



## UvA-DARE (Digital Academic Repository)

### BASIL DB

*bioactive semantic integration and linking database*

Jackson, David; Groth, Paul; Harmouch, Hazar

#### DOI

[10.1186/s13326-025-00336-3](https://doi.org/10.1186/s13326-025-00336-3)

#### Publication date

2025

#### Document Version

Final published version

#### Published in

Journal of Biomedical Semantics

#### License

CC BY-NC-ND

[Link to publication](#)

#### Citation for published version (APA):

Jackson, D., Groth, P., & Harmouch, H. (2025). BASIL DB: bioactive semantic integration and linking database. *Journal of Biomedical Semantics*, 16, Article 14.  
<https://doi.org/10.1186/s13326-025-00336-3>

#### General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

#### Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

DATABASE

Open Access



# BASIL DB: bioactive semantic integration and linking database

David Jackson<sup>1\*</sup>, Paul Groth<sup>1</sup> and Hazar Harmouch<sup>1</sup>

## Abstract

**Background** Bioactive compounds found in foods and plants can provide health benefits, including antioxidant and anti-inflammatory effects. Research into their role in disease prevention and personalized nutrition is expanding, but challenges such as data complexity, inconsistent methods, and the rapid growth of scientific literature can hinder progress. To address these issues, we developed BASIL DB (BioActive Semantic Integration and Linking Database), a knowledge graph (KG) database that leverages natural language processing (NLP) techniques to streamline data organization and analysis. This automated approach offers greater scalability and comprehensiveness than traditional methods such as manual data curation and entry.

**Construction and content** The process of constructing the BASIL DB is divided into four fundamental steps: data collection, data preprocessing, data extraction, and data integration. Data on bioactives and foods are sourced from structured databases. The relevant randomized controlled trials (RCTs) were extracted from PubMed. The data are then prepared by cleaning inconsistencies and structuring them for analysis. In the data extraction phase, NLP tools, including a large language model (LLM), are utilized to analyze clinical trials and extract data on bioactive compounds and their health impacts. The integration phase compiles these data into a knowledge graph, which consists of the entities Foods, Bioactives, and Health Conditions as nodes and their interactions as edges. To quantify the relationships/interactions between these entities, we generate a weight for each edge on the basis of empirical evidence and methodological rigor.

**Utility and discussion** The BASIL DB incorporates 433 compounds, 40296 research papers, 7256 health effects, and 4197 food items. The database features query and visualization capabilities, including interactive graphs and custom filtering options, that showcase different aspects of the data. Users are able to explore the relationships between bioactives and health effects, enhancing both research efficiency and insight discovery.

**Conclusion** The BASIL DB is a knowledge graph database of bioactive compounds. This study provides a structured resource for exploring the relationships among bioactives, foods, and health outcomes, representing a step toward a more systematic and data-driven approach to understanding the health effects of bioactive compounds. Future work will focus on expanding the database and refining the utilized methods. Extending the BASIL DB will help bridge the gap between traditional and conventional approaches to nutrition, guiding future research in bioactive compound discovery and health optimization.

**Availability** Users can access and explore the data via <https://basil-db.github.io/info.html> or fork and run the respective script via <https://github.com/basil-db/script>.

\*Correspondence:  
David Jackson  
d.i.jackson@uva.nl



**Keywords** Bioactive compounds, Knowledge graph, Natural language processing (NLP), Data integration, Clinical trials, PubMed, Evidence-based health

## Background

Bioactive compounds, which are found in foods, plants, and other natural sources, can contribute to health benefits [1]. These compounds, including but not limited to polyphenols, flavonoids, alkaloids, and carotenoids, have been studied for their potential interactions with biological pathways. These interactions may contribute to antioxidant activities, reduced inflammation, and enhanced immune responses [2]. As research continues, the role of these substances in health maintenance and disease prevention is being explored, particularly within the contexts of chronic disease management and personalized nutritional strategies [3]. This growing body of findings indicates that bioactives are pivotal in the evolution of nutritional science and the optimization of health strategies [4].

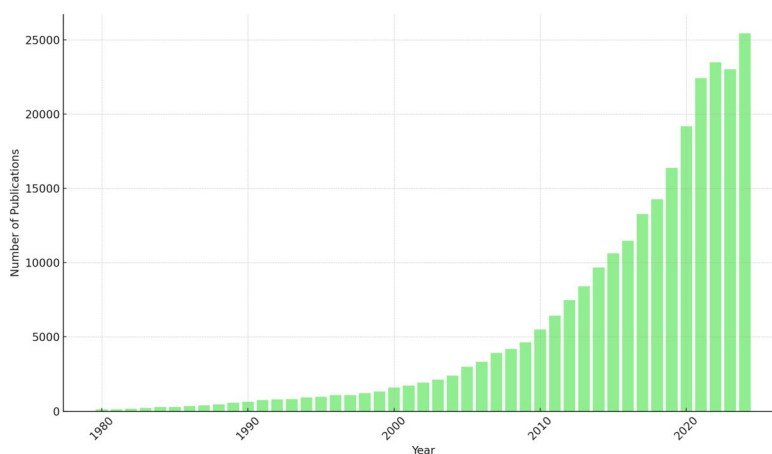
Nutritional data have inherent challenges due to their multifaceted nature, involving a variety of complex biological, environmental, and behavioral factors that make comprehensive analysis and interpretation difficult [5]. In addition, the field is often plagued by problems of confounding variables and heterogeneity in research methods [6], which further impedes the possibility of making definitive conclusions about the health effects of specific bioactive compounds. The endeavor of understanding such health outcomes is all the more complex considering the diversity of bioactive compounds in food products and dietary supplements, together with the often unclear or exaggerated claims about their health benefits [7]. The scale of the challenge is amplified by the amount of scientific data available (see Fig. 1); for example, PubMed is

expanding at a double-exponential rate, now comprising over 36.6 million publications [8]. Furthermore, a lack of standardization and consistency in the scientific literature complicates both consumers and industry specialists in interpreting useful information that can be drawn from available data.

In light of these complexities, there is a strong need for a structured system that can organize this information effectively. Knowledge Graphs (KG) offer a solution for addressing some of these challenges, particularly those related to the heterogeneity of data and the lack of standardization in the scientific literature [9]. By structuring the heterogeneous and multidimensional data related to bioactive compounds, a KG-based approach allows mapping their correlations with different health outcomes and demographic populations in a semantically rich and standardized format. Graph databases have been used in the medical domain for tasks such as discovering drug relationships, supporting diagnoses, providing personalized recommendations, aiding decision-making, and enabling intelligent predictions [10].

In this context, we introduce the BASIL DB (BioActive Semantic Integration and Linking Database), a database envisioned to enhance querying and retrieval capabilities for bioactive compounds, uncovering insights for better health strategies and nutrition recommendations.

To contextualize our work, it is important to recognize the established resources that underpin and precede our database. Established resources like FooDB [11], PhenolExplorer [12], and ChEMBL [13] provide substantial information concerning bioactive compounds and have



**Fig. 1** Annual publication counts on PubMed for research using the query 'bioactives OR phytochemicals OR dietary polyphenols' from 1980 to 2024, highlighting the increasing volume of research in this field

been instrumental in the construction of the BASIL DB. However, while they provide thorough information on the chemical properties of bioactive compounds and a few of their health effects, they rely mainly on labeled data and lack substantial clinical trial information. This includes missing demographic data, meaning that no information is provided on the specific populations that were studied in the respective research.

EBasis [14] is a database of plant-based bioactive compounds in European dietary sources and their biological effects. However, despite its comprehensive coverage, EBasis is limited by its dependence on expert hand-labeled data, which may restrict the scalability of updates and the integration of new insights. BASIL DB employs state-of-the-art NLP (natural language processing) techniques (see “Construction and content” section) to automatically extract and score the literature data, thus accelerating data gathering and reducing manpower costs associated with manual processing. While EBasis integrates data from 1,147 peer-reviewed articles and 567 investigations into human bioeffects [14], our automated approach enabled the assessment of over 100,000 publications.

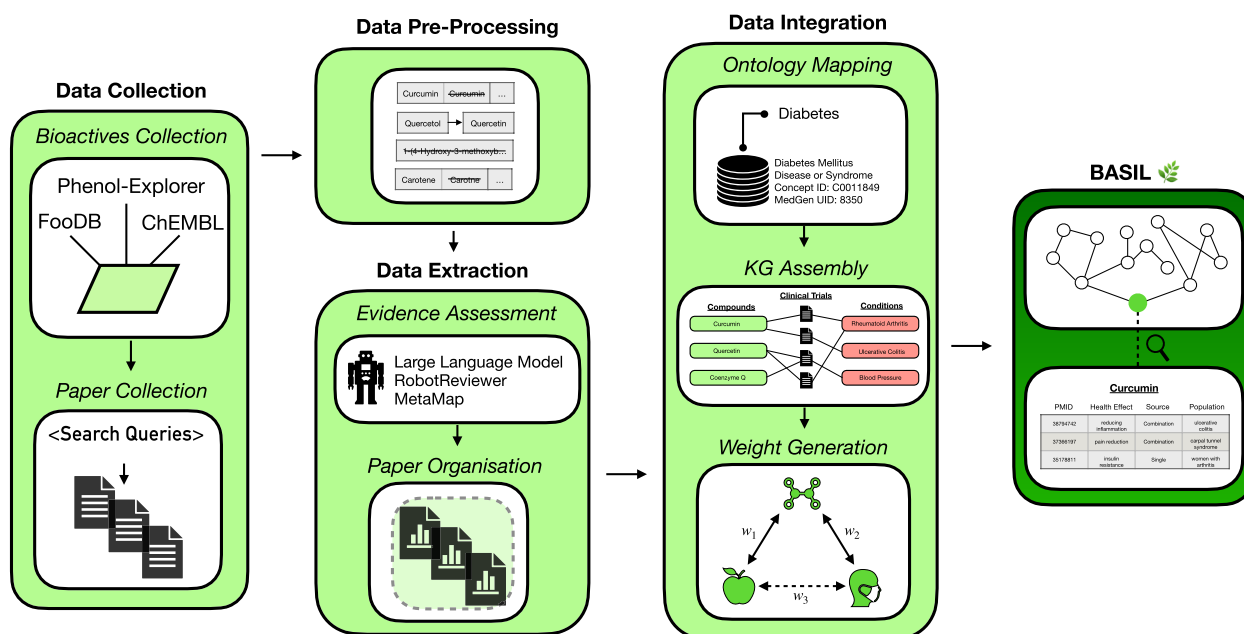
Consequently, the BASIL DB improves the range and diversity of health-related outcomes assessed, generating a more comprehensive understanding of recent literature. Furthermore, human-curated analyses such as

those in EBasis may introduce bias, particularly when data selection and interpretation are influenced by subjective judgments [15]. While acknowledging that the complete elimination of bias is challenging, we aim to elucidate the assessment process and potentially reduce such biases by leveraging algorithmic consistency. An additional goal of this work is the assessment of data quality issues inherent in the health informatics domain. By addressing these challenges, highlighted in “Limitations and future work” section, we aim to increase the reliability of our database and emphasize the need for future research on this topic.

The complexities and challenges discussed in this section underscore the necessity for a novel solution for efficient and accurate data analysis in functional food research.

### Construction and content

In this section, we outline the data sources, tools, and methods for developing the database. This process, which is visualized in Fig. 2, is divided into four components, namely Data Collection (“Data collection”), Data Pre-processing (“Data pre-processing”), Data Extraction (“Data extraction”), and Data Integration (“Data integration”).



**Fig. 2** Visualization of the pipeline used for the construction of the BASIL DB. The process begins with data collection from various sources, followed by preprocessing. The data extraction phase involves collecting relevant papers, assessing evidence with tools such as a large language models and MetaMap, and organizing this information systematically. Integration involves assembling a knowledge graph linking compounds, clinical trials, and conditions and generating weights for these links. The pipeline culminates in a queryable knowledge graph interface, allowing for detailed searches on compounds such as “Curcumin”, which display associated health effects and clinical data

### Data collection

Bioactive information is extracted from three sources: (1) FooDB, which provides detailed information on food ingredients, including bioactives, together with their chemical properties, biological activities, and concentrations in different foods; (2) PhenolExplorer, which contains comprehensive information on polyphenols, a key group of bioactive compounds; and (3) ChEMBL, a database of bioactive molecules with drug-like properties. Food-related data, including food names and bioactive compound contents, were also obtained from FooDB.

Papers are systematically extracted from PubMed via the NIH's E-Fetch utility [16] to identify and gather all relevant randomized controlled trials (RCTs) involving bioactive compounds. The search was optimized by using search queries that combined the bioactive terms and their synonyms with filters for RCTs.

An exemplary query looks like the following:

```
("bioactive term" OR "synonym") AND "randomized controlled trial"[Publication Type].
```

When only RCTs are considered, a major fraction of studies, such as case reports or cohort studies, are disregarded. The high importance of such studies should be noted; however, we restrict ourselves to RCTs for two reasons. First, RCTs have the highest rank of all study types in the evidence pyramid [17], meaning that the evidence presented in RCTs is more reliable than that presented in any other study type. Second, as each publication type is defined by a specific set of traits, the process of extracting and evaluating evidence is eased by standardizing the input type. A total of 101910 papers were extracted. Table 1 summarizes the key data sources used for the creation of the BASIL DB.

### Data pre-processing

To streamline our dataset and improve the efficacy of information extraction, we execute the following pre-processing measures.

A total of 1500 bioactive terms were considered for the information extraction. For simplicity, any bioactive names that exceeded a length of 35 characters or included more than 5 nonalphabetic characters were excluded. Synonyms were grouped using their respective IDs, duplicates were removed, and spelling checks were performed. Furthermore, bioactives with fewer than 3 associated papers, as well as papers without abstracts, were omitted to ensure credibility.

For some food entities, multiple values are reported for the concentration of a compound within a specific food, which may reflect either differences between sources or measurement methods. To perform a weighting (see “Data integration” section), these data are normalized to obtain a single representative value for each compound entry. For example, “European cranberry” is represented

by a single value (8.454 mg/100 g) for the “quercetin” content, which is achieved by averaging the values reported from different sources (e.g., 0.515 mg/100 g and 16.392 mg/100 g). Moreover, food entities with data specific to different parts of the food, such as the fruit, root, or shoot of a tomato, are split into distinct entries to ensure content granularity. Finally, to ensure comparability across different foods (see “Data integration” section), we standardize the measurement units of different content information (e.g., mg/kg, mg/g) to mg per 100 g. For detailed information on the unit conversion please refer to our GitHub page at <https://github.com/basil-db/script>.

### Data extraction

Here, we detail the data extraction process and the technologies deployed to assess and categorize findings from our collected data.

We use NLP tools, including a large language model (LLM), RobotReviewer [18], MetaMap [19], and

**Table 1** Overview of key data sources utilized in the construction process, sorted alphabetically. All listed data sources are structured, with the exception of PubMed, which primarily consists of textual, unstructured data

Data source	Type of data	Access	Primary use	Source location
ChEMBL	Bioactive molecule activities	Open	Drug discovery and development	<a href="https://www.ebi.ac.uk/chembl">https://www.ebi.ac.uk/chembl</a>
FooDB	Chemical compounds in foods	Open	Nutritional and dietary research	<a href="https://foodb.ca">https://foodb.ca</a>
Phenol-Explorer	Polyphenol content in foods	Open	Nutritional epidemiology	<a href="http://phenol-explorer.eu">http://phenol-explorer.eu</a>
PubMed	Biomedical literature	Open	Medical and life science research	<a href="https://pubmed.ncbi.nlm.nih.gov">https://pubmed.ncbi.nlm.nih.gov</a>
UMLS	Health and biomedical vocabularies	Restricted	Healthcare information integration	<a href="https://uts.nlm.nih.gov">https://uts.nlm.nih.gov</a>

rule-based approaches, to systematically extract and categorize data from clinical trial abstracts. Each tool in our data extraction process is selected for its strengths and capabilities: (a) An LLM, specifically GPT-4.0 Turbo [20], is employed to analyze the nuanced connections between concepts within clinical studies. More precisely, the LLM aids in determining what role a specific compound plays in a study, assessing the characteristics of study populations, and identifying observed effects of the respective compounds. The decision to use an LLM was driven by their ability to handle a wide range of unstructured textual data flexibly, which is supported by previous studies evaluating their performance within the medical and scientific domain [21]. (b) RobotReviewer uses the Cochrane Risk of Bias tool [22] to automatically evaluate study bias while also extracting quantitative data such as the number of participants in each study. (c) MetaMap identifies and categorizes medical concepts, particularly conditions such as diseases and symptoms, through Named Entity Recognition (NER). It maps the identified terms to the Unified Medical Language System (UMLS) [23], ensuring that medical terms are standardized and accurately represented. MetaMap frequently returns several partly overlapping entities for just one concept. However, our framework uses only the most relevant and accurate entity. Thus, if two or more concepts overlap, the longest string match will be chosen as the representative concept. When there are several matches of equal length, a UMLS score function [19] calculates the most relevant match, which is then the representative concept. Matches that return a MeSH tree code are also prioritized toward matches that do not. (d) Finally, rule-based methods are applied to extract statistical measures such as hazard ratios and progression-free survival rates from the text, which may provide information on the efficacy and safety of treatments.

While the LLM could have also been used for some of the tasks assigned to RobotReviewer or rule-based methods, such as the extraction of study participants, the tools were chosen instead for cost and efficiency.

In summary, the following features are extracted for each clinical trial:

- **Study Focus:** Trials are categorized on the basis of the extent to which a specific compound was investigated, such as whether the focus was on a single compound, combination therapy, or other modalities. This allows for the identification of the nature and scope of the interventions being tested in different studies.
  - **Health Benefits:** The positive results in each study are recorded via keywords. This approach simplifies the retrieval and synthesis of data for particular health benefits across many studies, hence allowing a systematic comparison and pooling of results.
  - **Population Terms:** Keywords describing the study population are also extracted. This includes demographic and clinical characteristics relevant to the trial, providing insights into the applicability and generalizability of the study results.
  - **Population Size:** An integer value representing the number of participants in the study.
  - **Abstract Bias:** A float value, ranging from 0 to 1, is used to describe the bias in a paper. Bias refers to systematic errors inherent or implemented in the study design, execution, and analysis jeopardizing the validity of results. Specifically, it reflects selection, performance, detection, attrition, and reporting biases [24]. A value of 0 indicates no observed bias, whereas a value of 1 represents maximum bias.
  - **Study Dosage:** The specific dosage information mentioned in the study, including quantitative amounts (e.g., milligrams) and temporal details (e.g., frequency or duration of administration).
  - **Other Compounds:** Names of additional compounds identified in the “Substances” section of the clinical trial.
- The study focus was categorized into one of six designated categories:
- **Single Compound:** The specific bioactive compound was studied independently,
  - **Combination Therapy:** The bioactive compound was used in combination with other treatments,
  - **Derivative:** Studies involving a derivative of the specific bioactive,
  - **Comparative Analysis:** Comparing the specific bioactive with other compounds or treatments,
  - **No Involvement:** The specific bioactive was not mentioned in the study,
  - **Other Focus:** The study was not focused primarily on the specific bioactive compound.
- Moreover, recognizing that not all trials yield positive outcomes, we add a separate column titled “Non-Significant Health Effects”. This column lists the health conditions studied in those trials where no beneficial effects of the compounds under investigation were observed. This distinction is critical to keeping up with the scientific rigor of a dataset since it ensures the inclusion of both favorable and non-favorable results, thereby offering an overall outlook into the research environment. The systematic documentation of experiments where compounds fail to show efficacy helps to support more valid assessments in meta-analyses or systematic reviews.

It also helps to identify the factors that may influence the variability in treatment outcomes, hence leading the research funding and effort to more promising or less explored areas.

### Data integration

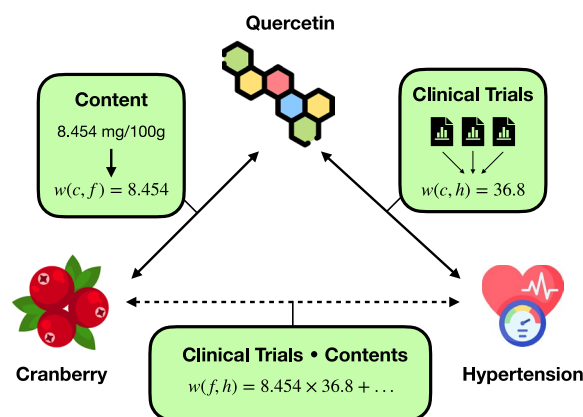
After outlining how data are extracted from clinical studies, we now describe the process of integrating these data into a unified knowledge graph format.

Both “Health Benefits” and “Population Terms” are mapped to the UMLS via MetaMap to structure the LLM output and to enrich it with additional information about each term. This process allows for a direct mapping from each bioactive to UMLS terms, which in this context represent conditions, utilizing UMLS semantic type filtering. By implementing this mapping strategy, we have constructed a KG with condition nodes and edges that link bioactives to these conditions, thus improving data connectivity and precision within the KG.

**Weight generation:** The relationships between entities, namely foods, bioactive compounds, and health conditions are quantitatively expressed through weights. These weights are intended to capture the extent and nature of the relationships between the entities. Each weight expresses the connection strength between a specific food  $f$ , compound  $c$ , or health condition  $h$  on the basis of their composition and empirical evidence. For this purpose, we define three distinct functions to calculate the weights: 1) between compounds and foods, 2) between compounds and health conditions, and 3) between health conditions and foods. Figure 3 illustrates the weight generation process using an example.

**Compound to food:** The weight  $w(c, f)$  between a compound  $c$  and food  $f$  is calculated by the normalized content of  $c$  within  $f$ , where the normalization refers to standardizing the measurement units as described in “Data pre-processing” section.

**Compound to health effect:** The weight  $w(c, h)$  captures the cumulative evidence of the relationship between a bioactive compound and a health condition across multiple studies. Each edge weight  $w(c, h)$  is calculated as the sum of contributions from individual papers that discuss both entities, taking into account the type of effect source, the number of participants, and the observed bias in each study. Established principles have been adapted to account for effect sources, participant numbers and study bias in determining trial significance [22, 25]. For each paper  $p$  that references both bioactive  $c$  and condition  $h$ , we define the following:



**Fig. 3** Visualization of the weight generation process using the example of the bioactive “Quercetin”, the food “Cranberry”, and the condition “Hypertension”. The weight between “Cranberry” and “Hypertension” is calculated using the weights  $w(c, f)$  and  $w(c, h)$  of all compounds  $c$  that are connected to each of the two concepts (such as “Quercetin”)

- $E_p$ : The weight of the effect source, assigned on the basis of the study type involved:
  - Single Compound: 6
  - Combination Therapy or Comparative: 4
  - Derivative: 2
  - Other: 1
- $N_p$ : The number of participants in the study.
- $B_p$ : The bias score, ranging from 0 (no bias) to 1 (maximum bias).

Given the varying scales of study sizes, we normalize participant counts to moderate the influence of large participant counts, thereby enhancing comparability across studies:

$$N'_p = \frac{\log(N_p + 1)}{\log(\max(N) + 1)},$$

where  $N_p$  is the participant count of the study  $p$ , and  $\max(N)$  is the maximum participant count observed across all studies. We also invert  $B$  to reflect a reduction in weight due to higher bias:  $B'_p = 1 - B_p^{0.5}$ . The exponent is applied to prevent a disproportionate impact of the bias score on the overall weighting, given the experimental nature of the bias estimation.

The weight for each paper  $p$  is then calculated as:  $w_p = E_p \cdot N'_p \cdot B'_p$ .

The total weight  $w(c, h)$  for the edge between bioactive  $c$  and condition  $h$  is the sum of the weights from all the papers that discuss both  $c$  and  $h$ :  $w(c, h) = \sum_i w_i$ , where  $i$  indexes the papers that include information of both

the bioactive  $c$  and the health condition  $h$ . This formula ensures that each edge weight in the KG is a quantitative measure reflecting the strength and credibility of the linkage on the basis of accumulated evidence from the literature. This approach not only captures the diversity and depth of the studies linking bioactives to health outcomes, but also appropriately scales the influence of each study on the basis of its methodological rigor and scale.

*Health effect to food:* In our model, there is no direct quantifiable link between foods and health conditions. Instead, the relationships are established indirectly through bioactive compounds. The quantification of the impact of a food  $f$  on a health condition  $h$  is therefore derived by aggregating the effects of all compounds  $c$  present in  $f$  that are linked to  $h$ . This is represented by the score  $w(f, h)$ , which integrates the compound-specific weights from foods to conditions through compounds, as previously defined. The formula for  $w(f, h)$  is as follows:

$$w(f, h) = \sum_{c \in C_f} \left( w(c, h) \times w(c, f) \times \frac{1}{\log(n_c + 1)} \right),$$

where  $C_f$  denotes the set of all the compounds found in food  $f$ ,  $w(c, h)$  is the weight representing the evidence of health effects of compound  $c$  on condition  $h$ , and  $w(c, f)$  is the normalized content of compound  $c$  in food  $f$ . The factor  $\frac{1}{\log(n_c + 1)}$  adjusts for the rarity of the compound across the food database.

The importance of rarer compounds (i.e., those occurring in fewer foods) is emphasized because they are more distinctive to a specific food profile. This enhances the specificity of  $w(f, h)$  scores for unique health effects, similar to IDF in information retrieval, where rarity increases informativeness and prevents ubiquitous compounds from dominating aggregated weights. Our approach aligns with biomarker research showing that rare, specific compounds may provide superior discriminative power compared with common compounds [26, 27].

In the current schema, our research is focused specifically on indicating the effects of bioactive food components, rather than the overall effects of nutrition. Therefore, basic nutritional components, such as calories and macronutrients (fats, proteins, and carbohydrates), are not directly assessed in terms of their health effects. This focus allows an in-depth examination of the specific health effects of bioactives, but it also underlines that our results are based on a rather limited view of the nutritional profiles of the foods in question. Moreover, all associations between entities, foods, bioactives, and health conditions, remain guided by our unique methodological approach. The definitions we have given are not

based on a standard framework; instead, they have been developed from existing works in the scholarly literature, such as Cochrane reviews, adapted to reflect what we consider the most trustworthy approach given the complex interrelationships that exist between diet and health. The complex interactions between dietary practices and health outcomes, which are influenced by many external factors, offer ample scope for additional research.

## Utility and discussion

Having outlined the data collection, data pre-processing, data extraction, and data integration steps, we now layout the capabilities of the BASIL DB. Subsequently, we discuss some limitations of our approach and explore potential avenues for future work, including a focus on data quality.

### Overview

The BASIL DB is designed to enable users to navigate through arrays of scientific literature regarding bioactive compounds. This enhancement is intended to assist users in understanding and processing complex clinical data more efficiently, providing a clearer and faster route to information in both nutritional science and medical research. It has the potential to support dietary intake assessments, aid in nutritional education for consumers, provide healthcare professionals with robust data for dietary recommendations, capture information on synergistic compound combinations, and assist research on diet-health relationships.

The process of discovering compounds with previously unrecognized therapeutic potential can be approached through various methods such as biodiversity-based, chemo-systematic, ecological, computational, and ethnopharmacological approaches [28]. The BASIL DB was designed to enhance the computational [29] and ethnopharmacological [30] approaches by providing a robust dataset of bioactive compounds and their documented health effects, simplifying data-driven insights, validations, and the exploration of traditional medical practices through modern scientific studies. Compared to the random model, leveraging ethnomedical knowledge and computational strategies is more cost-effective and time-efficient [31] and can even yield a higher success rate in identifying promising compounds [32, 33].

The database was constructed with 433 compounds, 40296 papers, 7256 health effects, and 4197 food items, available for access at <https://basil-db.github.io/info.html>.

The 3 bioactive entities with most associated papers are “Carotene” (979), “Docosahexaenoic acid (DHA)” (958), and “Arachidonic acid” (937). The conditions with the most associated papers are “Pain” (706) “Symptoms” (653), “Hypotension” (581), and “Continuance of

**Table 2** Three exemplary evaluation entries for the compound “Curcumin”. Some columns, such as “Other Compounds” were excluded for simplicity, and *N* refers to the number of participants

PMID	Health effects	Effect source	Population	N	Bias
38794742	‘reducing disease activity’, ‘reducing inflammation’, ‘improving quality of life’	Combination	‘mild-to-moderate active ulcerative colitis’	N/A	0.042
37366197	‘symptom improvement’, ‘pain reduction’	Combination	‘patients with mild to moderate Carpal Tunnel Syndrome’	147	0.036
35178811	‘insulin resistance’, ‘erythrocyte sedimentation rate’, ‘inflammatory markers’, ‘triglycerides’, ‘weight’, ‘BMI’, ‘waist circumference’	Single	‘women with rheumatoid arthritis’	48	0.447

**Table 3** Concentration of the compound “Curcumin” in various foods

Food name	Amount	Unit
Indian Saffron (Rhizome)	2800.450	mg/100g
Turmeric, dried	2213.571	mg/100g
Curry, powder	285.263	mg/100g

life” (531). The 3 highest weighted edges between bioactives and condition nodes are “Vinblastine” and “Non-Small Cell Lung Carcinoma” (657), “Ephedrine” and “Hypotension” (607.9), and “Artemisinin” and “Malaria, Falciparum” (497.3). Excluding food nodes, the graph has 5 components, the largest of which contains 12,844 nodes, and the average degree is 24.5.

### Web interface

We provide a web interface to browse the database. Importantly, we host the data and code on GitHub allowing for open development. We now describe the features of the web interface.

### BASIL DB search

Using the search function, users can query the categories “Compound”, “Condition”, “Food”, and “Other”, where the latter refers to any entities that were identified by the NER pipeline but do not fall within the semantic types classifiable as “Health Effect”. Examples include geographic locations, temporal concepts, or occupational activities.

To illustrate, when querying “Curcumin” under the “Compound” category, users can explore its associations with various health conditions, such as inflammation or pain, discover foods high in “Curcumin”, and examine other contextual factors recognized by the system but not directly related to health effects. Tables 2, 3, and 4 provide detailed examples of such a query.

**Table 4** Top six conditions with the highest edge weight for the compound “Curcumin”

Condition	Weight	UMLS ID
Pain	128.63	C0030193
Inflammation	128.57	C0021368
Obesity	124.94	C0028754
Non-alcoholic Fatty Liver Disease	103.02	C0400966
Oxidative Stress	80.84	C0242606
Diabetes (Type)	73.23	C0011849

### Visualization

We provide a number of visualizations that show various aspects of our dataset. Figure 4 is a heatmap showing the connection strengths between bioactive compounds and health conditions; Fig. 5 displays a graph of compound-to-condition links, indicating effect similarities; Fig. 6 features a Sankey graph of food-compound-condition relationships; and Fig. 7 presents a multidimensional scaling (MDS) plot of compounds on the basis of the Tanimoto similarity.

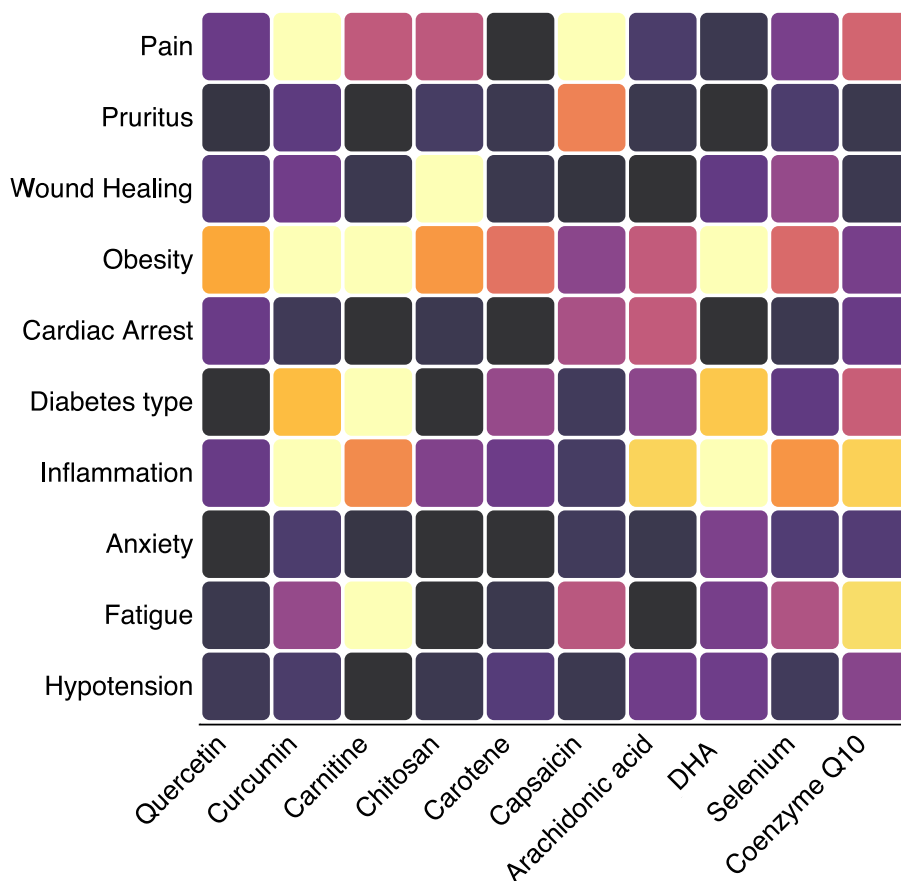
The Tanimoto coefficient, also known as the Jaccard index, is a measure of the similarity between two sets and is calculated via the following formula:

$$T(A, B) = |A \cap B| / |A \cup B|,$$

where *A* and *B* are sets of characteristics (in our case using the Simplified Molecular Input Line Entry System (SMILES)) for two different compounds. The coefficient ranges from 0 to 1, where 0 indicates no similarity and 1 indicates identical sets.

### Updates and scalability

Users have full access to the setup via GitHub, which they can fork and run independently, with the entire existing dataset available for download. Given the use of the described data sources in “Data collection” section users can easily scale the current version by adding new paper, bioactive, or condition entities. The



**Fig. 4** Heatmap presenting the connection strengths between 10 bioactive compounds and 10 health conditions, where lighter colors reflect a greater weight. For example, there seems comparably high evidence for the connection between “Capsaicin” and the concept “Pain” or “Curcumin” and the concept “Inflammation”

respective identifiers such as PubMed ID or UMLS concept ID can be used as criteria to avoid duplicating existing entries. For detailed instructions on configuring and running the setup, including versioning of the tools used, please refer to the GitHub repository.

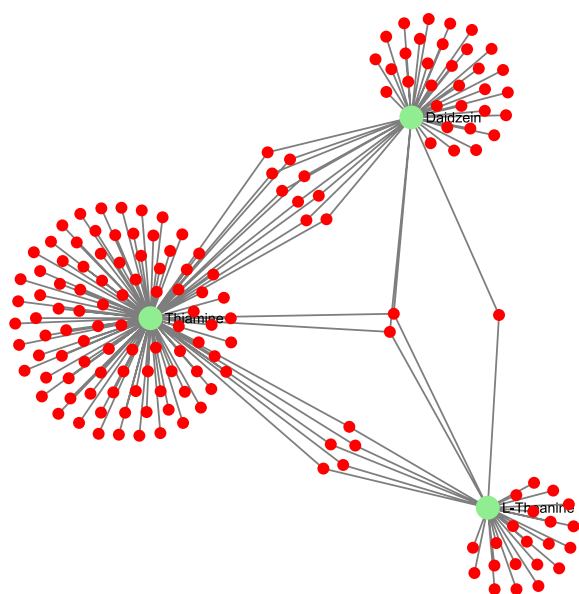
**Data extraction pipeline evaluation**

To evaluate the accuracy of our NLP-based pipeline, we conducted two complementary assessments: first, a manual review in which annotators examined a random sample of our outputs for correctness, and second, an automated benchmark against a publicly available, labeled dataset. For the manual review, we randomly selected 100 compound-paper pairs from our KG, ensuring that different compounds and effect sources were included. Two independent annotators were asked to assess the accuracy of the extracted information. For each compound-paper combination, annotators evaluated the following aspects:

- Compound involvement: Was the compound indeed studied in the trial? For example, was it actively investigated in the study rather than merely mentioned in the introduction?
- Effect source classification: Was the “Effect Source” classified correctly?
- Positivity of health effects: Were all positive health effects actually positive?
- Participant extraction: Was the number of participants correctly extracted?

To ensure the reliability of our evaluation, interannotator agreement was measured via Cohen’s Kappa, yielding scores of 0.78 for compound involvement, 0.78 for effect source classification, 0.82 for positivity of health effects, and 0.94 for participant extraction. Disagreements were resolved through discussion to establish a gold standard for final accuracy calculations.

For the assessment of whether the compound was actively studied in the trial, the pipeline achieved an



**Fig. 5** Graph linking compound nodes such as L-theanine or Dadzein (green) to various condition nodes (red). This visualization informs us of the effect similarity of two or more compounds. Users are able to adjust the minimum edge weight for overview

accuracy of 89% (89 out of 100 pairs correctly identified). Errors occurred in cases where compounds were mentioned in the introduction only for context, or compounds had a more complex function, for example, as a target for reduction rather than as a therapeutic agent.

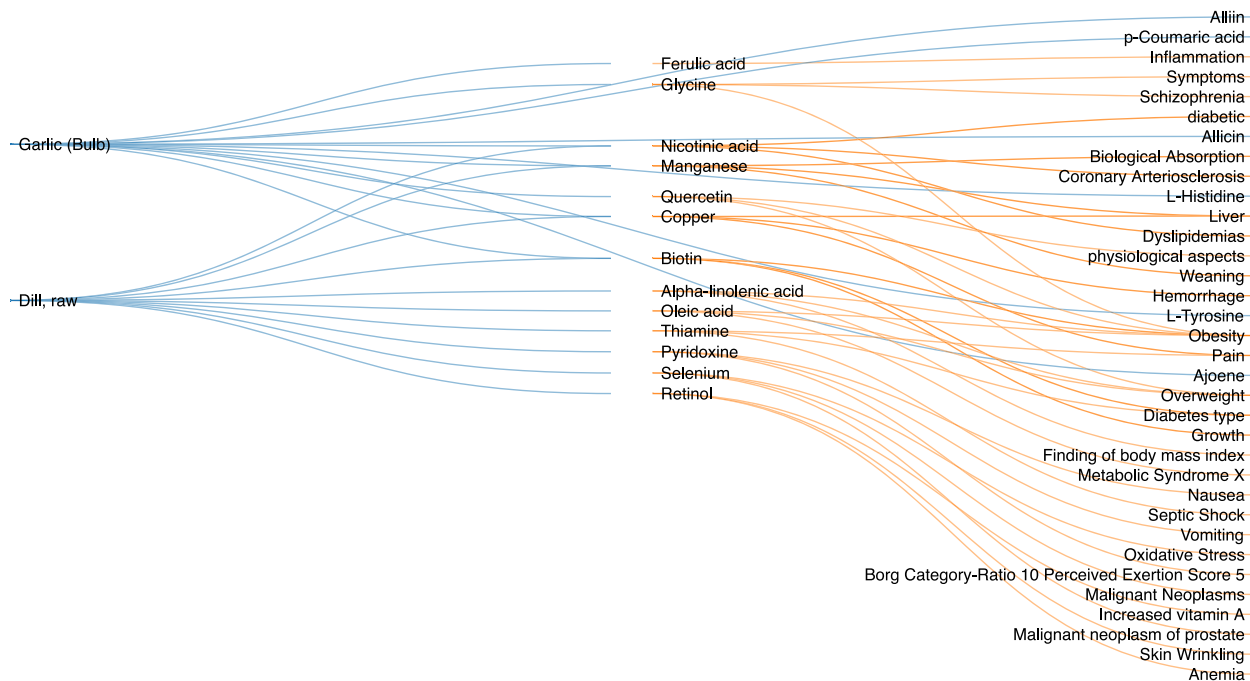
One included paper (PMID: 8741209) was a review, not an RCT. These cases were the most critical, as they could lead to follow-up errors.

The classification of “Effect Source” was correct in 80% of cases (80 out of 100). The most common error involved misclassification of “Combination” studies. For example, in study 34587702 the compound “monacolin K” was classified as “single”, although the study researched “red yeast rice with monacolin K”.

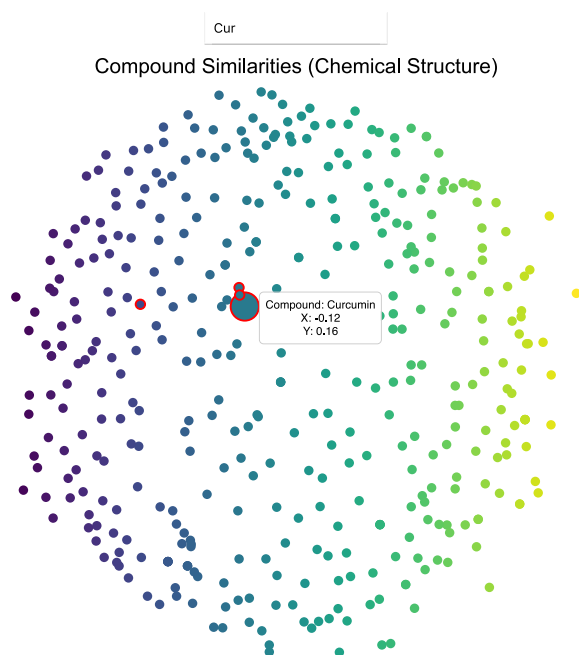
With respect to the classification of extracted health effects, three cases were incorrectly classified as positive, yielding a 97% accuracy. For instance, “increased cough reflex sensitivity” was classified as a beneficial effect of capsaicin in asthma patients, although this symptom is generally considered a negative outcome in asthma management.

The extraction of the number of participants was correct in 84% of the cases, with errors mainly stemming from abstracts that reported participant numbers ranges or subgroup breakdowns that RobotReviewer occasionally failed to parse correctly.

To further assess the generalizability of our pipeline in a fact-checking scenario, we evaluated our approach on the HealthFC dataset [34], which consists of 750 health-related claims labeled “health claim supported”, “health claim refuted”, and “not enough information”. Since our original pipeline focuses on extracting bioactive-specific information from RCT abstracts, we modified the LLM prompting strategy: instead of extracting health



**Fig. 6** Sankey graph displaying the connections between selected foods and their respective compounds and health effects. The blue links represent edges between foods and compounds, and the orange links represent edges between compounds and conditions



**Fig. 7** Multidimensional scaling (MDS) plot of the compound dataset as found on the BASIL DB website. Each pair of points is positioned to display, as accurately as possible, the dissimilarity between those two samples based on their Tanimoto similarity

benefits for predefined compounds, the prompts were adjusted to retrieve health claims from the HealthFC texts with predefined interventions (e.g., “masks”) and classify the respective claim as “supported”, “refuted”, or “inconclusive”.

Using the dataset’s labels as ground truth, the pipeline achieved an overall accuracy of 74.8% in claim verification (561 out of 750 correctly classified). The most common error type was the classification of inconclusive claims as supported claims, pointing to an overreliance on positive outcomes. These results demonstrate the pipeline’s ability to extract health effects, even for broader health fact-checking tasks, although further refinement could improve handling of ambiguous or multifaceted claims.

For more detailed evaluations of the individual NLP tools, we point to previous literature assessing the performance of LLMs [35], RobotReviewer [18], and MetaMap [36]. We provide further discussion in “[Limitations and future work](#)” section.

### Limitations and future work

By using various tools and sources for the creation of the BASIL DB, the quality of the final output is inherently dependent on the precision of these methods and the quality of the underlying data. In utilizing LLMs to extract the purported health effects of compounds from

scientific studies, we must acknowledge the limitations inherent in non-expert systems interpreting nuanced medical data. Assessing and improving the ability of these systems to understand the nuanced implications of clinical terms and relationships represents a critical area of future work [37, 38].

While developing BASIL, we frequently encounter data quality challenges commonly faced by large-scale medical datasets, including incorrectness, incompleteness, lack of standardization, and other issues that can compromise the robustness of data-driven insights. These problems are not unique to this work, as the literature shows that medical big data often suffer from errors, missing information, and inconsistencies [39]. For example, in PubMed, errors may arise from incorrect indexing of articles or misclassification of research topics. A notable case is the article with PMID 38474802, which was incorrectly classified as a randomized controlled trial despite not meeting the criteria for such a designation. Additionally, indexing issues are evident in the case of “Camptothecin”, where over 800 of the retrieved articles were tagged with this MeSH term, yet none of them discussed “Camptothecin” as a single compound, and 711 were classified as either “None” or “Derivative”. Such classification issues can mislead researchers and reduce the precision and efficiency of search results. In the BASIL DB, one can filter such instances by simply using the “Effect Source” feature.

Another illustrative example is the study with PMID 17939194, which highlights several data quality issues in PubMed. First, the article is translated from Russian, and the translation contains grammatical errors and incorrect drug spellings (e.g., “Metoclopramid” instead of the correct “Metoclopramide”). While this particular spelling error may not lead to significant confusion, similar errors involving drugs with nearly identical names could have serious consequences, such as misidentification of medications in clinical or research settings. Furthermore, the abstract is incomplete and ambiguous, lacking critical details about the study’s methodology, results, and conclusions. For instance, it mentions “vegetotropic therapy” and “hypobaric hypoxic adaptation” without providing sufficient context or definitions, making it difficult for researchers to assess the study’s relevance or validity. Compounding these issues, the full text of the article is unavailable in public or academic databases, limiting the ability of researchers to verify or build upon its findings.

The study is also indexed with a large number of MeSH terms, some of which may not be directly relevant to its focus or are redundant. For example, the terms “Biliary Dyskinesia/diagnostic imaging”, “Biliary Dyskinesia/drug therapy”, “Biliary Dyskinesia/physiopathology”, and “Biliary Dyskinesia/therapy” are all indexed separately, even

though they could potentially be consolidated under a single term like “Biliary Dyskinesia/therapy” to avoid redundancy. Over-indexing with broad or irrelevant terms can dilute the relevance of search results and make it harder for researchers to find studies that are truly focused on their topic of interest.

Furthermore, occurrences of empty abstracts (e.g., PMIDs 4922015, 19701267, 2573775) or incomplete abstracts (likely due to data entry errors, as seen in PMIDs 33075061, 33641356, 30453844) are common in PubMed. These issues hinder the ability of researchers to assess the relevance of studies quickly. Variations in terminology across different journals and disciplines [40] further exacerbate the lack of standardization, potentially leading to inconsistencies in article retrieval and cataloging. For instance, the Unified Medical Language System (UMLS), which integrates over 160 source vocabularies [41], may contain variations in definitions and usage that are not fully consistent. This can affect the accuracy of semantic searches and data integration tasks. An example is the distinction between terms like “Diabetes Mellitus” and “Diabetes Type”. While these terms serve different semantic roles, their overlapping usage in the literature can still lead to challenges in accurately retrieving and integrating relevant studies, particularly when automated systems such as MetaMap fail to account for their contextual differences.

Table 4 reveals a well-documented challenge in the biomedical knowledge graph literature, where generic health concepts such as “pain” can become high-degree nodes that may hinder knowledge discovery efficiency [42]. Several methodological approaches have been developed to address this problem. Van Haagen et al. [43] proposed systematic filtering on the basis of concept frequency thresholds and semantic specificity metrics, while recent work has employed advanced harmonization methods [44] and Kullback-Leibler Divergence reranking to weight concepts on the basis of their disproportionate representation in condition-specific versus general biomedical literature [45]. Future iterations of the BASIL DB could incorporate these methodologies to distinguish between general terms and specific health conditions. Additionally, MetaMap’s confidence scores can be leveraged to downweight low-confidence generic extractions. Finally, our separation of health effect and population terms may provide some mitigation by adding contextual specificity to generic concepts. For example, in PMID 29908031, while “pain” appears as a health effect, it is qualified by specific population descriptors including “post-surgical” and “impacted third molars”, enabling more targeted queries than the generic concept alone would allow.

FooDB aggregates food composition data from diverse sources, including scientific literature, government

databases, and industry reports. While this diversity enriches the dataset, it also introduces challenges such as inconsistent measurement methods, varying units, and terminological differences. As shown in “Data pre-processing” section, there are significant discrepancies between values reported from different sources, such as the quercetin content in “European cranberry”, which varies widely from 0.515 mg/100 g to 16.392 mg/100 g. Additionally, FooDB may not cover all food varieties, such as differences in nutrient content based on cultivation practices, geographical origin, or processing methods.

Having outlined some of the data quality challenges, we aim to address these issues more comprehensively in future work. This could involve developing more sophisticated data cleaning and harmonization techniques, as well as exploring advanced methods for integrating diverse data types to ensure comprehensive and accurate data representation. Strategies for improving data quality in health research typically involve the use of business intelligence models, statistical analyses, data mining techniques, and qualitative approaches [46].

Finally, future work could extend the “knowledge as a product” strategy by expanding the database to a wider range of papers, compounds, foods, and other entities representative of the domain. As only RCT abstracts were considered, a variety of study types such as cohort or case studies, as well as full trial texts, could increase cohesiveness of our approach. This could be further developed to include other forms of interdisciplinary knowledge, such as sustainability, market trends, or technology development, and create even more opportunity for innovation.

## Conclusion

In this work, we introduced the BASIL DB, a novel knowledge graph database for enhancing the understanding and analysis of bioactive compounds and their effects on health. By integrating data from several databases such as FooDB, PhenolExplorer, ChEMBL, and PubMed, and using advanced NLP techniques, BASIL DB provides a formalized and semantically enriched form for exploring the relationships among bioactive compounds, foods, and diseases. The database uses automatic data extraction and integration techniques to overcome the limitations of data heterogeneity, nonstandardization, and the sheer amount of scientific literature in the discipline of nutritional science.

The BASIL DB provides a number of important capabilities, such as querying and visualizing of intricate relationships between bioactive compounds and health effects, assisting in dietary intake evaluations, and promoting nutritional education and research. The web interface of the database enables navigation of the relationships

between compounds, foods, and diseases, whereas the visualization tools give an idea of the strength and direction of the relationships. Containing more than 43,000 papers, 433 compounds, and 7,256 health effects incorporated into the KG, BASIL DB is a significant leap forward in the computational exploration of bioactive compounds and their prospective health effects.

However, our work is not without limitations. The quality of the data in the BASIL DB is inherently dependent on the accuracy of the underlying sources and the tools used for data extraction and integration. Challenges such as incorrect indexing in PubMed, incomplete abstracts, and inconsistencies in terminology across different databases highlight the need for continued improvements in data quality and harmonization. Future work will focus on addressing these issues through more sophisticated data cleaning techniques, the inclusion of additional study types (e.g., cohort studies, case reports), and the expansion of the database to encompass a wider range of bioactive compounds, foods, and health conditions. Additionally, further research is needed to refine the interpretation of clinical data by NLP systems, particularly in understanding nuanced medical terms and relationships.

Despite these challenges, the BASIL DB represents a valuable resource for researchers, healthcare professionals, and consumers interested in the health effects of bioactive compounds. By providing a comprehensive and accessible platform for exploring the complex relationships between diet and health, the BASIL DB has the potential to support evidence-based dietary recommendations, identify novel therapeutic compounds, and advance our understanding of the role of bioactives in health maintenance and disease prevention. As the field of nutritional science continues to evolve, tools such as the BASIL DB will play an increasingly important role in bridging the gap between research and practical applications in health and nutrition [47].

In conclusion, BASIL DB is a step towards a more systematic and data-driven approach to understanding the health effects of bioactive compounds. By leveraging the power of knowledge graphs and NLP, we aim to provide a robust platform for researchers and practitioners to explore the complex interplay between diet, bioactive compounds, and health outcomes. Future enhancements to the database include improved data quality and expanded coverage.

#### Abbreviations

BASIL DB	BioActive Semantic Integration and Linking Database
FooDB	Food Database
KG	Knowledge Graph
LLM	Large Language Model
MDS	Multidimensional Scaling
NER	Named Entity Recognition
NLP	Natural Language Processing

PMID	PubMed Identifier
RCT	Randomized Controlled Trial
SMILES	Simplified Molecular Input Line Entry System
UMLS	Unified Medical Language System

#### Acknowledgements

Not applicable.

#### Authors' contributions

DJ conceived and led the development of BASIL DB, designing the methodology, implementing the data collection, extraction, and integration processes, and drafting the manuscript. PG and HH provided critical supervision throughout the project, offering guidance on the research design, technical implementation, and data interpretation, while also contributing to the writing and finalization of the manuscript. All authors read and approved the final version of the manuscript.

#### Funding

The authors received no specific funding for this work.

#### Data availability

Project name: BASIL DB.  
Project home page: <https://basil-db.github.io/info.html>.  
Archived version: 1.1.0.  
Operating system(s): Platform independent.

#### Declarations

##### Ethics approval and consent to participate

Not applicable, as this study uses publicly available, de-identified data from existing literature.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare no competing interests.

##### Author details

<sup>1</sup>University of Amsterdam, Amsterdam, The Netherlands.

Received: 5 March 2025 Accepted: 6 August 2025

Published online: 13 August 2025

#### References

- Šaponjac V, Čanadanović Brunet J, Četković G, Djilas S. Detection of bioactive compounds in plants and food products. In: Nedović V, Raspor P, Lević J, Tumbas Šaponjac V, Barbosa-Cánovas GV, editors. Emerging and traditional technologies for safe, healthy and quality food. Cham: Springer; 2016. pp. 81–109. [https://doi.org/10.1007/978-3-319-24040-4\\_6](https://doi.org/10.1007/978-3-319-24040-4_6).
- Miles E, Calder P. Effects of citrus fruit juices and their bioactive components on inflammation and immunity: a narrative review. *Front Immunol*. 2021;12: 712608. <https://doi.org/10.3389/fimmu.2021.712608>.
- van Breda S, de Kok T. Smart combinations of bioactive compounds in fruits and vegetables may guide new strategies for personalized prevention of chronic diseases. *Mol Nutr Food Res*. 2018;62(1):1700597. <https://doi.org/10.1002/mnfr.201700597>.
- Kussmann M, Abe Cunha D, Berciano S. Bioactive compounds for human and planetary health. *Front Nutr*. 2023;10: 1193848. <https://doi.org/10.3389/fnut.2023.1193848>.
- Jacobs Jr DR. Challenges in research in nutritional epidemiology. In: Temple NJ, Wilson T, Jacobs Jr DR, Bray GA, editors. Nutritional health: strategies for disease prevention. 4th ed. Cham: Springer; 2023. pp. 21–31. [https://doi.org/10.1007/978-3-031-24663-0\\_2](https://doi.org/10.1007/978-3-031-24663-0_2).
- Stern D, Ibsen D, MacDonald C, Chiu Y, Lajous M, Tobias D. Improving nutrition science begins with asking better questions. *Am J Epidemiol*. 2024;193(11):1507–10. <https://doi.org/10.1093/aje/kwae085>.

7. Yasmeen R, Fukagawa N, Wang T. Establishing health benefits of bioactive food components: a basic research scientist's perspective. *Curr Opin Biotechnol*. 2017;44:109–14. <https://doi.org/10.1016/j.copbio.2016.11.002>.
8. Hunter L, Cohen K. Biomedical language processing: what's beyond PubMed? *Mol Cell*. 2006;21(5):589–94. <https://doi.org/10.1016/j.molcel.2006.02.012>.
9. Tamašauskaitė G, Groth P. Defining a knowledge graph development process through a systematic review. *ACM Trans Softw Eng Methodol*. 2023;32(1):1–40. <https://doi.org/10.1145/3522586>.
10. Wang C, Zheng Z, Cai X, Huang J, Su Q. Overview of the application of knowledge graphs in the medical field. *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi*. 2023;40(5):1040–4.
11. Wishart D, Guo A. FoodDB: The Food Database. 2018. <https://foodb.ca>. Accessed 6 Dec 2024.
12. Neveu V, Perez-Jimenez J, Vos F, Crespy V, du Chaffaut L, Mennen L, et al. Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. *Database*. 2010;2010:baq027. <https://doi.org/10.1093/database/baq027>.
13. Gaulton A, Bellis L, Bento A, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 2012;40(D1):D1100–7. <https://doi.org/10.1093/nar/gkr777>.
14. Plumb J, Pigat S, Bompola F, Cushen M, Pinchen H, Norby E, et al. eBASIS (Bioactive Substances in Food Information Systems) and bioactive intakes: major updates of the bioactive compound composition and beneficial bioeffects database and the development of a probabilistic model to assess intakes in Europe. *Nutrients*. 2017;9(4):320. <https://doi.org/10.3390/nu9040320>.
15. Egger M, Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629–34. <https://doi.org/10.1136/bmj.315.7109.629>.
16. National Center for Biotechnology Information (NCBI). E-utilities: E-fetch; 2024. <https://www.ncbi.nlm.nih.gov/books/NBK25500/>. Accessed 6 Dec 2024.
17. Murad M, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *Evid Based Med*. 2016;21(4):125–7. <https://doi.org/10.1136/ebmed-2016-110401>.
18. Marshall I, Kuiper J, Wallace B. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc*. 2016;23(1):193–201. <https://doi.org/10.1093/jamia/ocv070>.
19. Aronson A. Metamap: mapping text to the UMLS metathesaurus. Bethesda: National Library of Medicine, NIH, DHHS; 2006.
20. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F, et al. GPT-4 technical report. 2023. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
21. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models in medical applications: a survey. 2024. [arXiv:2405.07468](https://arxiv.org/abs/2405.07468).
22. Higgins JPT, Green S, editors. *Cochrane handbook for systematic reviews of interventions*. Chichester: John Wiley & Sons; 2008. p. 649. <https://doi.org/10.1002/9780470712184>. ISBN: 978-0-470-69951-5
23. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(Suppl 1):D267–70. <https://doi.org/10.1093/nar/gkh061>.
24. Hopp L. Risk of bias reporting in Cochrane systematic reviews. *Int J Nurs Pract*. 2015;21(5):683–6. <https://doi.org/10.1111/ijn.12352>.
25. Boeije H, van Wesel F, Alisic E. Making a difference: towards a method for weighing the evidence in a qualitative synthesis. *J Eval Clin Pract*. 2011;17(4):657–63. <https://doi.org/10.1111/j.1365-2753.2010.01595.x>.
26. Qiu S, Cai Y, Yao H, Lin C, Xie Y, Tang S, et al. Small molecule metabolites: discovery of biomarkers and therapeutic targets. *Signal Transduct Target Ther*. 2023;8(1): 132.
27. Hedrick VE, Dietrich AM, Estabrooks PA, Savla J, Serrano E, Davy BM. Dietary biomarkers: advances, limitations and future directions. *Nutr J*. 2012;11(1):109.
28. Naghizadeh A, Salamat M, Hamzeian D, Akbari S, Rezaeizadeh H, Vaghaslou M, et al. IRGO: Iranian traditional medicine general ontology and knowledge base. *J Biomed Semant*. 2021;12:1–11. <https://doi.org/10.1186/s13326-021-00237-8>.
29. Basith S, Cui M, Macalino S, Choi S. Expediting the design, discovery and development of anticancer drugs using computational approaches. *Curr Med Chem*. 2017;24(42):4753–78. <https://doi.org/10.2174/0929867324666170725103520>.
30. Fakhrudin N, Waltenberger B, Cabaravdic M, Atanasov A, Malainer C, Schachner D, et al. Identification of plumericin as a potent new inhibitor of the NF- $\kappa$ B pathway with anti-inflammatory activity in vitro and in vivo. *Br J Pharmacol*. 2014;171(7):1676–86. <https://doi.org/10.1111/bph.12558>.
31. Lin X, Li X, Lin X. A review on applications of computational methods in drug screening and design. *Molecules*. 2020;25(6):1375. <https://doi.org/10.3390/molecules25061375>.
32. Pan S, Zhou S, Gao S, Yu Z, Zhang S, Tang M, et al. New perspectives on how to discover drugs from herbal medicines: CAM's outstanding contribution to modern therapeutics. *Evid Based Complement Alternat Med*. 2013;2013(1):627375. <https://doi.org/10.1155/2013/627375>.
33. Gyllenhaal C, Kadushin M, Southavong B, Sydara K, Bouamanivong S, Xaiveu M, et al. Ethnobotanical approach versus random approach in the search for new bioactive compounds: support of a hypothesis. *Pharm Biol*. 2012;50(1):30–41. <https://doi.org/10.3109/13880209.2011.634423>.
34. Vladika J, Schneider P, Matthes F. HealthFC: Verifying health claims with evidence-based medical fact-checking. 2023. arXiv preprint [arXiv:2309.08503](https://arxiv.org/abs/2309.08503).
35. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak*. 2025;25(1):117.
36. Jing X, Indani A, Hubig N, Min H, Gong Y, Cimino JJ, et al. A systematic approach to configuring MetaMap for optimal performance. *Methods Inf Med*. 2022;61(S 02):51–63.
37. Vladika J, Schneider P, Matthes F. MedREQAL: examining medical knowledge recall of large language models via question answering. 2024. [arXiv:2406.05845](https://arxiv.org/abs/2406.05845).
38. Han T, Kumar A, Agarwal C, Lakkaraju H. MedSafetyBench: evaluating and improving the medical safety of large language models. In: *Advances in neural information processing systems 37: Proceedings of the thirty-eighth conference on neural information processing systems; Vancouver, Canada: Neural Information Processing Systems Foundation; 2024.* [https://proceedings.neurips.cc/paper\\_files/paper/2024/hash/3ac952d0264ef7a505393868a70a46b6-Abstract-Datasets\\_and\\_Benchmarks\\_Track.html](https://proceedings.neurips.cc/paper_files/paper/2024/hash/3ac952d0264ef7a505393868a70a46b6-Abstract-Datasets_and_Benchmarks_Track.html).
39. Hoffman S. Medical big data and big data quality problems. *Conn Insur Law J*. 2014;21:289.
40. Parker R, Hayden J. Uncommon language: the challenges of inconsistent terminology use for evidence synthesis. 2011. *Cochrane Colloquium Abstracts*. <https://abstracts.cochrane.org/2011-madrid/uncommon-language-challenges-inconsistent-terminology-use-evidence-synthesis>. Accessed 6 Dec 2024.
41. Bodenreider O. Multi-lingual features of the Unified Medical Language System. In: *CLEF (Working Notes)*. 2013. Accessed 6 Dec 2024.
42. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J*. 2020;18:1414–28.
43. van Haagen HH, 't Hoen PA, Mons B, Schultes EA. Generic information can retrieve known biological associations: implications for biomedical knowledge discovery. *PLoS ONE*. 2013;8(11):e78665.
44. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. *Sci Data*. 2023;10(1):67.
45. Schäfer H, Idrissi-Yaghir A, Arzideh K, Damm H, Pakull TM, Schmidt CS, et al. Biokgrapher: initial evaluation of automated knowledge graph construction from biomedical literature. *Comput Struct Biotechnol J*. 2024;24:639–60.
46. Bernardi F, Alves D, Crepaldi N, Yamada D, Lima V, Rijo R. Data quality in health research: integrative literature review. *J Med Internet Res*. 2023;25:e41446. <https://doi.org/10.2196/41446>.
47. Okolo C, Chidi R, Babawarun O, Arowoogun J, Adeniyi A. Data-driven approaches to bridging the gap in health communication disparities: a systematic review. *World J Adv Res Rev*. 2024;21(2):1435–1445. <https://doi.org/10.30574/wjarr.2024.21.2.0451>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.