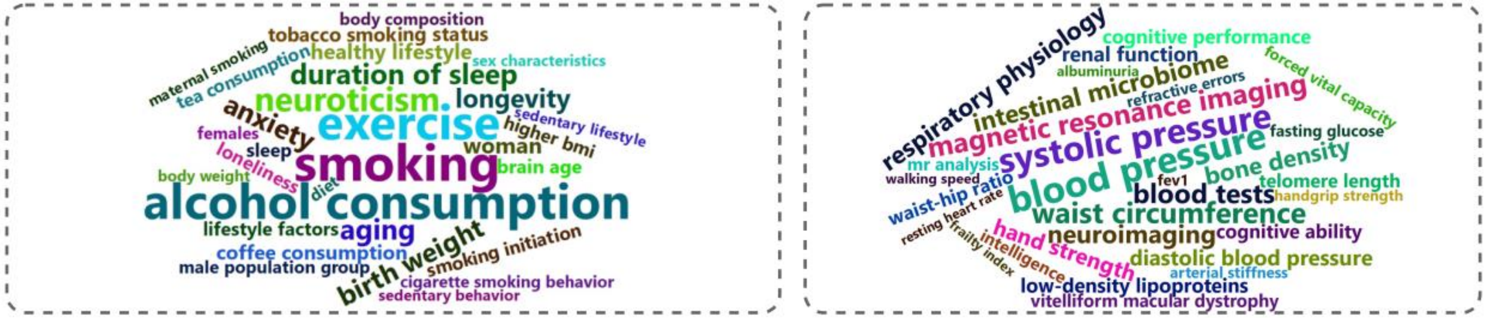
UKB-KG - a high quality medical knowledge graph constructed using LLMs

- ➢ The Entity Cloud which highlights key nodes in the UKB-KG across DISO, CHEM, LIFE, and FIND categories:



(a) DISO entity cloud

(b) CHEM entity cloud



(c) LIFE entity cloud

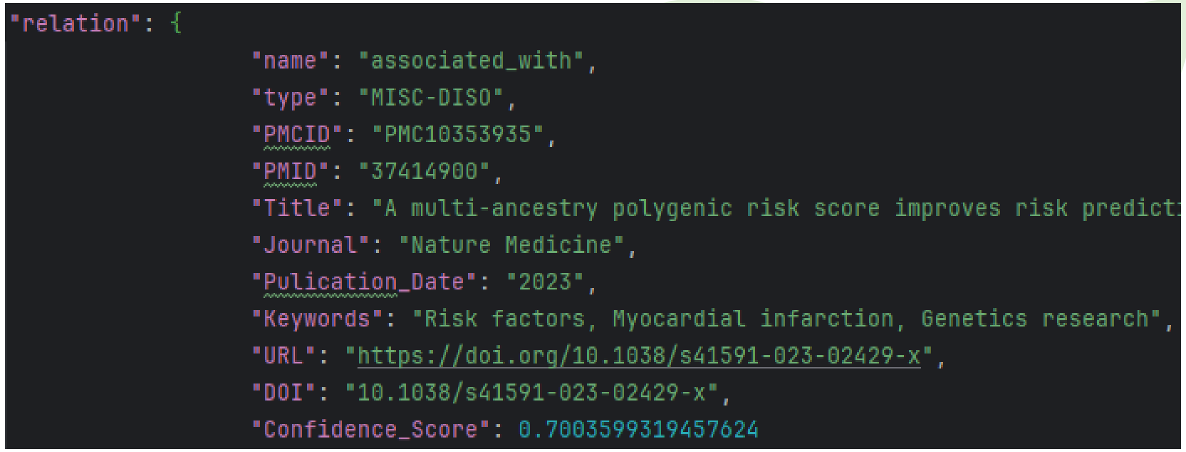
(d) FIND entity cloud

UKB-KG: RESULTS

UKB-KG - a high quality medical knowledge graph constructed using LLMs

A distinguishing feature of UKB-KG is the inclusion of extensive contextual information for the triple's relational properties

Source Info	Abstract, Affiliation, Article Identifier, Author, Author Identifier, Full Author, Date of Electronic Publication, Date of Publication, Grant Number, ISSN, Issue, Journal Title Abbreviation, Journal Title, Language, Location Identifier, Publication History Status, Publication Type, PubMed Central Identifier, PubMed Unique Identifier, Source, Title, Volume, Keywords, URL, Journal Impact Factor, Citations, If use UKB data
Baseline Info	Cohort name, Overall sample size, Average age, Age distribution, Gender ratio, Gender distribution, Major race, Racial distribution, Educational attainment, Employment status



```
"relation": {
                "name": "associated_with",
                "type": "MISC-DISO",
                "PMCID": "PMC10353935",
                "PMID": "37414900",
                "Title": "A multi-ancestry polygenic risk score improves risk predict
                "Journal": "Nature Medicine",
                "Pulication_Date": "2023",
                "Keywords": "Risk factors, Myocardial infarction, Genetics research",
                "URL": "https://doi.org/10.1038/s41591-023-02429-x",
                "DOI": "10.1038/s41591-023-02429-x",
                "Confidence_Score": 0.7003599319457624
```

UKB-KG: RESULTS

UKB-KG - a high quality medical knowledge graph constructed using LLMs

- Triple Precision Evaluation:

Macro Precision	Micro Precision	Number of Recalls
85.62	85.03	9.37

Evaluation method: A random sample of 150 abstracts is selected, from which triples are extracted using GPT-3.5 and evaluated for accuracy using GPT-4o.

- Graph Structure Evaluation:

Node Count	Average Clustering Coefficient	Betweenness Centrality.mean	Betweenness Centrality.max
1423	0.023	3.35	575.00

Evaluation is based on a random sample of 150 abstracts.

Content

1. Motivation
2. Methods
3. Results
4. **Applications**

UKB-KG: APPLICATIONS

UKB-KG enables diverse downstream applications across biomedical research, including:

- **Multi-Disease Prediction**
- **Retrieval-Augmented Generation (RAG)**
- **Interactive Platform for UKB-KG**

UKB-KG: APPLICATIONS

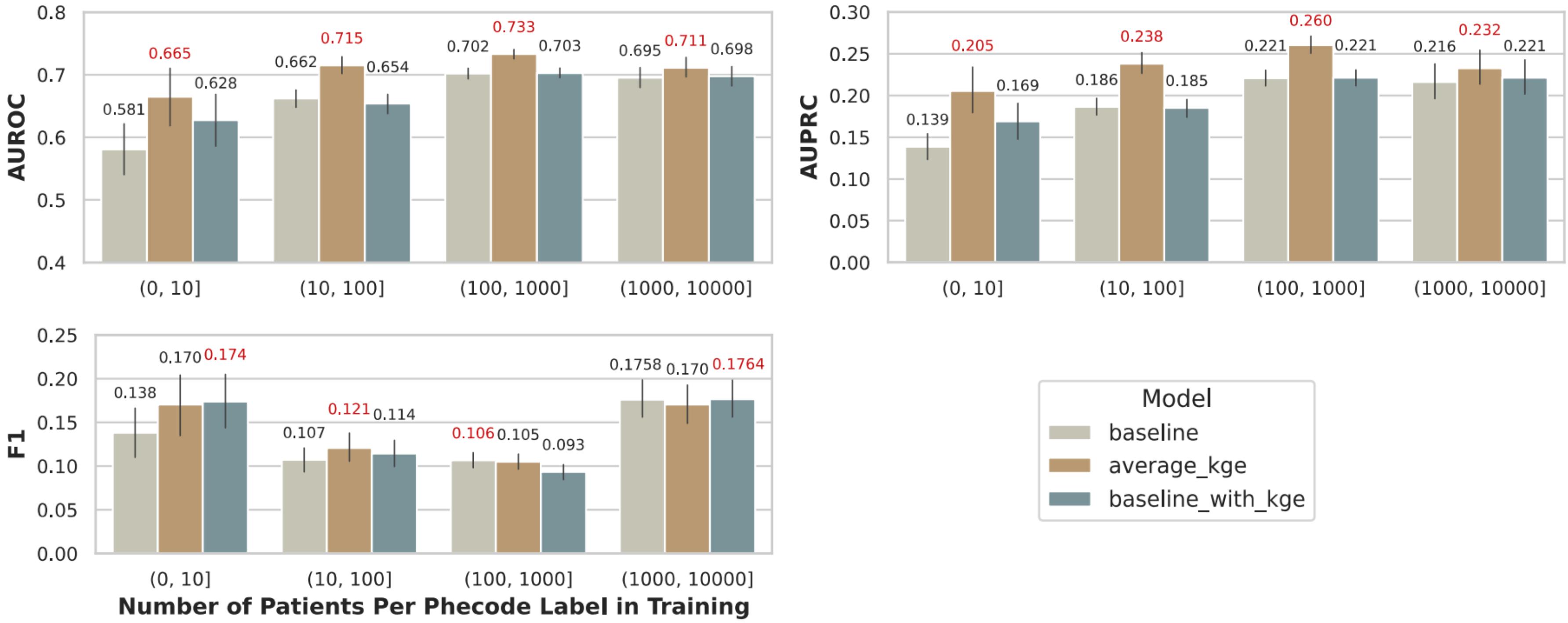
Multi-Disease Prediction

Incorporate domain knowledge through Knowledge Graph Embeddings (KGE) of UKB-KG to enhance disease prediction.

- **Task objective:** predict the occurrence of diseases in the next six months, based on prior medical histories.
- **Dataset**: 278,985 instances from UKB
- **Approaches**:
  - ***Baseline***: use binary indicators (1 for presence, 0 for absence) of medical history variables as input features (1560-d vector)
  - ***Average_KGE***: First, each medical history record is mapped to a node in the UKB-KG. Then use the average KGE of all mapped nodes as input features (1000-d vector)
  - ***Baseline_with_KGE***: the input combines the baseline features with the average KGE features (2560-d vector)

UKB-KG: APPLICATIONS

Multi-Disease Prediction



UKB-KGE significantly enhances multi-disease prediction accuracy, particularly for rare diseases with limited data, by mitigating data insufficiency through the integration of prior knowledge.

UKB-KG: APPLICATIONS

Retrieval-Augmented Generation (RAG)

Retrieve relevant knowledge from the UKB-KG to augment medical question answering

- **Task objective:** Answer medical questions with “Yes,” “No,” or “Maybe”
- **Medical QA Dataset**: PubMedQA
- **Approaches**:
  - ***LLM itself***
  - ***Zero/Few-shot CoT***
  - ***UKB-KG based RAG***: A graph-based RAG method tailored to UKB-KG, combining neighbor-based and path-based retrieval strategies while incorporating triple confidence scores derived from contextual features and semantic similarity.

UKB-KG: APPLICATIONS

Retrieval-Augmented Generation (RAG)

	GPT-4o	Zero-shot CoT	One-shot CoT	UKB-KG based RAG
Accuracy	0.508	0.600	0.668	0.720
Macro-F1	0.488	0.544	0.594	0.633
Macro-Precision	0.632	0.612	0.628	0.646

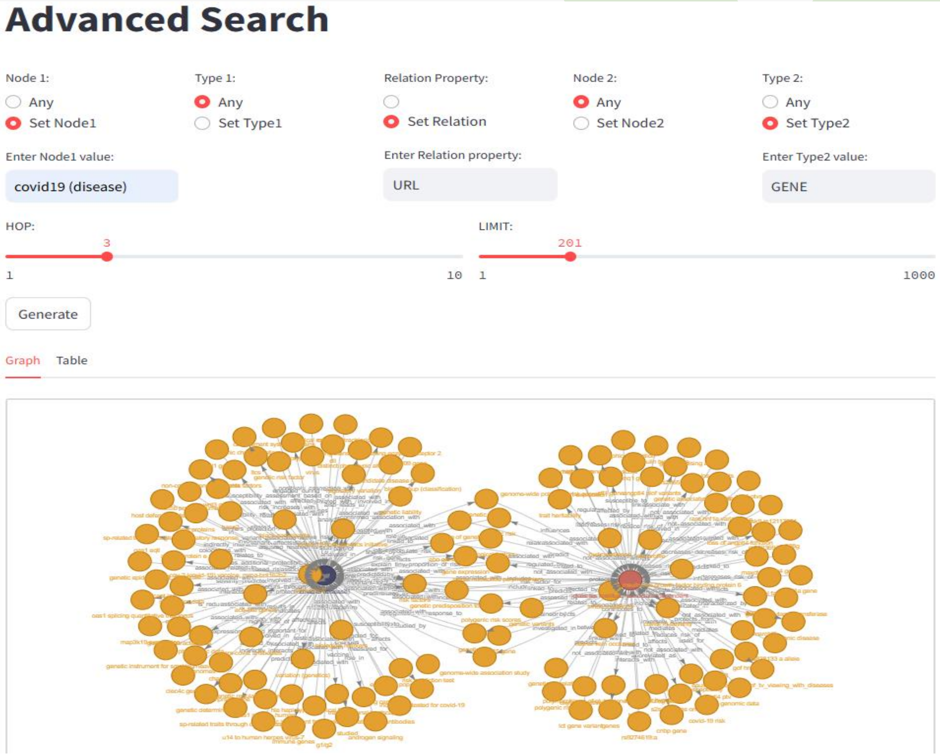
The UKB-KG based RAG leverages the UKB-KG as an external knowledge source to retrieve relevant information in real time, thereby enhancing the accuracy and reliability of medical question answering

UKB-KG: APPLICATIONS

Interactive Platform for UKB-KG - a user-friendly graph platform for viewing and analyzing UKB-KG data

The platform supports complex graph queries, contextual feature exploration, and a RAG Chatbot, enabling efficient access to essential information and insights, even for users with limited technical expertise.

Customized settings available for graph search:



Advanced Search

Node 1:
○ Any
● Set Node1

Type 1:
● Any
○ Set Type1

Relation Property:
○
● Set Relation

Node 2:
● Any
○ Set Node2

Type 2:
○ Any
● Set Type2

Enter Node1 value: covid19 (disease)

Enter Relation property: URL

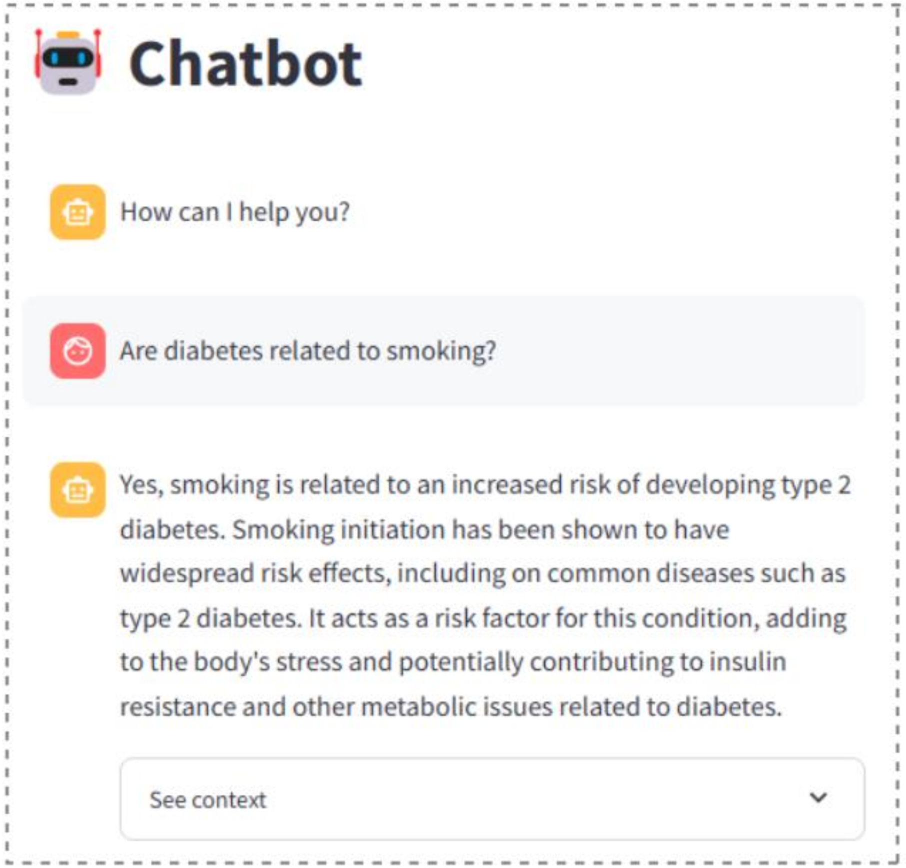
Enter Type2 value: GENE

HOP: 3 (1–10)

LIMIT: 201 (1–1000)

UKB-KG: APPLICATIONS

Interactive Platform for UKB-KG - a user-friendly graph platform for viewing and analyzing UKB-KG data

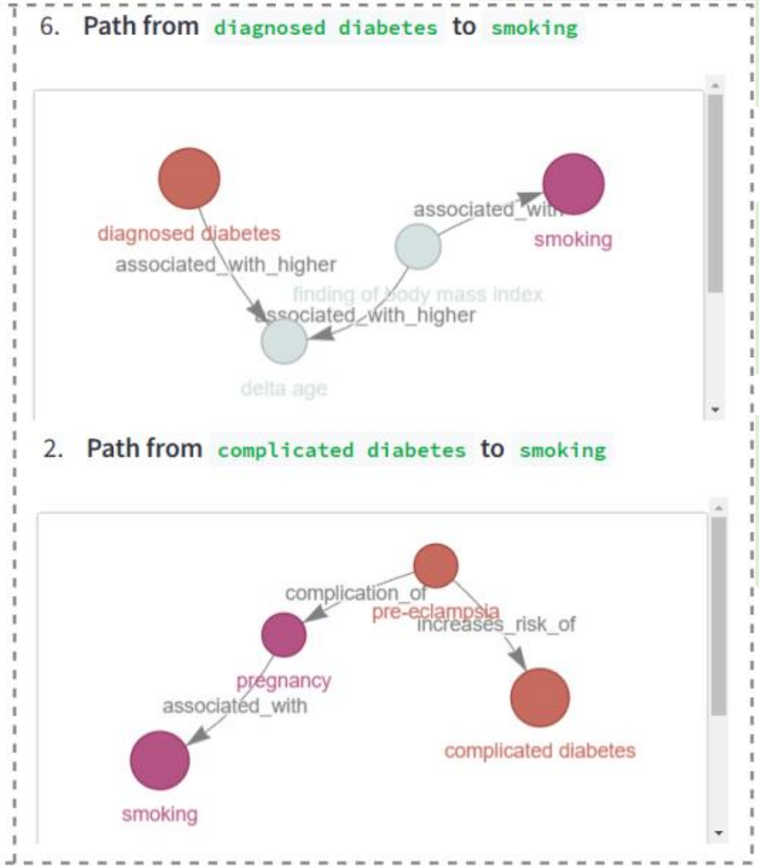


Chatbot

How can I help you?

Are diabetes related to smoking?

Yes, smoking is related to an increased risk of developing type 2 diabetes. Smoking initiation has been shown to have widespread risk effects, including on common diseases such as type 2 diabetes. It acts as a risk factor for this condition, adding to the body's stress and potentially contributing to insulin resistance and other metabolic issues related to diabetes.



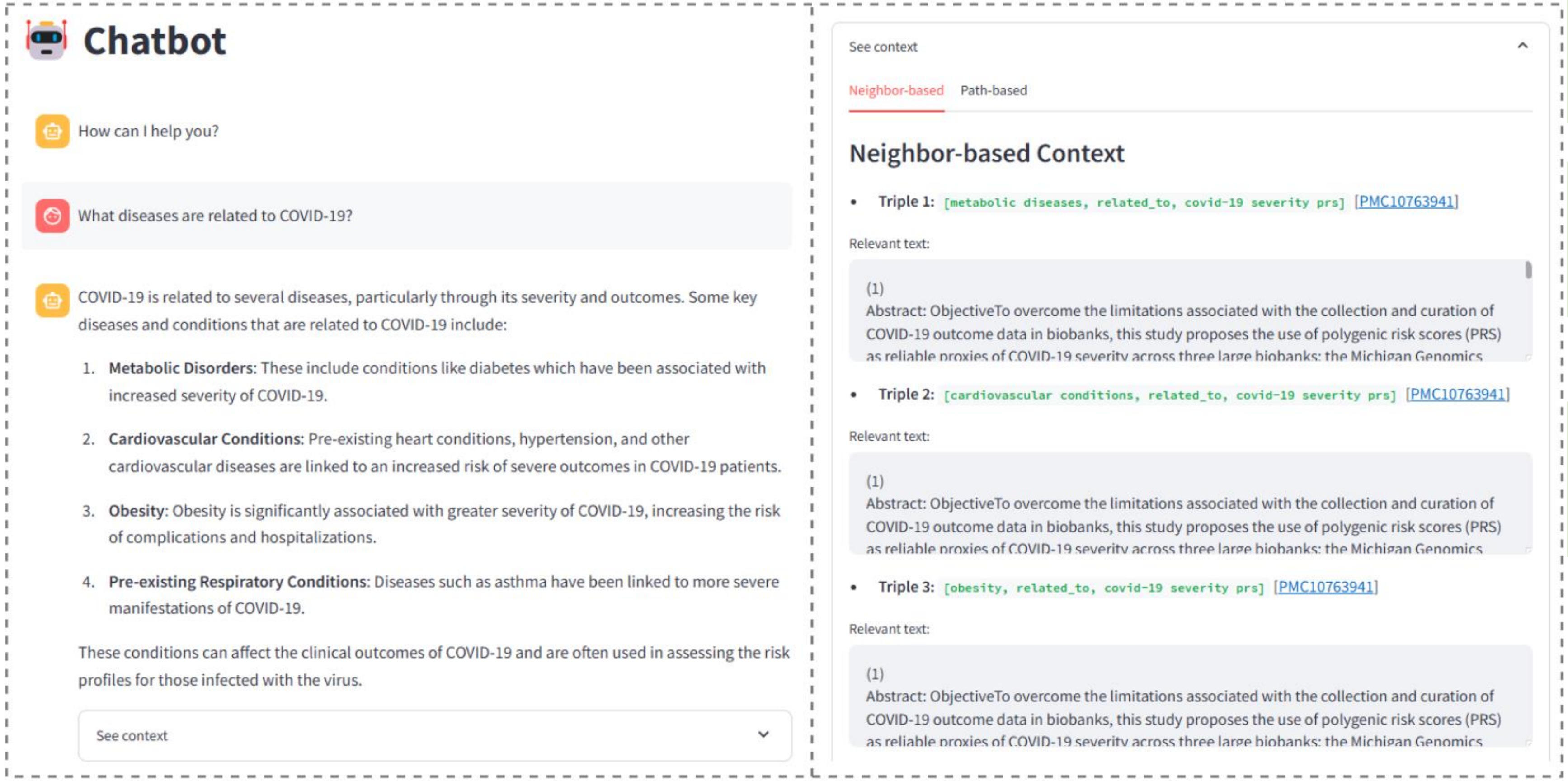
6. Path from `diagnosed diabetes` to `smoking`

2. Path from `complicated diabetes` to `smoking`

RAG Chatbot example: the answer to "Are diabetes related to smoking?" accompanied by Path-based Context, highlighting potential direct or indirect relationships between *diabetes* and *smoking*

UKB-KG: APPLICATIONS

Interactive Platform for UKB-KG - a user-friendly graph platform for viewing and analyzing UKB-KG data



Chatbot

How can I help you?

What diseases are related to COVID-19?

COVID-19 is related to several diseases, particularly through its severity and outcomes. Some key diseases and conditions that are related to COVID-19 include:

1. **Metabolic Disorders**: These include conditions like diabetes which have been associated with increased severity of COVID-19.
2. **Cardiovascular Conditions**: Pre-existing heart conditions, hypertension, and other cardiovascular diseases are linked to an increased risk of severe outcomes in COVID-19 patients.
3. **Obesity**: Obesity is significantly associated with greater severity of COVID-19, increasing the risk of complications and hospitalizations.
4. **Pre-existing Respiratory Conditions**: Diseases such as asthma have been linked to more severe manifestations of COVID-19.

These conditions can affect the clinical outcomes of COVID-19 and are often used in assessing the risk profiles for those infected with the virus.

Neighbor-based Context

- **Triple 1:** `[metabolic diseases, related_to, covid-19 severity prs]` [PMC10763941]

Relevant text:

(1)
Abstract: ObjectiveTo overcome the limitations associated with the collection and curation of COVID-19 outcome data in biobanks, this study proposes the use of polygenic risk scores (PRS) as reliable proxies of COVID-19 severity across three large biobanks: the Michigan Genomics

- **Triple 2:** `[cardiovascular conditions, related_to, covid-19 severity prs]` [PMC10763941]

Relevant text:

(1)
Abstract: ObjectiveTo overcome the limitations associated with the collection and curation of COVID-19 outcome data in biobanks, this study proposes the use of polygenic risk scores (PRS) as reliable proxies of COVID-19 severity across three large biobanks: the Michigan Genomics

- **Triple 3:** `[obesity, related_to, covid-19 severity prs]` [PMC10763941]

Relevant text:

(1)
Abstract: ObjectiveTo overcome the limitations associated with the collection and curation of COVID-19 outcome data in biobanks, this study proposes the use of polygenic risk scores (PRS) as reliable proxies of COVID-19 severity across three large biobanks: the Michigan Genomics

RAG Chatbot example: response to “What diseases are related to COVID-19?” accompanied by Neighbor-based Context, which includes *COVID-19*-related triples, associated PMCIDs, hyperlinks, and source texts.

Thank You!