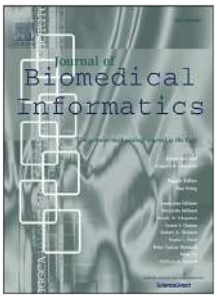
Journal Pre-proof



Creating an ignorance-base: Exploring known unknowns in the scientific literature

Mayla R. Boguslav, Nourah M. Salem, Elizabeth K. White, Katherine J. Sullivan, Michael Bada, Teri L. Hernandez, Sonia M. Leach, Lawrence E. Hunter

PII: S1532-0464(23)00126-0
DOI: https://doi.org/10.1016/j.jbi.2023.104405
Reference: YJBIN 104405

To appear in: *Journal of Biomedical Informatics*

Received date : 8 November 2022
Revised date : 18 May 2023
Accepted date : 21 May 2023

Please cite this article as: M.R. Boguslav, N.M. Salem, E.K. White et al., Creating an ignorance-base: Exploring known unknowns in the scientific literature, *Journal of Biomedical Informatics* (2023), doi: https://doi.org/10.1016/j.jbi.2023.104405.

Highlights

Creating an Ignorance-Base: Exploring Known Unknowns in the Scientific Literature

Mayla R. Boguslav, Nourah M. Salem, Elizabeth K. White, Katherine J. Sullivan, Michael Bada, Teri L. Hernandez, Sonia M. Leach, Lawrence E. Hunter

- We created the first ignorance-base (knowledge-base) to capture goals for scientific knowledge
- Our exploration methods provide analyses, summaries, and visualizations based on a query
- Ignorance enrichment provided fruitful avenues for future research
- Exploration by topic in vitamin D found three avenues to explore
- Exploration by experimental results for vitamin D and preterm birth found an emerging topic

Creating an Ignorance-Base: Exploring Known Unknowns in the Scientific Literature

Mayla R. Boguslav[a,*], Nourah M. Salem[a], Elizabeth K. White[a,b], Katherine J. Sullivan[a], Michael Bada[a], Teri L. Hernandez[c], Sonia M. Leach[a,b], Lawrence E. Hunter[a]

[a] *Computational Bioscience Program, University of Colorado Anschutz Medical Campus, E 17th Avenue, Aurora, 80045, CO, USA*
[b] *Center for Genes, Environment and Health, National Jewish Health, Jackson Street, Denver, 80206, CO, USA*
[c] *College of Nursing, Department of Medicine/Division of Endocrinology, Metabolism, Diabetes, University of Colorado Anschutz Medical Campus, E 17th Avenue, Aurora, 80045, CO, USA*

Abstract

Background: Scientific discovery progresses by exploring new and uncharted territory. More specifically, it advances by a process of transforming unknown unknowns first into known unknowns, and then into knowns. Over the last few decades, researchers have developed many knowledge bases to capture and connect the knowns, which has enabled topic exploration and contextualization of experimental results. But recognizing the unknowns is also critical for finding the most pertinent questions and their answers. Prior work on known unknowns has sought to understand them, annotate them, and automate their identification. However, no knowledge-bases yet exist to capture these unknowns, and little work has focused on how scientists might use them to trace a given topic or experimental result in search of open questions and new avenues for exploration. We show here that a knowledge base of unknowns can be connected to ontologically grounded biomedical

* *Email address:* Mayla.Boguslav@CUAnschutz.edu (Mayla R. Boguslav)

knowledge to accelerate research in the field of prenatal nutrition.

Results: We present the first ignorance-base, a knowledge-base created by combining classifiers to recognize ignorance statements (statements of missing or incomplete knowledge that imply a goal for knowledge) and biomedical concepts over the prenatal nutrition literature. This knowledge-base places biomedical concepts mentioned in the literature in context with the ignorance statements authors have made about them. Using our system, researchers interested in the topic of vitamin D and prenatal health were able to uncover three new avenues for exploration (immune system, respiratory system, and brain development) by searching for concepts enriched in ignorance statements. These were buried among the many standard enriched concepts. Additionally, we used the ignorance-base to enrich concepts connected to a gene list associated with vitamin D and spontaneous preterm birth and found an emerging topic of study (brain development) in an implied field (neuroscience). The researchers could look to the field of neuroscience for potential answers to the ignorance statements.

Conclusion: Our goal is to help students, researchers, funders, and publishers better understand the state of our collective scientific ignorance (known unknowns) in order to help accelerate research through the continued illumination of and focus on the known unknowns and their respective goals for scientific knowledge.

Keywords: Natural Language Processing, Knowledge Representation, Knowledge-base, Information Extraction, Epistemology

1. Introduction

Research begins with a question. It progresses through accumulating knowledge such that a previously unexplored subject (an unknown unknown) becomes an active research area exploring the questions (known unknowns), until a body of established facts emerges (known knowns) [1, 2, 3]. We aim to help illuminate this process using biomedical natural language process-

ing (BioNLP) to identify, categorize, classify, and explore known unknowns while highlighting their entailed goals for scientific knowledge (*i.e.*, actionable next steps). These known unknowns are discussed in the scientific literature as statements about knowledge that does not exist yet, including goals for desired knowledge, statements about uncertainties in the interpretation of results, discussions of controversies, and many others; collectively we call them **statements of ignorance**, borrowing the term from Firestein [1] and our prior work [4]. Our goal is to help researchers find the most pertinent questions to ask. For example, "these inconsistent observations point to the complicated role of vitamin D in the immune modulation and disease process" (PMC4889866) is a statement of ignorance. The entailed knowledge goal is to determine the correct role of vitamin D in the immune modulation and disease process by creating novel methods or conducting new experiments to study the complicated role. We also used **biomedical concept recognition**, the identification of biomedical vocabulary terms from ontologies or controlled vocabulary in text, to understand the biomedical subjects of these known unknowns. In the above example, these concepts include "vitamin D" and "immune". Therefore, we aim to reveal these statements of ignorance, the entailed knowledge goals, and the entailed biomedical concepts to help students, researchers, funders, and publishers better understand the state of our collective scientific knowledge and **ignorance** (known unknowns).

While these ideas and methods are generally applicable across biomedical research, we chose to focus on the prenatal nutrition field. Due to ethical and legal considerations and complexities in studying pregnant mothers and fetuses, the field of prenatal nutrition is understudied and poised to benefit from the identification of questions that are well studied in other fields [5, 6, 7, 8]. Fetal development is a critical period and exposure to nutrition has a lifelong impact [9]. For example, the micronutrient vitamin D is very important for maternal and fetal health, affecting the immune and musculoskeletal systems, neurodevelopment, and hormones [10, 11, 12, 13, 14] (see

Figure 1). Abnormal vitamin D levels can lead to gestational diabetes mellitus, preterm delivery, frequent miscarriages, adipogenesis, pre-eclampsia, obstructed labor, Cesarean sections, reduced weight at birth, respiratory issues, postpartum depression, and autism [10]. If we can identify the known unknowns or questions raised, even just with regard to the role of vitamin D, then we can search other fields for answers to inform the design of future studies. The prenatal nutrition field is a good case study for these ideas because it contains a diverse literature and a variety of studies from all over the world. Thus, applying an ignorance-based approach to this area is likely to generalize beyond prenatal nutrition, and more specifically to facilitate new interdisciplinary interactions that could advance the study of an underserved population and potentially help accelerate research to benefit mothers everywhere.

This work provides the necessary methods and tools to create a knowledge-base containing representations of known unknowns and their associated knowledge goals, an **ignorance-base**. This architecture allows scientists to explore the landscape of ignorance around a topic or a set of experimental results at scale and to find insights about related concepts across disciplines, resulting in an accelerated and interdisciplinary research process. (A list of the formal terms we have introduced here and their definitions are shown in Table 1.) We highlight its power by providing analyses, summaries, and visualizations that help researchers find knowledge goals to pursue in future work. Such an automated system could be useful to a wide variety of scientific stakeholders ranging from graduate students looking for thesis projects (*e.g.*, [15]) to funding agencies tracking emerging research areas (*e.g.*, [16]). It could help facilitate interdisciplinary interactions amongst researchers by finding questions from another field that bear on a topic or a set of experimental results (*e.g.*, [17]). It could also help track the evolution of research questions over time as a longitudinal analysis (*e.g.*, [18]). Furthermore, automatically identifying questions would allow us to query existing databases for

relevant information (*e.g.*, [19]). Thus, there is a need for such an automated system to capture questions or known unknowns.

There is only one similar system, the COVID-19 challenges and direction search engine (COVID-19 search engine) developed by Lahav *et al.*, [20]. They focused on creating a search engine to help researchers find two known unknown categories, scientific challenges and directions, and compared their work to a standard PubMed search. The search engine provides a relevant (high-confidence) table of challenge or direction sentences based on an input query of Medical Subject Headings (MeSH) terms. MeSH is a controlled vocabulary that is part of the Unified Medical Language System from NLM used for indexing, cataloguing, and searching for biomedical information and documents [21]). Their known unknown categories were motivated by the fact that "research focuses on *fine-grained* specific challenges, *e.g.*, difficulties in functional analysis of specific viral proteins, or shortcomings of a specific treatment regime for children. Each challenge, in turn, is associated with potential directions and hypotheses" [20]. However, their work stopped at identification of such statements and did not identify the knowledge goals associated with them; in contrast, our explicit representations of knowledge goals provide the users with guidelines for next research steps. In addition, we address a broader set of known unknowns and knowledge goals.

The other main goal of our work is to provide scientists with tools to explore the landscape of ignorance surrounding a topic or set of experimental results. The COVID-19 search engine [20] can support queries of multiple MeSH terms, but does not permit investigation of other types of inputs such as experimental results. As for the output, their search engine did not go beyond the identification of relevant sentences to provide analyses, summaries, or visualizations. This limits the ways users can explore the outputs to prioritize relevant areas of research. Lahav *et al.*, [20] posed as future work "to build more tools to explore and visualize challenges and directions across science." Thus, their system could benefit from prior work focused on knowl-

edge goals [4], the addition of other input types such as experimental results, and methods to explore and visualize known unknowns.

Our work extends the functionality described in [20] by means of the first ignorance-base. We compare our ignorance approach to the COVID-19 search engine [20] and standard methods. Adding in prior work [4] that identifies statements about unknowns based on their entailed knowledge goals (**ignorance taxonomy**) provides actionable next steps for the users. For instance, we describe the specific challenges discussed above as *difficult tasks*, with the corresponding knowledge goals to create new tools or methods to overcome the difficulties and shortcomings. We create the ignorance-base based on this ignorance taxonomy by extending our prior work [4] to create more robust and high-quality classifiers of ignorance statements. Like many other knowledge modelers [22], we chose to ground our biomedical concepts in the open biomedical ontologies (OBOs) [23, 24] instead of MeSH for reproducibility, interoperability, and to avoid pitfalls in the modeling of knowledge [22]. Previous work using these ontologies has yielded state-of-the-art biomedical concept classifiers [25]. As a result of interweaving the ignorance-base with biomedical concept expansion, our work supports researchers in querying the literature for known unknowns, either by topic or with a list of experimental results, and then connecting this work to other knowledge-bases (*e.g.*, PheKnowLator [26, 27]) to find additional information relevant to the knowledge goals. Our system's ability to perform concept enrichment and ignorance classification simultaneously extends its reach far beyond prior work [20], allowing it to trace out connections across different publications and knowledge-bases for a more comprehensive picture of what is known and unknown about a given subject.

The second goal of our work is to extend prior work [20] by adding in analyses, summaries, and visualizations of the outputs to help a researcher find knowledge goals to pursue. To do so, we explored the most frequent and enriched biomedical concepts in the ignorance statements returned by the

input query in comparison to all sentences. This helped narrow researchers' search for a topic in vitamin D to find ignorance statements ripe for exploration with the concepts "feeding behavior", "immune system", "brain development", and "respiratory system" (see Figure 1). To help the researchers understand the general landscape of unknowns surrounding the topic of vitamin D and focus on the most interesting types, we summarized the ignorance categories around the enriched concepts and mapped out how they changed over time. We extended our prior work [4] by defining a total of 13 ignorance categories in the literature. For example, the researchers could choose a topic that is a complete unknown (*indication of unknown or novel research topic or assertion*) or a topic where there were alternate existing hypotheses to explore (*indication of alternative research options or controversy of research*). We demonstrate how this approach can help track the emergence of new research areas or produce a longitudinal analysis showing how research questions evolve over time. This is informative not only for scientists, but also for funding agencies and publishers [18, 28, 29, 30, 31, 32, 16, 33, 34, 35].

Once researchers have chosen a topic and want to evaluate experimental results, our goal is to help them contextualize those results in terms of statements of ignorance and understand what questions may bear on them, either within the same field as the topic or outside it. To do this, we conducted the same analyses as the input topic but also added canonical analyses based on the experimental results. Our motivating example was a gene list connecting vitamin D and spontaneous preterm birth (sPTB) from the literature [36]. If vitamin D plays a role in preventing sPTB, it would be relevant to all women of childbearing age. By comparing our ignorance approach to the standard approach for a gene list (functional enrichment analysis, gene list coverage, and the findings from the paper), we found ignorance enrichment of the concept "immune system", a topic also identified by the original authors, as well as a novel putative relationship with the concept "brain development", which implicates the field of neuroscience as a place to look for answers. We provide

the ignorance statements and suggest questions for future exploration.

Table 1: Term definitions.

Term	Definition
Ignorance	community/collective/scientific known unknowns
Knowledge-base	a database of known information
Ignorance-base	a knowledge-base, created from the literature, with additional annotations for the sentences that are ignorance statements
Statements of ignorance	statements of incomplete or missing knowledge categorized based on the entailed knowledge goal
Knowledge goal	the next actionable step based on the given unknown
Biomedical concept classification/recognition	automatically identifying and mapping biomedical entities to ontologies
Ontologies	controlled vocabularies with specified relationships
Open biomedical ontologies (OBOs)	an effort to create standardized ontologies for use across biological and medical domains
Lexical cue	words or phrases that signify a statement of ignorance
Taxonomy of ignorance	a categorization of ignorance statements based on the entailed knowledge goal
Exploration by topic	automatically find statements of ignorance related to a topic from the ignorance-base
Exploration by experimental results	automatically contextualize experimental results in terms of statements of ignorance from the ignorance-base to understand what questions may bear on them
Ignorance enrichment	a method to identify biomedical concepts that are over-represented in a set of ignorance statements as compared to all sentences, and thus may be a new promising avenue to explore in relation to the input topic
Ignorance-category enrichment	a method to identify ignorance categories that are over-represented in a subset of ignorance statements as compared to all ignorance statements in order to illuminate the types of knowledge goals to pursue and to map out how they change over time

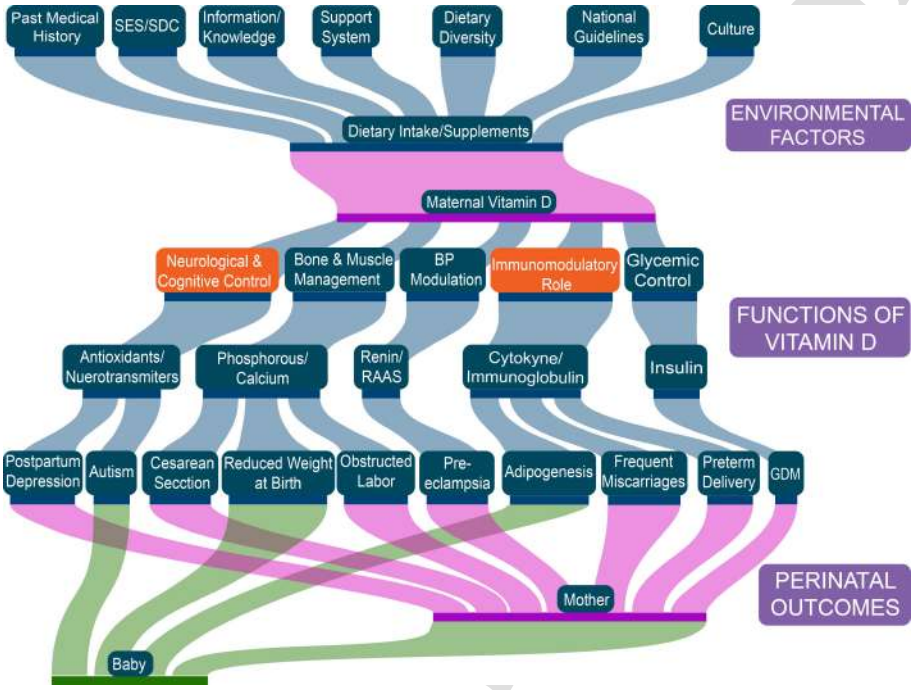


Figure 1: Relationship between society, maternal nutrition (vitamin D), and the effects on mother and offspring: a Sankey diagram created based on Figure 3 from [10]. The orange color represents the findings from the exploration methods that the concepts related to brain development and immune system are enriched in ignorance statements and possible novel avenues to explore. SES/SDC = socioeconomic status/sociodemographic characteristics; BP = blood pressure; GDM = gestational diabetes mellitus.

Statement of significance

Problem or Issue	What is Already Known	What this Paper Adds
No knowledge-base focused on scientific knowledge goals from the literature exists to provide insights for contextualizing topics and experimental results in our collective scientific ignorance.	Knowledge-bases exist to find information and contextualize experimental results in what is known. Known unknowns are important and studied under different focuses. One search engine exists to help scientists discover challenges and directions.	This study aims to create an ignorance-base focused on knowledge goals to provide insights to researchers interested in exploring a topic or contextualizing experimental results in the known unknowns. We found emerging and currently studied avenues ripe for future work.

2. Related Work

Many **knowledge-bases** (*e.g.*, the Reactome Pathway Knowledgebase [37]) exist to capture the known knowns from domain experts, the scientific literature, and other data sources such as experimental results [22]. These knowledge-bases have a variety of applications [22], including finding and interpreting information based on a single input topic, such as a concept, or a set of input topics that may be related, such as those from experimental results. In both cases the researchers want to find "relevant" information based on their query. For example, graduate students or researchers interested in learning about the field of prenatal nutrition might consult a database of dietary supplements [38]. Or researchers might perform a functional enrichment analysis to characterize a list of genes associated with vitamin D and preterm birth by finding relevant known biomedical concepts [36]. Many knowledge-base applications provide analyses to help researchers find and prioritize relevant information, which is our goal with the ignorance-base. Thus, we model the ignorance-base after knowledge-bases.

The aim of the ignorance-base is to help researchers find the most pertinent questions. Researchers gain these skills in graduate school, where

the goal is to identify and provide at least some solutions for a question that is unanswered. There are many books [39, 40, 41, 42, 43] and articles [15, 44, 45, 46, 47] discussing how to choose the most pertinent question or topic, and yet only one automated system has been developed, the COVID-19 challenges and directions search engine [20]. As explained above, our goal is to extend their work to provide the users with knowledge goals and insights based on their input query just as in the knowledge-bases. One of the main differences between our work and the COVID-19 search engine [20] is in the categorization of known unknowns. Lahav *et al.*, [20] classified known unknowns into two categories, namely challenge and research direction; most of the similar prior work, including ours [4], introduced more fine-grained categories.

The original linguistics phenomenon that sparked all these areas of research was hedging. Hedged statements in linguistics can be true or false to some extent [48]. Recognizing that scientific research articles included hedges, hedging was then defined more specifically within these articles as "any linguistic means used to indicate either a) a lack of complete commitment to the truth value of an accompanying proposition, or b) a desire not to express that commitment categorically" [49]. Hedging highlighted a focus on truth and facts. To help specify the levels of truth, research turned to uncertainty, and the ways that a writer can communicate what they do not know to the readers. One of the first attempts to understand uncertainty theoretically was for decisionmakers, especially for law [50]. Scientific uncertainty was defined as the "different kinds of potential error associated with descriptive scientific information" [50]. This resulted in a taxonomy of six categories of descriptive uncertainty: conceptual, measurement, sampling, modeling, causal, and epistemic, each characterized by its own kinds of errors. In the bioscience field specifically, prior work sought to explore speculative language by presenting many examples of the phenomenon and determining that it was feasible for humans to annotate [51]. They focused

on expressions of levels of belief including hypotheses, tentative conclusions, hedges, and speculations. Others have recast this phenomenon as factuality, alluding to a continuum that ranges from factual to counter-factual with degrees of uncertainty in between [52]. Still others [53] coined the term meta-knowledge to encompass different types of interpretive information including confidence levels, hypotheses, negation, and speculation. They [53] determined five categories of meta-knowledge including manner, source, polarity, certainty level, and knowledge-type. All of these works focused on these phenomena in relation to the current known knowledge (*i.e.*, how certain, speculative, hedged, factual, or meta the knowledge is). More recently research has refocused these categories on goals for future knowledge, anticipating the next actionable step research should take in future work [4]. For example, the statement "there can be a relationship between smoking and lung cancer" is uncertain [54], and also an ignorance statement. The knowledge goal is to gather more evidence to support the claim (*indication of proposed or incompletely understood research topic or assertion*). Boguslav *et al.*, [4] identified 13 categories of ignorance and showed preliminary evidence for this categorization. One aim of our work is to build on this foundation to show the value of categorizing the knowledge goals of known unknowns for the ignorance-base.

Other prior work related to known unknowns includes efforts to capture them through understanding the phenomenon [48, 49, 55, 56, 57, 58, 59, 52, 51, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87], creating taxonomies where a hierarchy of terms is linked by specified relationships [88, 62, 89, 52, 90, 91, 92, 50] and ontologies specifying relationships among controlled vocabularies [93, 94, 95, 96], annotating literature to create corpora [97, 98, 58, 56, 99, 100, 66, 101, 72, 82, 102, 103, 104], and automating identification of unknowns through classification tasks [105, 106, 52, 60, 61, 63, 64, 65, 67, 68, 69, 70, 71, 73, 76, 107, 77, 78, 79, 108, 81, 109, 83, 84, 110, 111, 85, 87, 59, 112, 113].

Some efforts have also sought to capture unknowns completely by creating theoretical frameworks, determining if the task is feasible for humans to perform, and automating it [4, 114, 62, 115, 116, 71, 75, 117, 80, 57, 86, 118, 51, 4, 20]. Only one work has created a formal search engine [20], and we create the first knowledge-base (ignorance-base) with added analyses, summaries, and visualizations that rely on a more fine-grained categorization of known unknowns.

Grounding our ignorance-base in the open biomedical ontologies (OBOs) [23, 24] also made our added analyses, summaries, and visualizations possible. Ontologies are vital to knowledge-based biomedical data science because they describe a knowledge representation in a way that preserves the definitions of biomedical entities and the relations between them [22]. Additionally, ontologies license "the ontological commitments a knowledge representation makes (*i.e.*, what it can or cannot describe), which inferences are possible within it, and, sometimes, which of those inferences can be made efficiently." [22] Within the biomedical domain, ontologies are "community consensus views of the entities involved in biology, medicine, and biomedical research, analogous to how nomenclature committees systematize naming conventions" [22]. Knowledge-bases grounded in and created from community-curated ontologies provide significant advantages for reproducibility in scientific research, for interoperability, and for avoiding pitfalls in the modeling of knowledge [22]. Knowledge-bases grounded in terminological resources, including UMLS (Unified Medical Languages System which includes MeSH - Medical Subject Headings), SNOMED CT (Systematized Nomenclature of Medicine, Clinical Terms), and the National Cancer Institute Thesaurus, lack some aspects of a computational ontology [22].

The COVID-19 search engine [20] used MeSH terms from UMLS, which lacks a common architecture and thus produces mappings that do not meld their terms together consistently into a single system [23]. UMLS [21] combines many vocabularies based only on the identification of synonymy rela-

tions between terms, resulting in potential loss of the intended meaning of concepts and distortion of the relationships between them during ontology mapping [119]. However, it can be easily applied to most currently existing databases [119]. Another effort to support biomedical data integration was the OBO foundry [23, 24], which sought to establish a set of principles for ontology development. These principles maintain the intended meaning of concepts, reduce the number and redundancy of ontologies, and require the cooperation and coordinated work of ontology developers [119]. Many OBOs, especially the Gene Ontology [120], are "specifically devoted to representing the biological knowledge underlying the reuse of data within new research contexts: in other words, it defines the ontology that researchers need to share to successfully draw new inferences from existing data sets" [121]. The goal of our work, using the OBOs, and that of Lahav *et al.*, [20], using UMLS, is to find new insights from the literature on existing biomedical concepts. The OBOs contain many more terms/classes and asserted (nontaxonomic) relationships than MeSH (*e.g.*, the Gene Ontology [122, 120, 23]). The OBOs are generally semantically richer and allow for more semantic/logical entailments [122, 120, 23, 123]. Further, systems were created to help integrate the ontologies (*e.g.*, BioPortal [124]). The downside of the OBOs is that they only integrate well with other databases derived from OBO Foundry ontologies [119]. We chose to use the OBOs because their richness and interoperability provide assurance that future work based on the OBOs can continue to build on our work. In time, the OBO model is also likely to remedy some of the flaws in UMLS [119], allowing future work to combine these efforts at standardization.

Our main novel contribution is providing analyses, summaries, and visualizations to help researchers find the next areas to study (biomedical concepts) and pertinent questions to ask (ignorance statements). Note that our analyses are made possible by and rely on both the knowledge goal categorization of known unknowns [4] and the OBOs [23]. Lahav *et al.*, [20] posed

future work to "build more tools to explore and visualize challenges and directions across science" [20]. This work is the first step towards those goals. Further, many knowledge-base applications for experimental results are used to provide the researchers with a "list of 'interesting' biomolecules" [125]. Functional enrichment analysis is the standard method for obtaining such lists and has become one of the most frequently used tools in computational biology [125, 126, 127, 128, 129]. For example, functional enrichment analysis provides valuable insight into a collective biological function underlying a list of genes "by systematically mapping genes and proteins to their associated biological annotations ... and then comparing the distribution of the terms within a gene set of interest with the background distribution of these terms" to identify statistically over- or under-represented terms within the list of interest [125]. The set of enriched terms then describes some important biological process or behavior [125]. Our work aims to provide a similar list of enriched terms with regards to ignorance (ignorance enrichment and ignorance-category enrichment) based on an input topic or set of experimental results, and use it to create summaries and visualizations to help researchers narrow in on the next areas to study and the pertinent questions to ask.

The goal of our system is to use the ignorance-base and exploration methods to go beyond the usual reach of a search engine, namely to provide summaries and visualizations for the numerous sentences and articles returned from an input topic or set of experimental results. We used ideas and techniques from the field of document visualization, which "transforms textual information such as words, sentences, documents, and their relationships into a visual form, enabling users ... to lessen their mental workload when faced with a substantial quantity of available textual documents" [130]. Gan *et al.*, [130] provided an overview of the field with design principles and examples. They discussed visualization techniques for both single document and document collection visualizations as well as vocabulary-based visualizations to

visualizations of document similarity [130]. We utilize Tag Clouds [130] to visualize the frequency of both words and biomedical concepts to compare them for ignorance statements versus all sentences. Other vocabulary-based visualizations include Wordle, TextArc, and DocuBurst [130]. Visualizations based on semantic structure include Semantic Graphs and visualizations based on document content include WordTree and Arc Diagram [130]. Visualizations for collections of documents can illustrate document themes, document core content, changes over different versions, document relationships, and document similarity [130]. All of these can help researchers gain an overview of the entire collection [130]. This work provides preliminary visualizations to help researchers digest the output of the ignorance-base. Future work can add and evaluate more visualizations.

The ultimate goal of this work is to provide analyses, summaries, and visualizations of ignorance statements resulting from an input topic or set of experimental results. For an input topic, similar works are search engines including PubMed and the COVID-19 search engine [20]. We compare our results to them. For experimental results, methods for standard functional enrichment analyses (contextualizing experimental results) use knowledgebases and ontologies [125, 126, 127, 128, 129], and natural language processing (NLP) tools over the biomedical literature [17, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140]. Some of this prior work not only aimed to characterize genes but also to help define new research areas (*e.g.*, [17] as one of a few goals), generate new hypotheses (*e.g.*, [141]), find information about genes of unknown function and fill gaps in knowledge (*e.g.*, a preprint [139] using manual curation). Thinking beyond a gene list, if we consider pathway models as experimental results, tools exist to associate pathway models to the literature (*e.g.*, [142]) and some of these take uncertainty into account (*e.g.*, [60, 143]). These works however focus on confidence and relevance to current knowledge, respectively, rather than focusing on the role they play in future knowledge and explicitly representing statements of known unknowns.

Thus, instead we compare our results to the standard functional enrichment analysis [144, 145] to highlight the difference and power of ignorance enrichment. We build upon all of this previous work to create an ignorance-base grounded in knowledge goals and OBOs to explore by topic and experimental results, providing researchers with tools and visualizations to explore the landscape of our collective scientific ignorance.

3. Methods

We created an ignorance-base grounded in knowledge-goals (ignorance) and OBOs to provide analyses, visualizations, and summarizations to researchers to help them find pertinent questions to explore in future work. We combined the best-performing ignorance classifiers (extending the work of [4] to create a corpus of 91 articles) with state-of-the-art biomedical concept classifiers [25] to create the ignorance-base and explore it by a topic and by experimental results. The ignorance-base can be queried by ontology concepts, ignorance categories, specific lexical cues, or any combination of the three. We compared our results to standard methods and to the COVID-19 search engine [20].

The rest of this section is organized into the following subsections:

1. Creating the ignorance-base
2. Exploration by topic
3. Exploration by experimental results

3.1. Materials

The inputs for all systems were scientific prenatal nutrition articles. We used full-text articles from the PubMed Central Open Access (PMCOA) subset of PubMed [146], allowing us more data beyond the abstract and the ability to share it publicly. 1,643 prenatal nutrition articles (1939-2018) were gathered from querying PMCOA for 54 regular expressions (keywords such as {prenatal, perinatal and antenatal} paired with keywords like {nutrition,

vitamin and supplement} determined in consultation with a prenatal nutrition expert, Teri L. Hernandez. All articles were provided in XML format, which was parsed and converted to text format using a script in Java. All subsequent computation was implemented in Python 3, with its associated packages. The continued annotation effort used Knowtator [147] and Protege [148] as in previous work [4], allowing the ignorance taxonomy (see Table 2) to be easily browsable like an ontology, and helping the annotators select the correct level of specificity for each lexical cue. The classification frameworks and models were also from our previous work [25, 4].

To connect the ignorance statements to the biomedical concepts, the ignorance-base was built upon the PheKnowLator knowledge graph (PheKnowLator_v3.0.2_full _subclass_relationsOnly_OWLNETS_SUBCLASS_purified _NetworkxMultic DiGraph.gpickle), which semantically integrates eleven OBOs [26, 27]. For exploration by topic, we compared our results to a PubMed literature search and the COVID-19 search engine [20]. For exploration by experimental results, the gene list (our motivating example of experimental results) was gathered from a PMCOA article (PMC6988958) [36]. We also used DAVID (a tool for functional annotation and enrichment analyses of gene lists) [145] as a standard approach for functional enrichment analysis to compare to ignorance enrichment.

Computation used a contemporary laptop (MacBook Pro) and an NIH-funded shared supercomputing resource [149] that included:

- 55 standard compute nodes with 64 hyperthreaded cores and 512GB of RAM
- 3 high-memory compute nodes with 48 cores and 1TB of RAM
- GPU nodes with Nvidia Tesla k40, Tesla k20, and Titan GPUs
- A high-speed Ethernet interconnect between 10 and 40 Gb/s

We used both CPUs and GPUs to train, evaluate, and predict statements

of ignorance. We also used CPUs and GPUs for predicting annotations of textual mentions of OBO concepts.

Code for the ignorance-base and exploration methods can be found at: https://github.com/UCDenver-ccp/Ignorance-Base. The expanded ignorance corpus can be found at: https://github.com/UCDenver-ccp/Ignorance-Question-Corpus, with all associated code and models at: https://github.com/UCDenver-ccp/Ignorance-Question-Work-Full-Corpus. Code for concept recognition of the OBOs can be found at: https://github.com/UCDenver-ccp/Concept-Recognition-as-Translation.

3.2. Creating the ignorance-base

We created the ignorance-base grounded in knowledge goals and OBOs. We expanded our corpus of ignorance statements based on knowledge goals to train and evaluate high-quality ignorance classifiers [4] and combined them with biomedical concept classifiers [25].

3.2.1. Expanding the ignorance corpus

The goal was to create an ignorance corpus to show that ignorance statements can be reliably identified and automatically classified. We produced a gold-standard corpus consisting of articles with labeled sentences as **statements of ignorance** along with the **lexical cue(s)** (words or short phrases) that distinctly signify it as such mapped to a categorization of **knowledge goals** (**ignorance taxonomy**). This was done by examining spans of text each in the form of a word, short phrase, or whole sentence. Following the example above, "<these inconsistent observations point to the complicated role of VITAMIN D in the IMMUNE modulation and disease process>" (PMC4889866), the ignorance statement and entailed knowledge goal were identified based on the underlined words that communicate knowledge is missing, **lexical cues**, which map to an **ignorance taxonomy**, a formal categorization of knowledge goals. The cue inconsistent is an *indication of*

alternative research options or controversy of research (abbreviated as *alternative/controversy*), and complicated is an *indication of difficult research task* (abbreviated as *difficult task*) (see Table 2). Our preliminary previous work [4] created a corpus of 60 articles annotated with lexical cues and ignorance categories. The goal was for an annotator to identify or an algorithm to classify that our example sentence was a statement of ignorance as shown by the brackets around the sentence. From there, once the sentence was deemed a statement of ignorance, the goal was to identify or classify that all underlined words including inconsistent, observations, etc. were the lexical cues that signified it as such. Note that we conducted two different classification task on the sentence and word levels respectively using the same data. Also note that one sentence can have multiple lexical cues that signify ignorance. The ignorance taxonomy helped to distinguish between different lexical cues: the annotator and classifier also needed to map the underlined cues to the specific ignorance category they deemed to capture the knowledge goal of the sentence. Here we expanded that corpus to 91 articles to provide enough data to evaluate the classifiers on a held-out set of gold-standard data. We used the same methodologies as in our previous work [4], aside from a few minor changes.

Two new independent annotators, Katherine J Sullivan (K.J.S.) and Stephanie Araki (S.A.), both computational biology researchers, were provided with one to four articles, chosen randomly, in the Knowtator platform [147]. Each article was preprocessed such that lexical cues were automatically highlighted and linked to their corresponding classes of the ignorance taxonomy (Table 2), since prior efforts [4] showed that the annotation task was prohibitively difficult in unmarked documents. The annotators read through each article independently, deciding for each cue highlighted whether it signified an ignorance statement or not, and then either confirmed the ignorance taxonomy category or deleted the cue. Note that a lexical cue can map to multiple categories depending on the context (*e.g.*, the cue however can map

to *anomalous/curious* or *alternative/controversy*). The annotators were also asked to add any new cues that signified ignorance and were not already highlighted by mapping them to the correct ignorance category. In the next annotation round, these new cues were added to the ignorance taxonomy. To capture the scope of each ignorance cue, we adapted the guidelines from BioScope [58], highlighting the whole sentence as the scope capturing all encompassed lexical cues due to difficulties with capturing only parts of the sentences. The annotators reviewed all annotations together, and Mayla R. Boguslav (M.R.B.) adjudicated any disagreements as they arose to create gold-standard articles that achieved an inter-annotator agreement (IAA) of at least 70-80%. The IAA is a measure of how well the annotations agree, and here we calculated the F1 score between the two annotations [150, 151]. An exact IAA was calculated on the exact text span of lexical cues or scopes as well as the ignorance category assignments. We also calculated a fuzzy IAA when the category assignments matched but not the text span of the cue or scope, or vice versa. For example, one annotator may highlight only need in a sentence containing the phrase need to be. In this case, we adopt the larger text span. (See [4] for more details).

K.J.S. and S.A. were trained first on eight random articles chosen from the 60 previous gold-standard articles. Any changes made to these articles (due to more experience with the task) were marked accordingly. After reaching the required IAA, new articles were chosen randomly using seeded randomness. For the first eight new articles (two batches of four), both annotators annotated the same articles as usual. After reaching IAAs of 80% or higher, we decided to divide the work: each annotator separately annotated one or two different articles and then adjudicated all annotations with M.R.B. Since the classic IAA could not be calculated because there was only one annotator, we calculated an "F1 score" between the original set of annotations and the adjudicated version to see how reliable the single annotation was compared to the final adjudicated version. We continued annotation when this score

stayed above 80%, indicating a sufficient level of accuracy.

Table 2: Ignorance Taxonomy: definitions, knowledge goals, example cues, and total cue count. The categories in bold are only narrow categories. Abbreviations are in italics.

Ignorance Category	**Definition**	**Knowledge Goal**	**Example Cues**	**Total Cues**
indication of *answered* research *question*	A statement of a goal or objective of a study that is attempted or completed during the study.	to find the answer(s) in the article; determine if the question(s) is (are) fully answered in the article	aim, goal, objective, our study, sought, to determine	64
indication of *unknown* or *novel* research topic or assertion	A statement that indicates something is not known (a lack of information), or information is presented for the first time (new or novel) and a significant amount of research is needed; not a statement about the absence of something.	to explore the unknown further to gain any insights	could not find, don't know, elusive, not...established, uncertain, still unclear	155
indication of *explicit* research *inquiry*	An explicit statement of inquiry (with a question mark or question word such as how, where, what, why).	to find answers to the question and/or discover methodologies that will help answer the question	?, what, where, wondered, why	19

indication of proposed or *incompletely understood* research topic or assertion	A positive or negative statement proposing a possible/feasible explanation for a phenomenon on the basis of limited evidence as a starting point for further investigation OR a statement that information is needed to support an assertion or claim, including both positive and negative statements. Either a statement that some evidence already exists, explaining how current findings support previous work, adding confidence to a claim OR a statement that information is limited, more research is needed or is ongoing including limitations – biases or shortcomings related to the study design and execution.	to gather more evidence to support the claim OR conduct more research to determine the validity of the claim; complete the partial picture; consider the shortcomings and biases for the next experiment and how it can be addressed.	a good understanding, believe, evidence...limited, has been suggested, hypothesis, no studies, possibly, preliminary stage, remains under investigation, still being discovered, support, trend	797
indication of *indefinite relationship* among research variables	A statement about a connection, link, or association between at least two variables; connectedness between entities and/or interactions representing their relatedness or influence.	to confirm the connection, link, or association between variables; determine the full underlying relationship between variables	affect, associated, correlate, factor, influence, interact, link, pattern, tend	198

indication of *largely understood* research topic or assertion	A statement staking a claim to the most likely explanation, relationship, or phenomenon; assumes that there is a good chance this understanding is correct.	to determine if the most likely option is correct or if another option is more feasible	almost all, assumed, concluding, evident, it is clear, most likely, thus	202
indication of *anomalous* or *curious* research finding	A statement of a surprising result, conclusion, observation or situation; the researchers were not expecting the result, conclusion, observation or situation but are intrigued by it.	to explore the surprising result, conclusion, or situation more and determine if the result, conclusion, observation, or situation is repeatable	appeared to be, interestingly, noteworthy, surprisingly	113

indication of *alternative* research options or *controversy* of research	Either an explicit statement of multiple (at least two) choices, actions, approaches, or methods that need to be experimentally determined, including statements with an implied second option, such as "whether". This includes a statement of disagreement amongst researchers OR a lack of consensus OR at least two possible answers presented as results from different researchers, usually in reference to previous results and stated when results contradict or disagree with each other.	to determine the correct option or a better option and if there are disagreements, to determine the truth to break any disagreements	cannot rule out, claims, has been challenged, whether, whilst	221
indication of *difficult* research *task*	A statement of something not easily done, accomplished, comprehended, or solved; or a complicated thing with a multitude of underlying pieces or parts; heterogeneity; excludes medical complications.	to create methods to study the complicated system and to better understand any piece of the complicated system; potentially requires new experiments or better techniques	not feasible, remains...challenge, variability, rarely able to	98

indication of research *problem* or *complication*	A statement of issues, problems, mistakes, or medical complications that are cause for anxiety and/or worry.	to determine the gravity of the concern and determine if it needs to be dealt with before the next experiment or study	issue, error, insufficient, lack of reproducibility, publication bias, underestimated	98
indication of *future* research *work*	A statement of extensions, including next steps, directions, opportunities, approaches, or considerations of the described work that may be implemented at some future time point. This also includes a statement of suggestion or a proposal as to the next best course of action, especially one put forward by an authoritative body; advice telling someone the best action to take.	to determine the next course of action based on this future work proposal	additional research, are needed, continue to explore, further study, more...studies, recommend, warrants, worthy of closer attention	258
indication of *future* research *prediction*	A statement of extrapolation of given data into the future and/or from past observations, without reference to next steps.	to run the simulation or experiment to determine if the prediction is correct; publicize the outcomes of the study to the correct people	allow, expect, if so, serve as a basis, will	27

indication of *important consideration* for future research work	A statement calling for attention including an action needed to be taken immediately or information that needs to be disseminated immediately OR is critical: being in or verging on a state of crisis or emergency OR urgently needed OR absolutely necessary.	to take the urgent action ASAP or distribute the knowledge ASAP	call for action, cautious, crucial, emphasis, global problem, high on the agenda, necessary, relevant to note, vital	263

3.2.2. Training and evaluating high-quality ignorance classifiers

With our full corpus of 91 articles, new classifiers were trained and optimized using a training set of 65 articles (approximately 2/3) and ultimately evaluated against a held-out test set of 26 articles (approximately 1/3). Ignorance classification can be made at either the sentence or the word level and as either a binary or a multi-classification problem. (Note that the COVID-19 search engine [20] only classified at the sentence level as a multi-label classification problem recognizing that one sentence can be both a challenge and a direction.) At the sentence level, the binary task determines whether or not a sentence's scope contains a statement of ignorance. Since each statement of ignorance has at least one lexical cue labeled, the sentence can also be labeled using the ignorance categories implied by its lexical cues. Following the example above this would include the categories *alternative/controversy*, *difficult task*, etc. This now created a multi-classification problem of mapping sentences to the specific ignorance categories of their lexical cues. Conversely, we can also focus the binary task on the lexical cues to classify whether a word in an article was part of a lexical cue or not as labeled in the corpus. For the multi-classification task, the words would be mapped to their specific ignorance categories. Note that the test data included 501 unique lexical cues with no sentence examples in the training data. To avoid batch effects based

on the different annotators, we split each batch separately (see Table 3).

For both the sentence- and word-level multi-classification tasks, we tested both one true multi-classifier and an ensemble that split each task up into 13 smaller binary tasks in which the ignorance category of interest was the positive case and all other sentences/words belonging to a different category were negative cases. Combining all 13 binary classifiers into an ensemble gave the full categorization for each sentence and avoided the problem of overlaps between categories. (Note that one sentence can map to multiple categories as the example above and as in [20].) In all cases, we split the training data 90:10 for training and validation, and then evaluated separately on the held-out test set. We report the F1 scores for both sentence and word classification tasks on the held-out test set of 26 articles for (1) the one binary task (ALL CATEGORIES BINARY), (2) an ensemble multi-classifier composed of the 13 separate binary tasks (each category reported individually), and (3) the one multi-classifier (ALL CATEGORIES COMBINED - the macro-average). The best performing models were used for the ignorance-base. We further compared the ignorance binary sentence classifier to the binary COVID-19 search engine PubMedBERT model [20] (a challenge or direction) on our held-out test set to understand the relationship between the two systems. Lahav *et al.,* [20] calculated a probability for each class label and used a threshold of 0.99 to ensure high-confidence sentences. We thus used their model in the same way. We calculated an F1 score between their predictions and ours (the F1 score is the same no matter the reference - it only switches precision and recall). A low F1 score means the classifiers identified different sentences and a high score means the classifiers captured a similar phenomenon.

For all classification tasks, each article was segmented into sentences and then words used in the respective tasks. All models were chosen based on our prior work [4] and on an evaluation of several canonical models for concept recognition [25]. As our taxonomy was very similar to an ontology, we

Table 3: Data split for automatic classification: The table is in order of completion of annotation batches. Note that E.K.W. is Elizabeth K. White, M.R.B. is Mayla R. Boguslav, E.D. is Emily Dunn, Gold-standard is the previous gold-standard up to that point (the first row), K.J.S. is Katherine J. Sullivan, and S.A. is Stephanie Araki. *M.R.B. is an annotator along with the others. **E.D. only annotated one article along with the other annotators and then stopped. ***M.R.B. was the adjudicator in these batches.

Annotation batch	Total Articles	Training Articles	Testing Articles
Prior work: E.K.W., M.R.B.*, (E.D.**)	52	37	15
Training: Gold-standard, K.J.S., S.A.	8	6	2
K.J.S., S.A., (M.R.B***)	8	6	2
K.J.S., M.R.B.***	11	8	3
S.A., M.R.B.***	12	8	4
Total Articles	91	65	26
Total Sentences	12,055	8,281	3,774
Total Words	416,866	285,439	131,427

used our prior work in concept recognition [25] applied to a different type of linguistic phenomenon. In this work [25] we explored and evaluated some of the canonical algorithms for concept recognition over many different ontologies and found that the CRF [152] and BioBERT [153] achieved the best performance for the task of span detection.

For sentence classification, we first built a simple Feed-Forward Artificial Neural Network (ANN), consisting only of three layers as our baseline model. We then fine-tuned both BERT [154] and BioBERT [153], so that we could compare a basic deep learning model to state-of-the-art language models. Our ANN consisted of a flattened layer followed by several dense layers to allow for arbitrary non-linear transformations of the input, with early stopping callbacks to avoid over-fitting (additional details have been previously published [4]). For BioBERT, we used its domain-specific vocabulary to train the base BERT model. The same hyper-parameters were used for both BioBERT and BERT: batch size of 16, patience of 5, and a learning rate of 1×10^{-5}. The number of epochs was tuned using truncating functions to avoid overfitting. We did not freeze the layers of the pre-trained BERT

model and allowed the weights to keep updating during training for better performance.

For word classification, the goal was to automatically identify all lexical cues per our annotation task setup. Note that our input was all sentences not only ignorance statements so that we could predict on any new sentences for the ignorance-base. We represented the underlying data using BIO- tagging: a word at the beginning of a lexical cue was marked 'B'; a word inside a multi-word cue was marked 'I'; a word outside of a cue (*i.e.*, not a word in the lexical cue) was marked 'O'. (All words in non-ignorance statements, no lexical cues, were marked as 'O'.) If the lexical cue contained a discontinuity (*e.g.*, no...exist where the "..." signifies a discontinuity), we labeled the words that exist between the lexical cues as 'O-' (BIO-tagging scheme from [25]). CRF models were tuned with L1 and L2 regularization to avoid overfitting using the sklearn-crfsuite Python package [155]. For BioBERT, the named entity recognition baseline parameters performed quite well, most likely because it is a similar task, and so we did not tune any other parameters [153]. For a more thorough discussion of all data preparation and representation, the performance metrics, and the classification algorithms please refer to our prior works [4, 25].

Our ignorance statement identification task relied on the identification of lexical cues. To determine the role they played in classifying sentences, we conducted an ablation study. We deleted all ignorance-category annotations of the sentences and then re-trained the sentence classifiers using the best performing model for each category. We tested the models' performance on our held-out test data set. Poor performance would indicate that the sentence classifiers relied heavily on lexical cues, while good performance would point to the existence of other features beyond lexical cues that could identify ignorance statements. We report the results of these classifiers and discuss the use of lexical cues in related work to understand how well they apply beyond our work here. Additionally, we compared our cue list to

some canonical work in the field to show generalizability beyond our prenatal nutrition corpus. We compared our list to the Bioscope [58] lexical cue list for clinical abstracts and articles, the lexical cue set for the meta-knowledge annotations of the GENIA project [156, 98], and the COVID-19 search engine keyword list [20]. We allowed for partial matching between the cues from each list to consider all forms that the cue can take within a phrase. Note that these works are similar to ours, but not the exact same task. Since each of these works focused on different domains than ours, the number of overlapping cues between our work and theirs may indicate generalizability beyond our work.

3.2.3. Combining ignorance and biomedical concept classifiers

Creating an ignorance-base grounded in knowledge goals and OBOs allowed us to explore it and provide analyses, summaries, and visualizations of the outputs. Further, grounding the ignorance-base in the OBOs allowed us to connect our work to many other knowledge-bases [22]. To create the ignorance-base we combined the ignorance and biomedical concept classifiers over all 1,643 prenatal nutrition articles. The ignorance-base included all sentences from all articles to capture all biomedical concepts for comparison of our ignorance approach (only ignorance statements) to the standard literature search approach (all articles).

For ignorance classification, we used the 91 gold-standard corpus articles, and ran the best ignorance classifiers over the other 1,552 articles. Similarly, we ran our state-of-the-art biomedical concept classifiers [25] over all 1,643 articles to automatically identify biomedical concepts represented in ten OBOs, taking the best-performing models in terms of F1 scores, (CRFs [152] and BioBERT [153]). Most F1 scores ranged from 0.7-0.98 with the exception of PR at 0.53 (see Table 5 in [25]). Note that even though all of the classifiers performed close to the state of the art for the task at hand, they are still automated, so we draw conclusions cautiously. Because of this, we manually reviewed a random sampling of the identified biomedical con-

cepts (a few hundred of each). The ten OBOs used for our work were the same used to manually annotate the CRAFT Corpus [157, 158], (a corpus of full-text articles annotated along multiple syntactic and semantic axes, including extensive concept annotations):

1. Chemical Entities of Biological Interest (ChEBI)
2. Cell Ontology (CL)
3. Gene Ontology (GO):
   (a) Gene Ontology Biological Process (GO_BP)
   (b) Gene Ontology Cellular Component (GO_CC)
   (c) Gene Ontology Molecular Function (GO_MF)
4. Molecular Process Ontology (MOP)
5. NCBI Taxonomy (NCBITaxon)
6. Protein Ontology (PR)
7. Sequence Ontology (SO)
8. Uber-anatomy Ontology (UBERON)

For each of these ontologies, two sets of concept annotations were created for CRAFT (and appear in the public distribution): only proper classes of these OBOs and another adding in extension classes to better integrate the OBOs (created by the semantic annotation lead but defined in terms of proper OBO classes). We employed automatic concept recognition of our prenatal nutrition corpus with both the core OBOs and with the corresponding extended OBOs (suffixed with "_EXT"). Note that classification performance on the OBOs_EXT was lower in general compared to the OBOs, especially for PR and PR_EXT, so caution should be taken in interpreting those results. We focused on the proper OBO classes going forward, but have the data and results for both. (PheKnowLator does not currently have the OBOs_EXT, but it is easily extendable.) Combining ignorance and biomedical concept classifiers automatically captured all ignorance statements and biomedical concepts for the 1,643 prenatal nutrition articles.

For clarification, the underlying data for the ignorance-base included sentences like the example above and another example: "it has an important role in BONE HOMEOSTASIS, BRAIN DEVELOPMENT and MODULATION OF the IMMUNE SYSTEM and yet the impact of ANTENATAL VITAMIN D deficiency on infant outcomes is poorly understood" (PMC4072587). The lexical cues important mapped to *important consideration*, role and impact mapped to *indefinite relationship*, yet mapped to *anomalous/curious*, and poorly understood to *unknown/novel*. BONE HOMEOSTASIS mapped to GO:0060348 (bone development), BRAIN DEVELOPMENT mapped to GO:0007420 (brain development), BRAIN also mapped to UBERON:0000955 (brain), MODULATION OF...IMMUNE SYSTEM mapped to GO:0002682 (regulation of immune system process), IMMUNE SYSTEM also mapped to UBERON:0002405 (immune system), ANTENATAL mapped to GO:0007567 (parturition), and VITAMIN D mapped to ChEBI:27300 (vitamin D). All of these mappings were identified by the classifiers. Note that we also identified biomedical concepts in non-ignorance statements. The entailed knowledge goal was to explore the relationship between prenatal vitamin D deficiency and infant outcomes through the important role of vitamin D. Note that all the ignorance categories interact to form the final knowledge goal, making this task quite difficult.

The power of the ignorance-base is in its potential to be used for exploratory analyses, summaries, and visualizations to help researchers choose a topic to study or contextualize their experimental results in the known unknowns. Thus, in order to explore these data, we created a network representation of the ignorance-base to connect all sentences from these articles using both the ignorance lexical cues and biomedical concepts (see Figure 2). We combined all the literature data to connect sentences that have the same ignorance lexical cues, such as poorly understood, and then used PheKnowLator to compile all assertions mentioning the same given set of biomedical concepts, such as VITAMIN D. The semantic integration of PheKnowLator

allowed us to not only connect our sentences to the biomedical concepts, but also to related ones; these connections were used in exploration by experimental results. This network can be used to search for all sentences that include the biomedical concept VITAMIN D, the lexical cue poorly understood, sentences with the ignorance category *unknown/novel*, or any combination of these features. Each sentence also related back to an article with its own metadata to be used for summaries and visualizations. For instance, the publication date was used to map how ignorance categories changed over time for a topic. Note that all sentences in all articles were included whether or not they contained ignorance statements, allowing for the ignorance enrichment comparison to the background information.

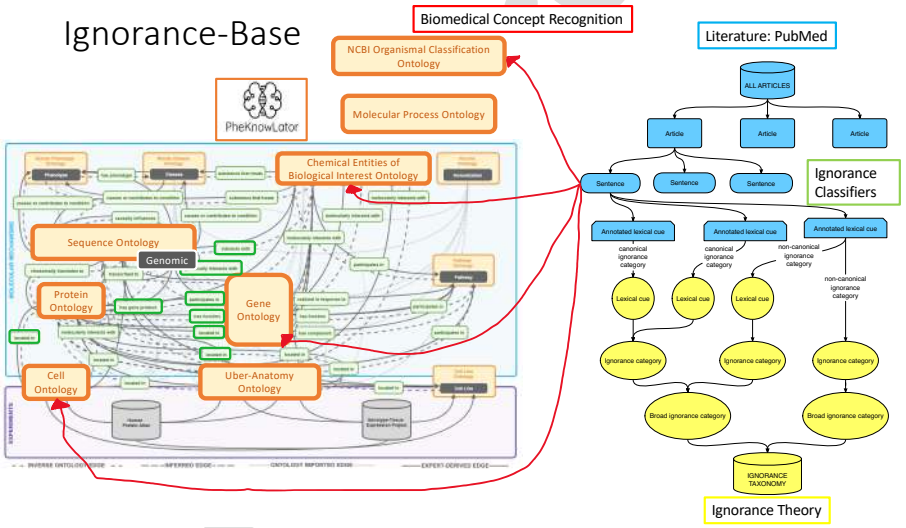


Figure 2: Network representation of the ignorance-base: The top right corner is the literature connecting the articles via segmented sentences (in blue) to the ignorance taxonomy (in yellow) through the ignorance classifiers (the annotated lexical cues). The sentences also connect to the biomedical concepts on the left with PheKnowLator [26, 27] using the biomedical concept classifiers with the ontologies of interest in bold and larger font.

3.3. Exploration by topic

The goal of exploration by topic was to explore the ignorance statements surrounding a topic to reveal novel insights. We compared our approach to a standard literature search (using biomedical concept expansion without ignorance expansion) and the COVID-19 search engine [20]. These comparisons provide a direct and informative comparison of results at the sentence level.

For our ignorance approach, an input topic consisted of a list of ontology concepts in PheKnowLator. To illustrate our approach, we explored the topic of vitamin D in consultation with a prenatal nutrition specialist (T.L.H.). We mapped the topic of vitamin D to four OBO concepts narrowed from 38 exact matches (280 partial matches): VITAMIN D (ChEBI:27300), D3 VITAMINS (ChEBI:73558), CALCIOL/VITAMIN D3 (ChEBI:28940), and VITAMIN D2 (ChEBI:28934). Note that going forward, when we refer to vitamin D, we mean the union of these four search terms. For the standard literature approach, we gathered all sentences from the ignorance-base that included terms from this vitamin D OBO concept list. For the ignorance approach, we only took the sentences that contained a vitamin D OBO list concept and had an ignorance lexical cue. For the COVID-19 search engine [20], we conducted two comparisons based on their findings that their models generalize beyond Covid-19. For a full comparison to our task, we ran their model [20] over our vitamin D sentences to calculate agreement between the two tasks and to determine the most frequent ontology concepts based on their method. Lahav *et al.*, [20] calculated a probability for each class label and used a threshold of 0.99 to ensure high-confidence sentences. (Note that they [20] used MeSH terms and not OBOs for concept recognition). Since our comparisons were not exactly the same between OBOs and MeSH, we also used their online search engine [20] to find relevant MeSH concepts and sentences to compare to our results. We searched for the concepts "vitamin D" and "pregnancy" (to simulate our corpus theme). It provided a list of sentences and a drop-down list of the most frequent MeSH concepts in

relation to the topic.

For our ignorance approach, we not only provided a list of sentences and the most frequent concepts, but also an analysis of the most ignorance-enriched concepts (ignorance enrichment). The goal was to help researchers narrow in on a specific topic to explore in future research. The ignorance statements surrounding the topic were available to explore. We also provided visualizations of the most frequent concepts: we present both biomedical concept clouds and word clouds along with frequency tables to explore them. For the COVID-19 search engine [20], we compared the frequency lists for both their model run on our data and their search engine results. For enrichment, we used the hypergeometric test for over-representation with both Bonferroni (family-wise error rate) and Benjamini-Hochberg (false discovery rate) multiple testing corrections [159, 125] to find concepts enriched in ignorance statements compared to all vitamin D sentences. We compared these results to the standard literature search approach using concept enrichment (concepts enriched in all vitamin D sentences compared to all sentences). In comparing our ignorance approach to the standard literature approach (see Figure 3), if a concept was more frequent or enriched in the standard approach but not in ignorance, then it could be established information. If a concept appeared in both approaches, then it might currently be under study (currently studied). If a concept only appeared in the ignorance approach then it could be an emerging topic. Note that concepts that were not frequent or enriched in either approach were not of interest.

Our ignorance approach can further help researchers narrow their research topic to one with the right scope of ignorance. We demonstrate this by visualizing how known unknowns are described (ignorance-category enrichment) and how they change over time. We bubble plotted the ignorance categories per article over time, with the bubble size representing the percentage of sentences in an article scaled by the total number of sentences of that category. Using these methods, the researchers continued to deep dive into

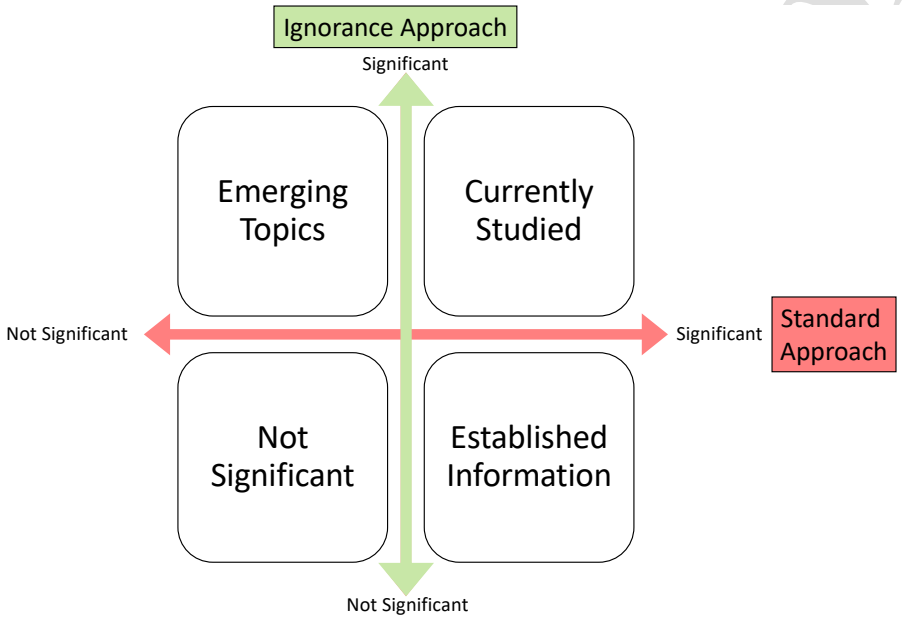


Figure 3: Ignorance vs. Standard Approach Results Chart: The interpretation of the results comparing the ignorance approach to the standard approach.

ignorance statements that included their topic, enriched concepts, and enriched ignorance categories to find knowledge goals to pursue for research. In order to determine if we found *novel* avenues to explore using our ignorance approach, we consulted both our prenatal nutrition specialist (T.L.H.) and PubMed (after our data collection time period) to corroborate our findings.

3.4. Exploration by experimental results

The goal of exploration by experimental results was to identify questions (ignorance statements) that may bear on the results, providing new avenues for exploration (biomedical concepts), potentially from other fields. Exploration by experimental results used the same methods as exploration by topic with some added pre-processing steps and analyses based on the relationship between the inputs. As before, the input topic was still an OBO concept list,

but an extra pre-processing step connected the experimental results to OBO concepts in PheKnowLator [26, 27]. In general, as long as the experimental results can be mapped to OBO concepts in and through PheKnowLator, we can connect them to the ignorance-base.

To illustrate our approach, we used our motivating example, the vitamin D and sPTB gene list (Entrez genes) [36], and mapped it to the genomic part of the sequence ontology (SO) and the corresponding proteins in the protein ontology (PR). This initialized the list of ontology terms to use for our search. To add more terms, we utilized the relations ontology (RO) which connects the different ontologies together in PheKnowLator. For instance, the relation "interacts with" (RO:0002434) connects proteins or genes to chemicals (ChEBI). This yielded a large list of ontology terms; we then found all the sentences that contained these terms (our sentences of interest). Note that not all OBO concepts connected to a sentence. From here, we performed all the same analyses as exploration by topic, including finding articles, sentences, ignorance categories, and concepts to investigate. Further, we added three more analyses: (1) gene list coverage (prioritizing the OBO concepts that connect to the most genes), (2) comparisons to other enrichment analyses such as DAVID, and (3) comparisons to any other findings about the gene list such as findings from a paper. (See figure 4 for the exploration by experimental results pipeline.) Note that neither a standard literature search nor the COVID-19 search engine can currently support queries by experimental results.

Gene list coverage can help prioritize which OBO concepts are most critical to examine. As we mapped the gene list to the OBO concepts, some OBO concepts had many genes map to them, implying that these OBO concepts were potentially more relevant to the gene list than concepts with fewer genes mapping to them. Thus, we sorted the OBO concept list by these high coverage ones and looked to see if those were enriched in all of our sentences of interest and/or in ignorance sentences. This provided a smaller and more

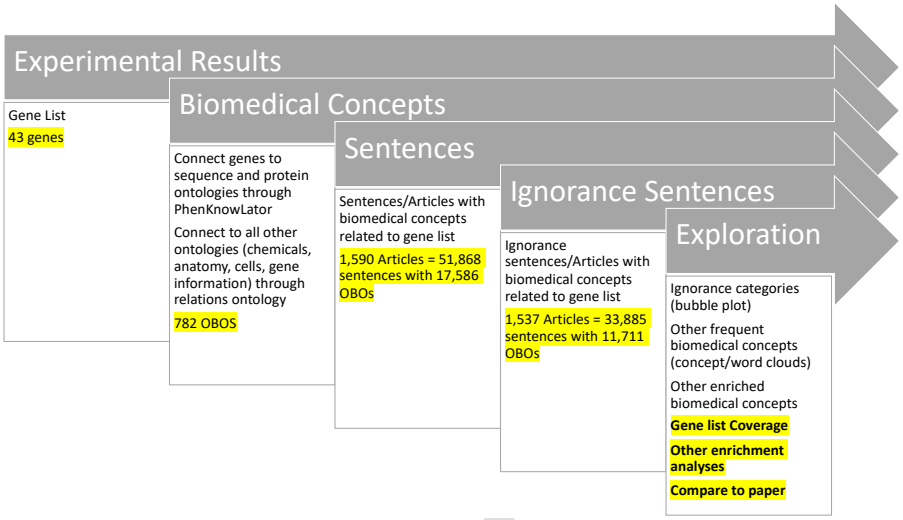


Figure 4: Exploration by Experimental Results (gene list) pipeline: The results are in yellow highlights for the example presented here. For exploration at the end of the pipeline, the three not highlighted are the same as exploration by topic and the three highlighted are the new additions based on a gene list.

refined list to start exploring.

Since canonical enrichment methods can also help prioritize OBO concepts, we compared our ignorance enrichment method to them, allowing us to both enhance the canonical methods and find new lines of investigation. From tools such as DAVID [145], we got a list of enriched OBOs (GO concepts) by using the gene list and functional annotations from their entailed knowledge-bases. We then found and examined any ignorance statements that contained the concepts linked by DAVID. Further, our method provided a list of OBO concepts enriched in ignorance statements. Thus, we compared and examined these two lists to add the ignorance layer to the classic enrichment analysis and to find new concepts that may be currently unknown to the knowledge-bases but potential emerging topics related to the gene list.

Given that our gene list came from a paper, we compared our ignorance-

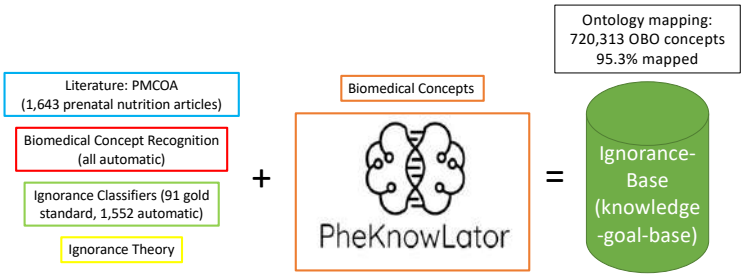
approach findings to the paper findings to identify questions that may bear on it, providing new avenues for exploration from other fields. Yadama *et al.*, [36] focused on the immune system in their main findings. For the resulting biomedical concepts of interest and their corresponding sentences and articles from the ignorance approach, we determined if they were mentioned or cited in the paper. If not, we looked in the literature to corroborate our findings.

4. Results

4.1. The ignorance-base: The power of combining ignorance and biomedical concept classifiers

The ignorance-base captured the connection between our collective scientific ignorance (ignorance taxonomy) and knowledge (PheKnowLator) through sentences from the literature and yielded a wealth of data (see Figure 5) for future study via the network (see Figure 2). The short manual review of some random sentences from the ignorance-base suggested that both the ignorance and biomedical concept classifiers correctly identified concepts (data not shown). Combining these two types of classifiers enhanced the exploration methods.

Creating an expanded gold-standard ignorance corpus (see Table 4) yielded ignorance classifiers achieving F1 scores around or above 0.8 with many closer to 0.9, on both the sentence- and word-levels (see Tables 5 and 6). The ensemble of 13 different binary classifiers performed the best for both classification tasks and was used for all ignorance classification for the ignorance-base. Further, the COVID-19 search engine [20] PubMedBERT model (binary) achieved an F1 score of 0.66 on our held out test data. Note that 105 sentences were excluded from our test set due to an input length limitation of their model (only 1900 sentences were evaluated). This compares to our ignorance BioBERT binary sentence classification model at 0.95. The COVID-19 search engine [20] cannot predict all the ignorance statements. Overall, our



Nodes	857,766	780,753	1,638,519
Edges	1,815,645	5,072,064	6,887,709

Figure 5: Summary information for the ignorance-base. The ignorance-base is a combination of biomedical concept classifiers and ignorance classifiers over a corpus of prenatal nutrition articles. The network representation connected the literature to the ignorance theory and biomedical concepts via PheKnowLator [26, 27].

high-quality classifiers, built on the expanded ignorance corpus, allowed us to scale up our system for the ignorance-base.

For the corpus, the annotation guidelines may be generalizable, with five different annotators over all annotation tasks. We can trust the annotations and reliability of the guidelines as the IAA, as measured by an F1 score, was near or above 80% for the classic annotation task [150, 151, 160] and for the split annotations, the annotators were correctly identifying statements of ignorance around 90% of the time. The task was quite difficult, requiring the pre-processing of the documents, extensive training, and many examples of ignorance statements. Disagreements involved the annotators choosing different lexical cues that signified ignorance for a given sentence, different semantic interpretations of a sentence, and the need for clarification on the ignorance categories. Consensus was reached during the adjudication process and the annotation guidelines and ignorance taxonomy were updated based on those discussions. Our annotation guidelines were robust and reproducible in two different annotation tasks. (For more information on the corpus itself see the Supplementary File on Corpus Information.)

Table 4: Interannotator Agreement (IAA): IAA is calculated as F1 score for all annotation tasks. The IAA for the training is between the two annotators, not including the previous gold-standard. *F1 score between annotator and the final gold-standard version after adjudication with M.R.B.

Annotation batch	category IAA	scope IAA	fuzzy category IAA	fuzzy scope IAA
Prior work (60 articles)	78%	87%	79%	90%
Training (8 articles)	77%	66%	78%	87%
K.J.S. and S.A. (8 articles)	81%	82%	81%	93%
K.J.S. with M.R.B adjudicator* (12 articles)	88%	92%	89%	95%
S.A with M.R.B adjudicator* (12 articles)	89%	92%	90%	96%
All combined	82%	87%	83%	92%

Table 5: Sentence Classification: the best model for sentence classification for each approach: (1) ALL CATEGORIES BINARY: binary classification ignorance or not, (2) AN ENSEMBLE OF BINARY CLASSIFIERS: binary classification for each class (reported) combined to create the ensemble, and (3) ALL CATEGORIES COMBINED: one multi-classifier to all categories.

Ignorance Category	Model	testing F1 score	testing support
ALL CATEGORIES BINARY	BioBERT	0.95	2005
answered question	BERT	0.97	168
explicit inquiry	BioBERT	0.9	92
unknown/novel	BioBERT	0.88	63
incompletely understood	ANN	0.83	225
indefinite relationship	BERT	0.87	1072
largely understood	BERT	0.9	312
anomalous/curious	BERT	0.96	149
alternative/controversy	BioBERT	0.79	441
difficult task	BERT	0.95	93
problem/complication	BioBERT	0.9	202
future work	BioBERT	0.85	195
future prediction	BERT	0.88	55
important consideration	BERT/BioBERT	>0.99	491
ALL CATEGORIES COMBINED	BioBERT	0.12	2005

Our results suggest that ignorance statements proliferate throughout the ignorance-base. The 1,643 articles, spanning years 1939 to 2018 (see Fig-

Table 6: Word Classification: the best model for word classification for each approach: (1) ALL CATEGORIES BINARY: binary classification ignorance or not, (2) AN ENSEMBLE OF BINARY CLASSIFIERS: binary classification for each class (reported) combined to create the ensemble, and (3) ALL CATEGORIES COMBINED: one multi-classifier to all categories. *Reporting the average F1 score of all the categories for one multi-classifier.

Ignorance Category	Model	testing F1 score	testing support
ALL CATEGORIES BINARY	BioBERT	0.89	7601
answered question	BioBERT	0.89	320
unknown/novel	CRF	0.98	155
explicit inquiry	BioBERT	0.97	43
incompletely understood	BioBERT	0.93	2809
indefinite relationship	BioBERT	0.97	1205
largely understood	BioBERT	0.94	618
anomalous/curious	BioBERT	0.96	399
alternative/controversy	BioBERT	0.91	598
difficult	CRF	0.93	128
problem/complication	BioBERT	0.9	238
future work	BioBERT	0.89	391
future prediction	BioBERT	0.94	100
important consideration	BioBERT	0.93	608
ALL CATEGORIES COMBINED*	BioBERT	0.82	6239

ure 6), contained 327,724 sentences with over 11 million words. Just over half of those sentences had an ignorance lexical cue (182,892), with articles averaging a total of 111 cues (with a median of 93). Every section of the articles had ignorance cues aside from the title, with the most in the discussion and conclusion sections and the fewest in the abstract and results. Our collective scientific ignorance is abundantly represented throughout the literature.

Further, our ignorance taxonomy is not only a categorization system of ignorance statements via knowledge goals, but also a depiction of the research life-cycle and how researchers discuss our collective scientific ignorance (see Figure 7 with proper definitions and example lexical cues in Table 2). (Note that we renamed our ignorance categories from our previous work [4] to

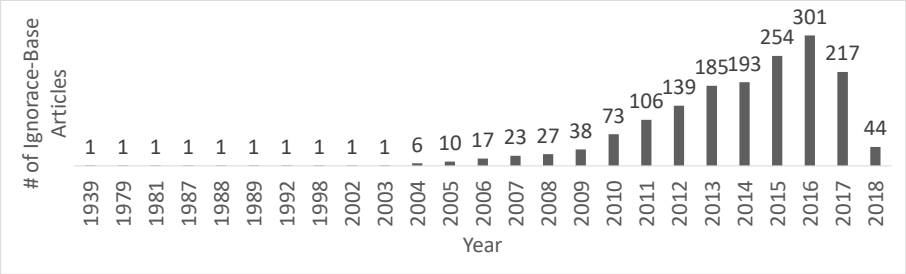


Figure 6: Article date distribution for the ignorance-base (1939-2018).

Table 7: Ablation study: Results of the ablation study on the sentence classification level.

Ignorance Category	Model	testing F1 score	testing support
ALL CATEGORIES BINARY	BERT	0.33	2005
answered question	BERT	0.47	168
explicit inquiry	BERT/BioBERT	0.35	92
unknown/novel	BERT/BioBERT	0.25	63
incompletely understood	BERT	0.27	225
indefinite relationship	BERT	0.52	1072
largely understood	BERT	0.49	312
anomalous/curious	BERT	0.58	149
alternative/controversy	BERT	0.23	441
difficult task	BERT	0.35	93
problem/complication	BioBERT	0.43	202
future work	BioBERT	0.61	195
future prediction	BERT	0.22	55
important consideration	BERT/BioBERT	0.46	491

Table 8: Lexical cue generalizability study: Results of overlaps between our lexical cue list (2,513 cues) and other similar works from different domains.

Similar work	cues support	matching support
BioScope corpus [58]	46	45 (98%)
GENIA project (meta-knowledge) [156, 98]	1374	591 (43%)
COVID-19 search engine [20]	286	166 (60%)

be more ontologically precise based on discussions with an ontologist, Mike Bada.) Underneath the 13 categories of ignorance (with 3 broader ones in all caps in Figure 7) were 2,513 unique lexical cues collected from related work or added during our annotation tasks. 1,822 of them had examples in our ignorance-base. Further, the ignorance classifiers found 5,637 new unique lexical cues that signify ignorance. These new cues were added to the ignorance-base and noted as such. Many of these cues were variations of ones already captured and others were new, such as "not as yet", "interplay", and "have begun to illuminate". Our ignorance classifiers recognized more complex language than just a dictionary match. In addition, the classifiers quite heavily relied on the lexical cues as features (see Table 7). Our ablation study showed poor performance without them. We found that the performance on the held-out test set dropped significantly for all ignorance categories, with an average F1 score of 0.39. Further, our cues seem to generalize beyond prenatal nutrition and our corpus based on the many overlaps between our cue list and prior work from different biomedical domains (see Table 8). Overall, there were 517,445 ignorance annotations involving 7,459 unique lexical cues; this reinforces the diversity of ways that ignorance is expressed in the literature.

Our ignorance-base also contained a wealth of different types of biomedical concepts. Our biomedical concept classifiers identified 720,313 concepts involving 19,883 unique concepts from all of the ontologies and almost all of them (95.3%) mapped to PheKnowLator. Note that we can only represent the biomedical concepts in PheKnowLator that were also captured by our biomedical concept classifiers. This overlap included six of our eight ontologies (missing MOP and NCBITaxon) or six of the eleven PheKnowLator ones (missing the human phenotype ontology, MONDO disease ontology, vaccine ontology, pathway ontology, and cell line ontology). Because our biomedical concept classifiers predict identifiers character by character (see our prior work for more details [25]), they can produce identifiers that do not exist.

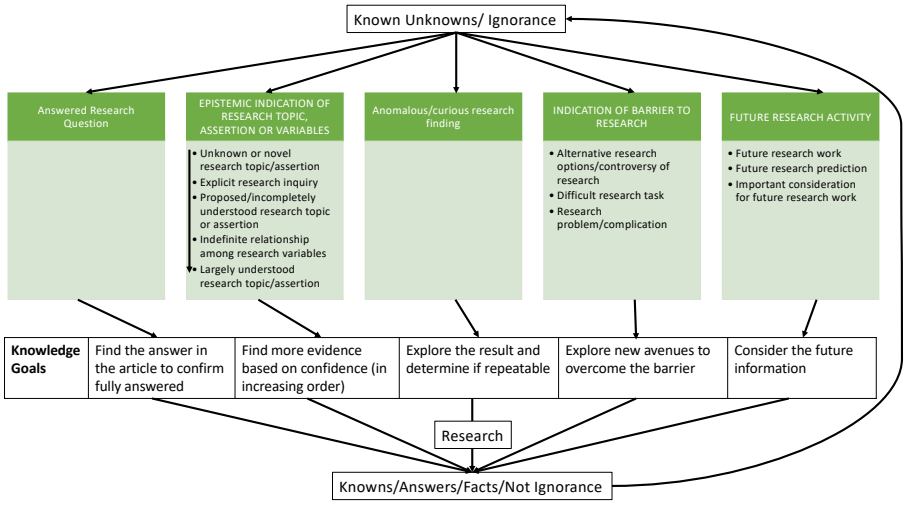


Figure 7: Ignorance taxonomy embedded in the research context: Starting from the top, research starts from known unknowns or ignorance. Our ignorance taxonomy is in green (an ignorance statement is an indication of each ignorance category) with knowledge goals underneath. Research is then conducted based on the knowledge goals to get answers; these then filter back to the known unknowns to identify the next research questions.

In terms of errors, our classifiers predicted 850 (4%) unique OBO concepts with non-existent OBO identifiers. The other 1,432 (7%) unique OBO concepts that did not map seemed to either be from the two ontologies not included in PheKnowLator (MOP and NCBITaxon) or were terms no longer used/depreciated from the ontologies. The ignorance-base captured many biomedical concepts. Overall, the ignorance-base contained a great deal of data consisting of all different types of lexical cues, ignorance categories, and OBO concepts.

4.2. Focusing on ignorance statements provides an alternative targeted exploration of a topic that is distinct from the standard approach

Focusing on ignorance statements provided researchers interested in the topic of vitamin D with new avenues of exploration that were distinct from the standard literature approach and the COVID-19 search engine [20] (see

Figure 8). We present results for the ignorance approach (vitamin D ignorance statements), the COVID-19 search engine [20], the standard approach (vitamin D sentences only), and a comparison of all three.

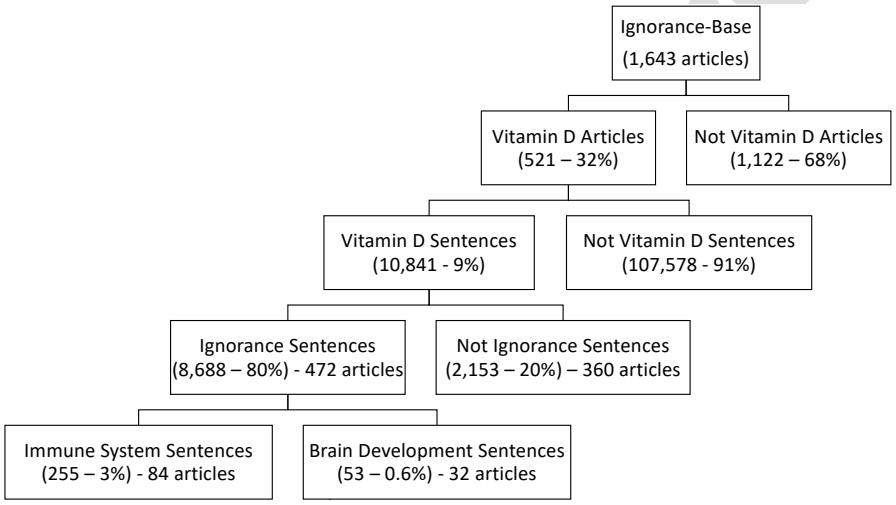


Figure 8: Exploring the ignorance-base by Vitamin D: Searching the ignorance-base for vitamin D yielded many articles and sentences that can be explored using ignorance statements to find new research questions, including immune system and brain development.

There is a great deal of research on vitamin D and more specifically a plethora of ignorance statements. Searching through the ignorance-base for the four vitamin D terms yielded 521 articles with 10,841 sentences mentioning vitamin D (9% of the 118,419 sentences in the 521 articles and 3% of all the sentences in the ignorance-base) (see Figure 8). Note that only the terms VITAMIN D and VITAMIN D2 pulled out sentences from the ignorance-base. These sentences included 17,584 unique biomedical concepts excluding the VITAMIN D concepts (88% of the total unique biomedical concepts). Of those VITAMIN D sentences, 8,688 sentences (80%) were ignorance statements spanning 472 articles. The COVID-19 search engine [20] model run on the 10,841 vitamin D sentences predicted 4,315 direction or

challenge statements (40%), achieving an agreement of 0.66 (F1 score) with our ignorance classification. We explored this data to differentiate between knowns and unknowns.

4.2.1. Term Frequency

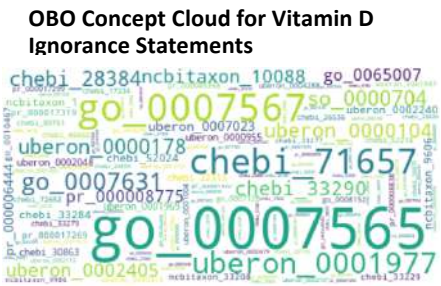
Focusing on term frequency provided some concepts of interest. The top five most frequent biomedical concepts for the ignorance approach included: FEMALE PREGNANCY, PARTURITION (giving birth), VERSICONOL ACETATE, BLOOD SERUM, and FEEDING BEHAVIOR (see Figure 9a). The first two aligned with the corpus theme of prenatal nutrition. VERSICONOL ACETATE is an intermediate in the biosynthesis of aflatoxin, a toxin produced by mold that may be toxic towards the vitamin D receptor in relation to rickets [161]. Vitamin D levels are mainly measured from the BLOOD SERUM, and FEEDING BEHAVIOR seems to highlight the importance of ingesting vitamin D. For the words, the most frequent terms were supplementation, maternal, status, levels, and women and they also fit with the theme: supplements are suggested for many people, maternal and women fit with the corpus theme, and status and levels are measurement terms for vitamin D. None of these terms were surprising, which was a good sign that we captured meaningful information.

Term frequency can help prioritize areas to explore. For example, FEEDING BEHAVIOR (GO:0007631), defined as the behavior associated with the intake of food, was an interesting concept in relation to vitamin D. Vitamin D is naturally absorbed through sunlight and digestion. To corroborate our findings, we searched for "vitamin D" and "feeding behavior" in the literature, and found that vitamin intake during pregnancy in general seems to affect both the metabolic system and food intake regulatory pathways in the offspring [162]. This result argues that concepts that appear frequently with a topic can provide useful keyword search terms for researchers.

Further, frequent concepts can lead to some pertinent questions for the researchers. The ignorance statements for FEEDING BEHAVIOR and VI-

TAMIN D, provided research ideas. Most of these ignorance statements discussed the ingestion of vitamin D mainly via supplements, but also with some foods. The recommendations for ingestion all varied by study (agreeing with the findings from a systematic review [163]). One ignorance statement stood out specifically, "the high prevalence of Vitamin D deficiency in PREGNANT women is a worldwide health problem regardless of latitude, FOOD INTAKE or socio-economic status [93]" (PMC5941617) [164], citing a systematic review and meta-analysis that looked at vitamin D status globally [165]. The ignorance categories were *important consideration* and *incompletely understood*, meaning it is an urgent topic and more evidence is needed to understand the reasons. Note how our ignorance categories build on each other and provide actionable next steps through our taxonomy. The COVID-19 search engine [20] classified this statement as a challenge, meaning "a sentence mentioning a problem, difficulty, flaw, limitation, failure, lack of clarity, or knowledge gap." [20] This does not provide specific actionable next steps. Looking at the review, all of the studies recommended vitamin D supplementation, but we could not find any studies that determine why supplementation is so low. This is an urgent matter based on the ignorance statement. How do supplements, specifically for vitamin D, fit into feeding behavior? A potential research topic could be to study what specific factors, beyond general socio-cultural factors, lead to women taking vitamin D supplements as part of their diet, especially for pregnancy. Studying this topic could lead to novel methods that help mothers stay vitamin D sufficient throughout pregnancy, resulting in fewer adverse outcomes for both mother and offspring (see Figure 1). Thus, biomedical concept frequency combined with ignorance statements and categories can lead to a high impact research topic that could affect mothers of childbearing age and their offspring globally.

Term frequency can also highlight when the context of a term is more known or unknown. Comparing the ignorance approach to the standard



OBO Concept Cloud for Vitamin D Ignorance Statements

OBO ID	OBO ID Label	Frequency
go_0007565	female pregnancy	100.00%
go_0007567	parturition	36.06%
chebi_71657	versiconol acetate	29.21%
uberon_0001977	blood serum	22.45%
go_0007631	feeding behavior	14.73%

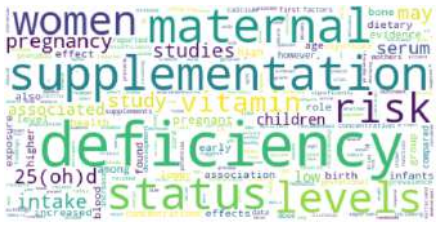
Word Cloud for Vitamin D Ignorance Statements

Word	Frequency
supplementation	14.34%
maternal	13.75%
status	13.31%
levels	12.36%
women	10.98%

(a) Ignorance Approach: Vitamin D ignorance statements

OBO Concept Cloud for Vitamin D

OBO ID	OBO ID Label	Frequency
go_0007565	female pregnancy	100.00%
go_0007567	parturition	37.17%
chebi_71657	versiconol acetate	29.50%
uberon_0001977	blood serum	23.73%
go_0007631	feeding behavior	16.38%



Word Cloud for Vitamin D

Word	Frequency
deficiency	13.79%
supplementation	13.48%
maternal	13.39%
status	12.25%
levels	11.78%

(b) Standard Literature Approach: Vitamin D sentences

Figure 9: Term frequency results: Frequent Biomedical Concepts and Words in (a) ignorance approach vitamin D ignorance statements and (b) standard literature approach vitamin D sentences. Word clouds using words and biomedical concepts are on the right and left respectively. Also underneath are frequency tables of the top 5 most frequent concepts or words.

literature approach, the most frequent concepts in the standard approach were the same as the ignorance approach (see Figure 9b). This suggests that term frequency may not capture the difference in biomedical subjects between the two approaches. However, the top five words slightly differed between them: the word "deficiency" was the top most frequent term in the standard approach and the term "women" disappeared (see Figure 9). This may signify that vitamin D deficiency was established information, resulting in a lack of ignorance. At the same time, all these terms may be more unknown than known. Recall that 80% of the vitamin D sentences were ignorance statements, so it is possible that much of the context around VITAMIN D was still unknown in general. The ignorance frequency term list not only provided an avenue for exploration, FEEDING BEHAVIOR, with a potential research topic, but in addition, may help distinguish between terms that describe probable knowns, like "deficiency", and those connected to more open questions, like FEEDING BEHAVIOR.

Comparing these ignorance results to the COVID-19 search engine [20] model run on our data found the same five most frequent OBO concepts as above, while the online search engine provided a different set of top frequent MeSH terms. It is interesting that all approaches agreed on the same OBO concepts. Maybe this actually corroborates our findings. We also looked at the online COVID-19 search engine [20] because of this: searching for "vitamin D" and "pregnancy" resulted in 26 sentences. Within those sentences, the five most frequent terms were vitamin D deficiency, asthma, autoimmune diseases, child, and placenta with only 2-3 sentences for each of them. Note that the small number of sentences was probably due to the differing underlying themes of the corpora (COVID-19 mainly vs. prenatal nutrition). In comparing all frequency lists, we found the word "deficiency" in the standard approach with almost 14% of the corpus containing it. The terms "child" and "placenta" fit the theme of pregnancy and were similar to the terms found in the ignorance approach. The terms "asthma" and "autoimmune disease"

raised possible avenues to explore, but were underrepresented with only two sentences per topic. Our ignorance approach provided more sentences, visualizations, and avenues to explore for researchers interested in prenatal nutrition topics including vitamin D. At the same time we acknowledge that all approaches provided interesting avenues for exploration.

4.2.2. Ignorance Enrichment

To go beyond term frequency and further distinguish between known and known unknowns, ignorance enrichment found at least three interesting new avenues to explore in relation to vitamin D that were captured by the standard approach, but buried amongst 275 concepts, and one avenue not captured by the standard approach at all. Note that the COVID-19 search engine [20] did not calculate enrichment and so we only compared to their frequency list. The ignorance approach found 11 ignorance enriched concepts, whereas the standard approach found 275, with an overlap of eight concepts (see Figure 10). However, only focusing on the overlapping concepts, in the standard approach most of them were buried far down the list of enriched concepts ordered by enrichment p-value (indicated by the parentheses next to the overlapped concepts in Figure 10). Further, in comparing the two different approaches, the ignorance approach found concepts from broader categories, including IMMUNE SYSTEM and BRAIN DEVELOPMENT, compared to the standard approach which were more specific entities, such as BLOOD SERUM and VITAMIN K. Ignorance enrichment provided the researchers with a smaller list of targeted statements of knowledge goals to potentially pursue or spark ideas from.

Focusing on the ignorance-enriched concepts also provided research topic ideas. T.L.H. and M.R.B. determined that IMMUNE SYSTEM, RESPIRATORY SYSTEM, and BRAIN DEVELOPMENT (all captured by the standard approach) were all interesting in relation to vitamin D. Intriguingly, the COVID-19 online search engine [20] frequency list included the terms "autoimmune diseases" and "asthma", which are subsets of IMMUNE SYS-

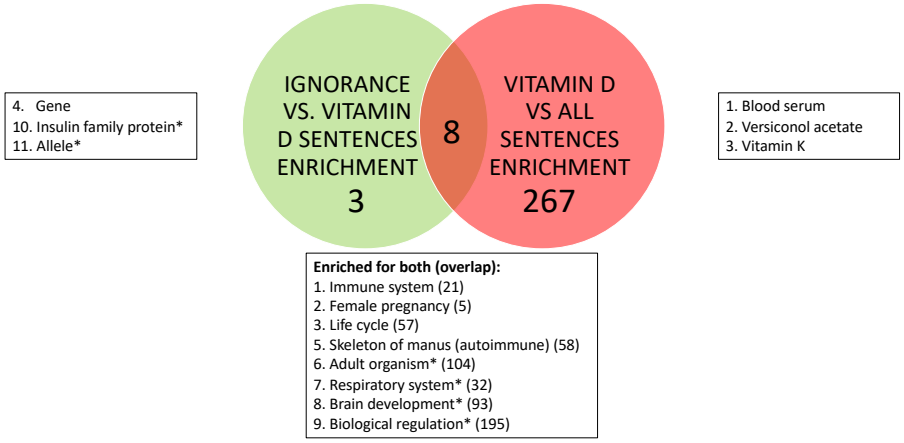


Figure 10: Comparison of standard and ignorance enrichment: A Venn diagram of biomedical concept enrichment between just vitamin D (pink) and ignorance vitamin D (green) sentences. Next to each bubble are concepts in their respective enrichment orders. The concepts in the middle are the overlap and the numbers correspond to the enrichment position for the ignorance vitamin D enrichment, with the overlap position in parentheses. Skeleton of manus is an error and is actually annotating autoimmune as in the parentheses. *Statistically significant with FDR but not family-wise error.

TEM and RESPIRATORY SYSTEM, respectively. This may be evidence that corroborates our findings. All of these concepts were currently studied, with more room for future work. We also found the insulin family protein (not captured by the standard approach) intriguing because many studies have attempted to determine the link between vitamin D and gestational diabetes mellitus [166]. All of these concept areas are ripe for exploration.

We explored the specific ignorance statements for the concepts of interest to narrow in on a research topic, just as with FEEDING BEHAVIOR. We chose to look at the IMMUNE SYSTEM ignorance statements, which provided the researcher with 255 ignorance statements plus their entailed knowledge goals, spanning 84 articles (see Table 9 for the top eight articles with the most ignorance statements). Note that only one article had no ignorance statements that included VITAMIN D and IMMUNE SYSTEM. Thus,

only using the ignorance statements themselves, we have already found a set of articles and sentences for the researchers to review for a potential research topic.

Table 9: Articles with the most ignorance statements: The top eight articles for vitamin D and immune system in order of the most ignorance statements.

PMCID	Article Title	Date	# of ignorance statements	# of non-ignorance statements
PMC4448820	Inflammation and Nutritional Science for Programs/Policies and Interpretation of Research Evidence (INSPIRE)	5/15	22	0
PMC4251419	Vitamin D and immunity	12/14	17	0
PMC3717170	Vitamin D: beyond bone	5/13	13	1
PMC3277098	Vitamin D and allergic disease: sunlight at the end of the tunnel?	12/11	13	0
PMC4889866	Maternal Vitamin D Level Is Associated with Viral Toll-Like Receptor Triggered IL-10 Response but Not the Risk of Infectious Diseases in Infancy	5/16	11	0
PMC3347028	Vitamin D and its role during pregnancy in attaining optimal health of mother and fetus	3/12	10	0
PMC5489519	Vitamin D Modulation of TRAIL Expression in Human Milk and Mammary Epithelial Cells	6/17	8	0
PMC4302429	Vitamin D deficiency decreases survival of bacterial meningoencephalitis in mice	1/15	8	1

Our ignorance taxonomy includes 13 categories of unknowns that can help researchers narrow their search. So we continued to narrow our search using ignorance-category enrichment, since choosing the IMMUNE SYSTEM was still quite a large topic with lots of ignorance statements. Understand-

ing and tracing ignorance categories over time can both help the researchers narrow in on a research topic and also show how the questions in a field are asked more broadly. VITAMIN D ignorance statements in general employed a wide-range of different ignorance categories (see Figure 11), spanning all the thirteen categories of ignorance. Ten were enriched in VITAMIN D ignorance statements as compared to all ignorance statements (see the green highlights in Figure 11). For example, *unknown/novel* was enriched in this domain, pointing to large unknowns about the context of VITAMIN D in pregnancy and fetal development. To also understand how these questions changed over time, the bubble plot for IMMUNE SYSTEM and VITAMIN D ignorance statements (see Figure 12) showed that *unknown/novel* was spread out amongst the different articles. This suggests that researchers have not resolved their broadest unknowns in this field over time, in which case it may be a good knowledge goal area for a research topic. Thus, the researchers can continue to narrow in on a research topic not only with a biomedical concept, such as IMMUNE SYSTEM, but also with an ignorance category, such as *unknown/novel*.

We chose to dive deeper into the *unknown/novel* category for VITAMIN D and IMMUNE SYSTEM, with the goal to find pertinent questions as an example of exploration by topic. There were many *unknown/novel* ignorance sentences to investigate here. Below are some ignorance sentences (lowercase) from this set with the biomedical concepts capitalized and the ignorance lexical cues underlined:

1. "in the last five years, there has been an explosion of published data concerning the IMMUNE effects of VITAMIN D, yet little is known in this regard about the specific IMMUNE effects of VITAMIN D during PREGNANCY." (PMC3347028)
2. "these results describe novel mechanisms and new concepts with regard to VITAMIN D and the IMMUNE SYSTEM and suggest therapeutic targets for the CONTROL of AUTOIMMUNE diseases." (PMC3717170)

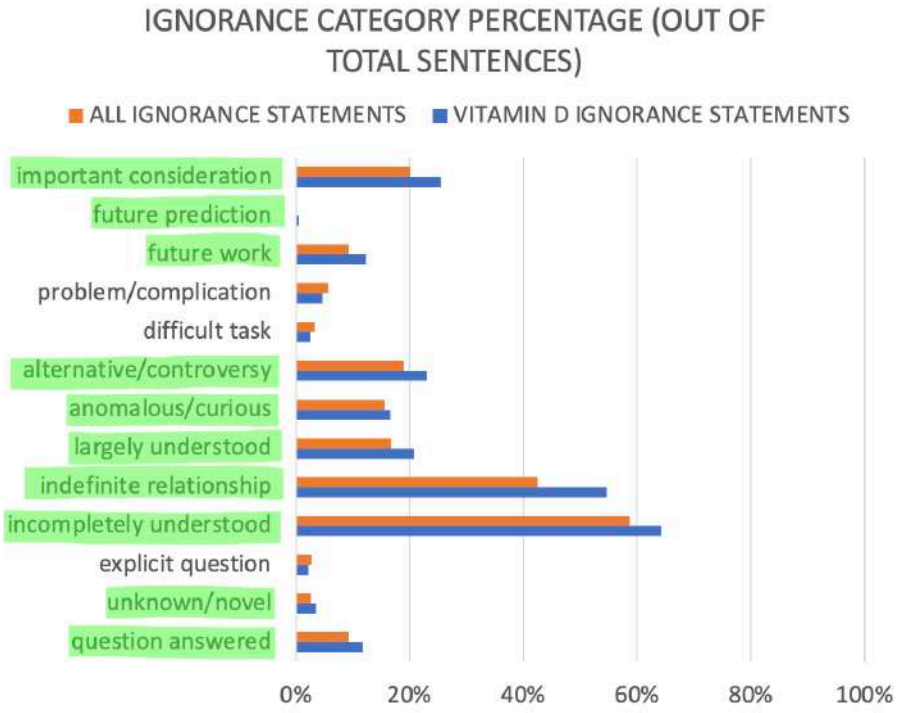


Figure 11: Ignorance-category enrichment: Ignorance vitamin D sentences compared to all ignorance sentences. The 10 categories highlighted in green were enriched.

3. "however, findings regarding the combined effects of PRENATAL and POSTNATAL VITAMIN D status on fs [food sensitization], two of the most critical periods for IMMUNE SYSTEM DEVELOPMENT (19,20), are unclear." (PMC3773018)
4. "it has an important role in BONE HOMEOSTASIS, BRAIN DEVELOPMENT and MODULATION OF the IMMUNE SYSTEM and yet the impact of ANTENATAL VITAMIN D deficiency on infant outcomes is poorly understood." (PMC4072587)
5. "background: VITAMIN D is known to affect IMMUNE function; however

it is uncertain if VITAMIN D can alter the IMMUNE RESPONSE towards the persistent herpesviruses, EBV and CMV." (PMC4113768)

(Note that not all biomedical concepts were recognized by the biomedical concept classifiers.) The overall research topic or knowledge goal based on these statements was the need to explore the relationship between VITAMIN D and the IMMUNE SYSTEM especially in pregnancy (ignorance categories *indefinite relationship* and *incompletely understood*). The same methods can be used for the other top enriched concepts including BRAIN DEVELOPMENT (data not shown). For BRAIN DEVELOPMENT, the overarching knowledge goal was the need to determine if VITAMIN D and BRAIN DEVELOPMENT were truly linked (ignorance categories *largely understood* and *indefinite relationship*). Thus, from querying the ignorance-base for the topic VITAMIN D, the researchers now have knowledge goals to pursue in specific concept areas. (see Figure 8). Our exploration by topic methods provided multiple starting points for this research, in more depth than the standard approach and the COVID-19 search engine [20] can supply.

*4.3. Connecting experimental results (*e.g.*, a gene list) to ignorance statements can identify questions that may bear on it, providing new avenues for exploration, potentially from other fields*

Similar to exploration by topic, exploration by experimental results provided the ignorance context for a gene list as possible future work for the researchers. Note that this was made possible by the OBOs and that neither the standard literature approach nor the COVID-19 search engine [20] have this capability. Connecting a vitamin D and sPTB gene list from a paper [36] to ignorance statements found a new avenue for exploration, BRAIN DEVELOPMENT, that was not mentioned in the paper, and pointed to an implied field, neuroscience, as a possible source for answers.

Following the exploration by experimental results pipeline (see Figure 4), the 43 genes mapped to 782 OBO concepts. These OBOs connected to 51,868

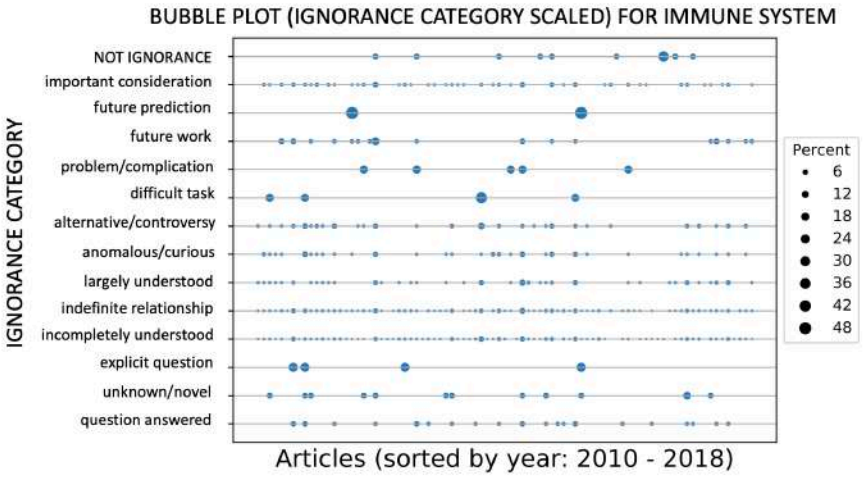


Figure 12: How ignorance changes over time: A bubble plot of vitamin D and immune system sentences (including non-ignorance sentences). The x-axis is the articles sorted by time. The y-axis is the ignorance categories. Each bubble represents the portion of sentences in each article in that ignorance category (scaled by the amount of total ignorance sentences in the category). For example, *future prediction* only appears in two different articles and is basically split in half between both.

sentences (1,590 articles) that included 17,586 unique OBO concepts (88% of the total unique OBO concepts), of which, 33,885 sentences (1,537 articles) were ignorance statements with 11,711 unique OBO concepts (59% of the total unique OBO concepts). This suggests that the majority of sentences connected to these genes were ignorance statements (65%). These data can be explored by topic using the OBO list, but we focused on the three new analyses. The three new analyses were helpful to digest both the many OBO concepts and the many statements of ignorance connected to the gene list to provide areas of research to explore in future work.

With the gene list connected to so many concepts (782), combining gene list coverage and ignorance enrichment helped prioritize concepts to explore (see Table 10). The highest covered OBO concept was PROTEIN CODING GENE (SO:0001217). In the top 25 most covered OBO concepts, all con-

cepts were established information, currently studied, or not significant (see Figure 3). (None were emerging topics.) Note that some had no information from the literature-side, meaning no conclusions could be drawn based on the current information. In fact, all 18 concepts enriched in ignorance, were also enriched in all gene list sentences, including: PROTEIN CODING GENE, INNATE IMMUNE RESPONSE, GENE, IMMUNE RESPONSE, and BRAIN in the top 25. As before, concepts related to IMMUNE SYSTEM and BRAIN surfaced after ignorance enrichment. These concepts were ripe for exploration in relation to the gene list to find knowledge goals that may bear on them.

Combining other canonical enrichment methods with ignorance enrichment also helped prioritize the many OBO concepts produced by our gene list (see Figure 13 focusing only on the gene ontology). DAVID [144, 145] found 42 of the 43 genes and mapped them to 159 GO concepts. 51 of those were enriched and 30 were contained in sentences found in the ignorance-base. Of those 30, 19 were contained in gene list statements and 11 had no information. Of those 19, 17 had at least one ignorance statement, and the concepts were mainly related to the immune system. (Two concepts, RESPONSE TO STRESS (GO:0006950) and MULTI-ORGANISM PROCESS (GO:0051704) had no ignorance statements.) The ignorance statements for the 17 concepts can be explored to provide more information to the canonical enrichment methods and their respective knowledge-bases.

To find out whether our ignorance lens could augment canonical methods, we compared our ignorance approach to the canonical approach. When comparing ignorance enrichment in GO to DAVID, we found more ignorance in general compared to established information. 3,173 GO concepts were enriched in ignorance with 159 in DAVID (see Figure 14). Intriguingly, the overlap between the two analyses was small: 60. If we look at enrichment it was even smaller: only two concepts overlapped. Potentially this makes sense as we were enriching for the opposite things: ignorance vs. established

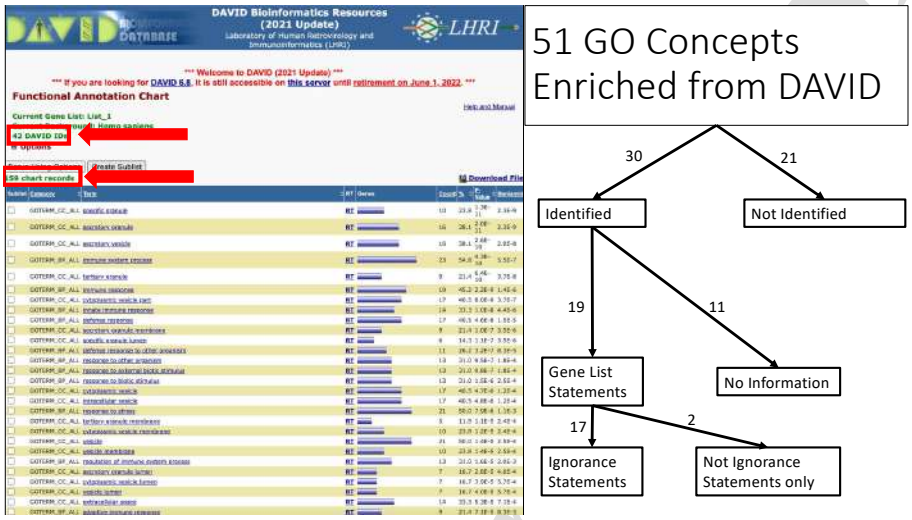


Figure 13: Enhancing canonical enrichment analysis using the ignorance-base: DAVID enrichment analysis for the gene ontology (GO) in relation to the ignorance-base. The DAVID initial analysis is on the left with 42 of the 43 genes found in DAVID mapping to 159 GO concepts. The right is a breakdown of where the 51 enriched GO concepts from DAVID fall within the ignorance-base.

knowledge. On the knowledge-side, most enriched concepts from DAVID pertained to the immune system, which was the main focus of Yadama *et al.*, [36]. The ignorance-side found more general biological processes, and two concepts from the overlap, IMMUNE RESPONSE and INNATE IMMUNE RESPONSE, also pertained to immunity. These concepts were currently studied.

Looking at the ignorance enriched concepts only, we achieved our goal of finding a new avenue to investigate that was not mentioned by the paper [36], namely the brain. Further, it also provided a different field to examine for answers, namely neuroscience. The top three ignorance-enriched GO concepts included FEMALE PREGNANCY, BIOLOGICAL REGULATION, and METABOLIC PROCESS (see Figure 14). Broadening this exploration beyond GO, there were 130 total ignorance enriched concepts for this gene

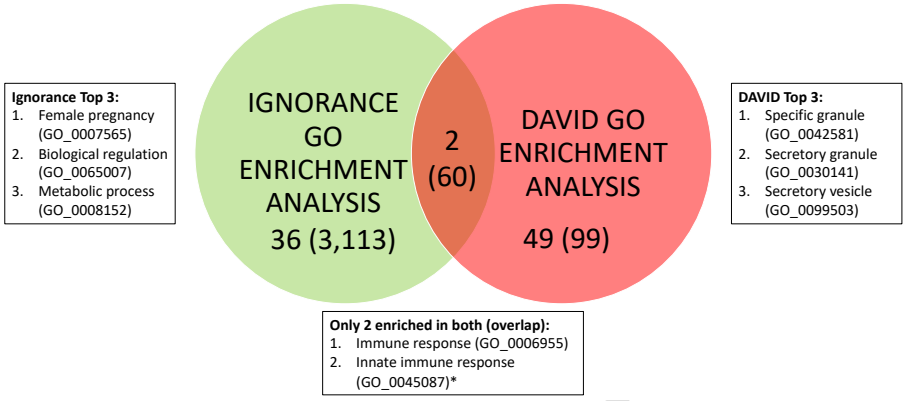


Figure 14: Comparison of DAVID and ignorance enrichment: A Venn diagram of gene ontology enrichment between DAVID (pink) and the ignorance-base (green). In parentheses are the total number of concepts found in each category without enrichment. Next to each bubble are the top three concepts for each enrichment method. The concepts in the middle are the overlap. *Statistically significant with FDR but not family-wise error.

list, including the 38 from GO. There were still some immune related concepts including (in order of enrichment): IMMUNE SYSTEM, IMMUNE RESPONSE, SKELETON OF MANUS (autoimmune), INTERLEUKIN-1 FAMILY MEMBER 7*, and INNATE IMMUNE RESPONSE*. (The * means that they were statistically significant with FDR but not family-wise error.) As mentioned above, there was still future work to understand the IMMUNE SYSTEM in relation to sPTB and VITAMIN D [36], and the ignorance approach provided specific ignorance statements to explore it. Even more striking though was the number of ignorance enriched concepts related to the BRAIN (12): BRAIN, BRAIN DEVELOPMENT, COGNITION, NERVOUS SYSTEM DEVELOPMENT, NEURON, LEARNING, SYNAPSE, NERVOUS SYSTEM, CENTRAL NERVOUS SYSTEM, NEUROTRANSMITTER*, NEURAL TUBE*, and NEUROGENESIS. Neither Yadama *et al.*, [36] nor DAVID [144, 145] derived any brain-related concepts from this gene list, signifying that this association was not yet established knowledge.

The developing brain may be an emergent topic related to vitamin D and spontaneous preterm birth, and neuroscience may shed light on these connections. Along these lines, we present some excerpts (lowercase) from five papers to explore below (the biomedical concepts are capitalized and the ignorance lexical cues are underlined):

1. "this AREA OF the BRAIN, specifically the FRONTAL CORTEX, is important for LANGUAGE, MEMORY and higher order COGNITIVE functioning, including purposeful, goal-directed behaviours which are often referred to as executive functions.[4] the importance of adequate dha [docosahexaenoic acid] during this key period of BRAIN DEVELOPMENT is indicated in studies of preterm infants who are denied the full GESTATION period to accumulate DHA" (PMC4874207)
2. "discussion: we review relevant literature suggesting in utero inflammation can lead to PRETERM labor, while insufficient development of the GUT-BLOOD–BRAIN barriers could permit exposure to potential neurotoxins." (PMC3496584)
3. "a major Intake of DHA in the BRAIN happens in the last TRIMESTER of PREGNANCY; therefore, preterm infants are disadvantaged and have decreased BRAIN concentration of this vital lcpufa [long-chain polyunsaturated fatty acid]." (PMC3607807)
4. "at present, preterm infants have a limit of viability (50% survival rate) of around 23–24 weeks ga [gestational age] so post-NATAL nutrition will always be introduced during the second major phase of BRAIN growth, resulting in differences mainly in WHITE MATTER." (PMC3734354)
5. "the most likely explanation seems to be related to the timing of the nutrition event, since the infants were BORN at term rather than preterm when different developmental processes are occurring in the BRAIN." (PMC3734354)
6. "the period between their PRETERM BIRTH and term BIRTH at 40

weeks, a time when the major BRAIN spurt is occurring), was spent ex utero in these infants; this exposure to environmental influences, at an early stage of BRAIN DEVELOPMENT, might be expected to increase their vulnerability to dietary effects." (PMC3734354)

7. "the authors conclude by saying that reducing the energy deficit by improving early nutrition in preterms may improve the growth and maturation of the BRAIN." (PMC3734354)
8. "iron status, more commonly assessed in PREGNANCY, is not only important in HEMATOPOESIS and NEUROLOGICAL and COGNITIVE DEVELOPMENT[9] but plays a crucial role in CARNITINE SYNTHESIS,[10] although CARNITINE precursors may be more important.[11] zinc is an important COFACTOR for more than 300 identified zinc metalloenzymes.[12] zinc insufficiency in late PREGNANCY disrupts NEURONAL REPLICATION and SYNAPTOGENSIS,[13] and maternal deficiency is associated with decreased dna, rna, and protein content of the f1 BRAIN.[14] zinc deficiency affects one in five world inhabitants.[14]ZINC supplementation reduces the risk of PRETERM BIRTH, though not sga [small for gestational age].[14]

   VITAMIN D deficiency is under investigation for its role in protection against dm [diabetes mellitus], cv [cardiovascular], some ca [cancers], osteoporosis, and optimization of IMMUNE function.[15] VITAMIN D might be an important mediator in GUT HOMEOSTASIS and in signaling between microbiota and host.[16] the INTESTINAL microbiome in both newborns and LACTATING mothers influences infant and childhood FOOD allergy and eczema." (PMC4268639)

(Note that not all biomedical concepts were recognized by the biomedical concept classifiers. Also, the numbers in the sentences represent citations, which were superscript in the original article but were flattened for processing.)

In the paper, Yadama *et al.*, [36] focused on the mother's immune sys-

tem, but it is possible that this brain connection is instead focused on the effects of the spontaneous preterm birth on the offspring. Thus, a potential knowledge goal for the authors based on our analysis was to explore the connections between maternal VITAMIN D levels and spontaneous preterm birth through the maternal IMMUNE SYSTEM and the effects on the BRAIN DEVELOPMENT of the offspring. Exploring these connections would greatly impact mothers and their offspring globally. The ignorance-base provided a novel avenue (and field), BRAIN DEVELOPMENT (neuroscience), along with specific knowledge goal statements, that the authors can explore in future work based on starting from their initial gene list. Our exploration by experimental results method contextualized experimental results in the ignorance landscape, providing multiple avenues for future research, the immune system and the brain.

5. Discussion

Focusing on ignorance statements through our ignorance-base and exploration methods led to new research avenues that could help accelerate research. Further, the ignorance-base is more than just a literature search engine similar to Lahav *et al.*, [20]; it also provided insights, summaries, and visualizations based on topics and experimental results. Its focus on knowledge goals and its grounding in the OBOs helped our ignorance-base find areas of research with many questions and identify fields of study that may contain answers. The knowledge goals of the thirteen ignorance categories provided actionable next steps based on the inputs beyond what Lahav *et al.* provided with their two categories. Our goal was to provide researchers, students, funders, and publishers with actionable next steps based on a query. We demonstrated that through the field of prenatal nutrition, where the ignorance-base predicted areas of research that were currently studied and an emerging topic with a corresponding field that may prove fruitful for answers.

The exploration by topic method showed that vitamin D may play an important role in the immune system, respiratory system, and brain development (see Figures 8- 12 and Table 9). Corroborating these findings after 2018, when our corpus ended, recent review articles [10, 11, 12, 13, 14] included these areas as future work. These review articles required many hours of reading and synthesizing the literature article by article, whereas the ignorance method automatically offered not only articles, but also specific sentences that discuss knowledge goals for future work. For example, the sentence "it has an important role in BONE HOMEOSTASIS, BRAIN DEVELOPMENT and MODULATION OF the IMMUNE SYSTEM and yet the impact of ANTENATAL VITAMIN D deficiency on infant outcomes is poorly understood" (PMC4072587) [167] showed that further research on the impact of vitamin D on infant outcomes was needed. The context of the sentence is also important: it comes from the abstract objective section of a 2014 study in Rural Vietnam. Because our ignorance approach allows sorting of statements by time and section, we showed that since 2014, more research has been conducted on this topic [12]. Further, we can track how research questions emerge using our ignorance taxonomy (see Figure 12). Even this smaller-scale effort, limited to one broad topic and the years 1939-2018, demonstrated that we can map the landscape of our collective scientific ignorance and track how research questions evolve over time. Ideally, future work would create an ignorance-base over the entire body of scientific literature to provide this resource to researchers, students, funders, and publishers.

We further demonstrated that the ignorance-base and exploration by experimental results method can find an emerging topic (see Figures 4, 13, 14, and Table 10). Ignorance enrichment of the 43 genes in common between vitamin D and spontaneous preterm birth (sPTB) [36] found many concepts that relate to the brain and some that relate to the immune system (as found by [36]). This suggested that brain development could be an emerging topic in relation to vitamin D and sPTB. For example, consider the ignorance

statement: "discussion: we review relevant literature suggesting in utero inflammation can lead to PRETERM labor, while insufficient development of the GUT-BLOOD–BRAIN barriers could permit exposure to potential neurotoxins" (PMC3496584) [168]. This sentence ties all the relevant concepts together by suggesting that "vitamin D may be causing in utero inflammation leading to preterm labor; due to the preterm labor the gut-blood-brain barrier may develop incompletely, which in turn exposes the fetus to potential neurotoxins". Although this article was not cited by Yadama *et al.*, [36], it may posit a new knowledge area that needs to be explored further. Lastly, researchers could look to the field of neuroscience to help find relevant information to some of these knowledge goals. In consultation with our prenatal nutrition expert, here are some potential questions that could be explored:

1. What is the association between development of the gut-blood-brain barrier and whole-body inflammation and neuroinflammation in the context of fetal development?
2. How do the 43 genes relate to offspring brain development? Are any of them specifically related to offspring brain function?
3. What are the effects of vitamin D on lifecourse brain development generally?
4. How does spontaneous preterm birth effect offspring brain development compared to those born at term? Are there any remedies for said effect? Does nutrition play a role?
5. How does the gestational timing of nutrition and supplement exposure affect offspring brain development? What role do iron and zinc play in brain development?

To corroborate our findings, looking at the literature further showed that vitamin D and brain development may in fact be an emerging topic since 2018, the last year of our data. The connection between vitamin D and brain development has only recently been studied extensively. Looking for recent papers on "vitamin D", "brain development", and "spontaneous preterm

birth" in the literature (Google Scholar search on 9/13/2022), many review articles appeared ([10, 11, 12, 13, 14]) that discuss the impact of vitamin D on maternal and fetal health (see Figure 3 in [10] and our adapted Figure 1). All of these studies drew links between vitamin D, the immune system, and sPTB, and acknowledge at least one link between vitamin D and brain development. A 2022 review article stated that "recently, extensive scientific literature has been published determining the role of vitamin D in brain development" [10]. Note that these articles contain mentions of controversies and other types of ignorance statements, which also point to other areas of investigation. There appears to be room for exploration around the connections between vitamin D, sPTB, and brain development. Our ignorance approach could help automate review articles in finding the emerging topics to study. Understanding the ignorance-context around a set of genes in combination with the knowledge-context can help push the boundaries of our current understanding.

In general, we showed that ignorance-bases and knowledge-bases can enhance and complement each other. The ignorance-base itself was built upon a knowledge-base, PheKnowLator [26, 27]. We also utilized DAVID [145] as a comparison knowledge-base (see Figures 13 and 14) as well as other canonical methods (gene list coverage) to help prioritize the most relevant biomedical concepts (see Table 10). These analyses were made possible by grounding our ignorance-base in the OBOs, which allowed us to connect our ignorance-base to other knowledge-bases. At the same time, we did not use any of these methods to their fullest potential. First, only six ontologies overlapped between the biomedical classifiers and PheKnowLator, which limited the expansion of relevant concepts. Second, the method to create the OBO concept lists from both the vitamin D topic and the gene list were not very sophisticated (using only one step via the relations ontology). Finally, we did not use the knowledge-bases to determine if any ignorance statements have been answered. This is quite a hard problem that Lahav *et al.*, [20]

also did not tackle. All of these limitations could be addressed in future work. Further, there are many other knowledge-bases and methods that can be explored in relation to the ignorance-base.

The goal of this work was to demonstrate feasibility of the ignorance methods and we recognize that more improvements can be made. Our ignorance-base was created from automatic classifiers run over 1,643 articles in the prenatal nutrition literature. Any automation of this kind adds errors and all classification tasks can be improved upon to minimize it. For the ignorance classifiers, other parameter tunings and other algorithms, such as PubMedBERT [169], may yield improved results. Lahav *et al.*, [20] used PubMedBERT as a multi-label classification model for their two categories along with other algorithms. We found that an ensemble of binary models for each ignorance category worked the best for our thirteen categories (see Tables 5 and 6). We note that our sentence multi-classifier performed very poorly (0.12 F1 score). We believe this is due to the complexity of identifying ignorance in general and more specifically within one sentence alone where all the ignorance categories build on and interact with each other. In fact, our word multi-classifier performed quite well (0.82 F1 score) potentially suggesting that ignorance can be distinguished on the word-level and not as easily on the sentence-level due to the interplay of all the ignorance categories. In general, multi-classification problems are known to be quite difficult. Future work can look at other methods to improve it, including reframing the problem as multi-label as in Lahav *et al.* [20]. Overall though, our performance was quite good.

There is also future work with regards to our biomedical concept classifiers. They were developed using CRAFT [158, 157], a corpus of mouse articles, not prenatal nutrition, and it only included ten ontologies. Applying biomedical classifiers with more similar training data and more ontologies (*e.g.*, MONDO disease ontology and the phenotype ontology) would be beneficial (*e.g.*, PubTator [170]), although all of them have their pros and cons.

We ran these classifiers over only 1,643 prenatal nutrition articles. The scale of the ignorance-base was small; ideally we would create an ignorance-base that included all articles (or at least PMCOA to start). We focused only on the prenatal nutrition literature, and future work will determine if the ignorance taxonomy and methods generalize outside of it. But the extent of overlap between our cue list and similar prior work (see Table 8) implies that our ignorance-base may translate to other biomedical domains. In terms of exploration methods, concept enrichment provided more fruitful concepts (see Figures 10 and 11) than concept frequency (see Figure 9). Another avenue to explore would be co-occurrence terms. The creation of a tool (similar to [20]) incorporating more data analyses and visualizations techniques into a user-interface that allows researchers to interact with the system could make the ignorance-base easier to adapt to new environments. Future work could combine these efforts. Even with all these limitations, the current ignorance-base showed its power to find new research avenues to explore, providing insights, summaries, and visualizations beyond prior work [20]. We have just barely scratched the surface of what it can do. Collaborating with experts on vitamin D, delving into the topics introduced here, creating new methods, exploring other topics, and contextualizing other experimental results are obvious extensions of this work.

There is also future work in relation to the ignorance corpus and classifiers. Highlighting the lexical cues for the annotators before annotation could have biased the annotators. We did not measure this effect because our annotators found the task infeasible when the lexical cues were not highlighted. Even still, with the highlights, the annotators continued to find new lexical cues, which further extended the reach of the classifiers. We conducted an ablation study that determined the importance of the lexical cues as features for classifying ignorance statements (see Table 7). As mentioned, our annotators found them helpful. Lahav *et al.*, [20] agreed that the task was quite difficult with misleading keywords and so they added in sentences

without any of them. In contrast, we found that a larger cue list (2,513) was more resilient to error and helped our classifiers discover many more lexical cues (added in 5,637). Our IAAs were comparable to [20], in the 80% range. More data can always be annotated both to improve the current annotations and to add data from other fields besides prenatal nutrition. Confirming the ignorance taxonomy and classifiers generalize beyond prenatal nutrition will allow for the creation of a larger ignorance-base. More work needs to be done, but we showed that our lexical cues overlapped with prior work in other domains (see Table 8), hinting at the generalizability beyond our work here.

We demonstrated that a focus on ignorance statements through our ignorance-base and exploration methods can lead students, researchers, funders, and publishers to research avenues that are currently being studied or are emerging topics. Research begins from a foundation of established knowledge, and many knowledge-bases and ontologies exist to provide that. However, research continues through a process of posing questions and creating hypotheses to analyze and explore what is not yet understood. To facilitate that, we present the first ignorance-base based on knowledge goals and OBOs, along with two new exploration methods that provided insights, summaries, and visualizations of statements of unknowns, controversies, and difficulties needing resolution in future work. Just as the literature contains both knowledge and ignorance, so too can both knowledge-bases and ignorance-bases help researchers navigate the literature to find the next important questions or knowledge gaps.

6. Conclusion

Our ultimate goal was to create an ignorance-base and exploration methods to enable students, researchers, funders, and publishers to find the next important scientific questions or knowledge gaps. By augmenting and streamlining the manual work of literature reviews, we can help direct research to

focus on important quesitons and possible answers. The exploration by topic method not only found new avenues for exploration for researchers interested in vitamin D using our novel method of ignorance enrichment (the immune system, respiratory system, and brain development), but also elucidated how questions were asked and how that changed over time using our novel method of ignorance-category enrichment. Our exploration by experimental results method found an emerging topic (brain development) with specific knowledge goal statements to pursue that bear on a sPTB and vitamin D gene list. Further, the findings suggested a field (neuroscience) to look to for answers. These questions (and subsequent answers) have high potential to positively impact the health of pregnant women and their offspring globally. The importance of questions and knowledge goals in research is well established, and our ignorance-base and exploration methods bring these to the forefront to help researchers explore a topic and experimental results in the context of our collective scientific ignorance. The scientific endeavor rests on our continuous ability to ask questions and push research farther as we learn more knowledge. To paraphrase Confucius,"Real knowledge is to know the extent of one's ignorance" (Analects 2:17). In the right context, ignorance is a source of wisdom.

Table 10: Gene list coverage enrichment information for the top 25: NO INFO stands for NO INFORMATION meaning that the ontology term exists in PheKnowLator and is connected to our gene list, but there were no sentences that contained it on the literature side. *Statistically significant with FDR but not family-wise error.

OBO ID	OBO label	Gene Coverage	Enriched in all gene list sentences	Enriched in Ignorance
SO:0001217	protein coding gene	37	YES*	YES
GO:0005515	response to virus	23	NO INFO	NO INFO
CL:0000094	granulocyte	19	YES	NO
GO:0005886	plasma membrane	18	YES	NO
UBERON:0000178	blood	17	YES	NO
UBERON:0002371	bone marrow	15	YES	NO
CL:0000576	monocyte	14	YES	NO
GO:0005576	extracellular region	14	YES	NO
GO:0005615	extracellular space	13	YES	NO
CL:0000775	neutrophil	13	YES	NO
CHEBI:2504	aflatoxin B1	12	YES	NO
CHEBI:39867	kidney	11	NO INFO	NO INFO
GO:0045087	innate immune response	11	YES	YES*
GO:0070062	extracellular exosome	11	YES	NO
GO:0016021	bone marrow	10	NO INFO	NO INFO
SO:0000704	gene	9	YES	YES
GO:0005829	cytosol	9	YES	NO
CL:1001608	foreskin fibroblast	7	NO	NO
UBERON:0001332	prepuce of penis	7	YES*	NO
GO:0005737	cytoplasm	7	YES	NO
GO:0006955	immune response	7	YES	YES
CL:0000765	erythroblast	7	NO	NO
UBERON:0000955	brain	6	YES	YES
GO:0035580	lead(0)	6	NO INFO	NO INFO
CL:0000771	eosinophil	6	YES	NO

Declaration

Acknowledgements

We would like to acknowledge the BioFrontiers Computing Core for computing resources and support, especially Jonathon Demasi. The authors would like to thank Harrison Pielke-Lombardo for his updates to the Knowtator tool; William A. Baumgartner Jr. for his help with the literature and taxonomy work; and Tiffany J. Callahan for many discussions about this work.

Funding

This work was supported by the National Institutes of Health [R01LM013400].

Availability of data and materials

Code for the ignorance-base and exploration methods can be found at: https://github.com/UCDenver-ccp/Ignorance-Base. The expanded ignorance corpus can be found at: https://github.com/UCDenver-ccp/Ignorance-Question-Corpus with all associated code and models at: https://github.com/UCDenver-ccp/Ignorance-Question-Work-Full-Corpus. Code for concept recognition of the OBOs can be found at: https://github.com/UCDenver-ccp/Concept-Recognition-as-Translation. Our previous related work that we built upon for this work can be found at: https://github.com/UCDenver-ccp/Ignorance-Question-Work.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Mayla R. Boguslav: Conceptualization, Methodology, Data Curation, Software, Validation, Writing - Original Draft. **Nourah M. Salem:** Methodology, Software, Validation, Writing - Review Editing. **Elizabeth K.**

White: Data Curation, Validation, Writing - Review Editing. **Katherine J. Sullivan:** Data Curation, Validation, Writing - Review Editing. **Stephanie P. Araki:** Data Curation, Validation. **Michael Bada:** Data Curation, Writing - Review Editing. **Teri L. Hernandez:** Supervision, Writing - Review Editing. **Sonia M. Leach:** Supervision, Writing - Review Editing. **Lawrence E. Hunter:** Supervision, Writing - Review Editing.

References

[1] S. Firestein, Ignorance: How it drives science, OUP, USA, 2012.

[2] T. S. Kuhn, The structure of scientific revolutions, [2d ed., enl Edition, International encyclopedia of unified science. Foundations of the unity of science, v. 2, no. 2, University of Chicago Press, Chicago, 1970.

[3] Z. O'leary, The essential guide to doing research, Sage, Great Britain, 2004.

[4] M. R. Boguslav, N. M. Salem, E. K. White, S. M. Leach, L. E. Hunter, Identifying and classifying goals for scientific knowledge, Bioinformatics Advances 1 (July 2021). doi:10.1093/bioadv/vbab012.

[5] A. Holdcroft, Gender bias in research: how does it affect evidence based medicine? (2007).

[6] N. Slawson, 'women have been woefully neglected': does medical science have a gender problem? (2019).
URL http://www.theguardian.com/education/2019/dec/18/women-have-been-woefully-neglected-does-medical-science-have-a-gender-problem

[7] A. C. Mastroianni, L. M. Henry, D. Robinson, T. Bailey, R. R. Faden, M. O. Little, A. D. Lyerly, Research with pregnant women: new insights on legal decision-making, Hastings Center Report 47 (3) (2017) 38–45.

[8] M. Meadows, Pregnancy and the drug dilemma, FDA Consumer magazine 35 (3) (2001) 16–20.

[9] M. Hanson, , P. Gluckman, Early developmental conditioning of later health and disease: physiology or pathophysiology?, Physiological reviews (2014).

[10] R. Arshad, A. Sameen, M. A. Murtaza, H. R. Sharif, S. Dawood, Z. Ahmed, A. Nemat, M. F. Manzoor, Impact of vitamin d on maternal and fetal health: A review, Food Science & Nutrition (2022).

[11] M. Tous, M. Villalobos, L. Iglesias, S. Fernández-Barrés, V. Arija, Vitamin d status during pregnancy and offspring outcomes: a systematic review and meta-analysis of observational studies, European journal of clinical nutrition 74 (1) (2020) 36–53.

[12] M. Todorova, D. Gerova, B. Galunska, Vitamin d deficiency during pregnancy, Scripta Scientifica Medica 54 (1) (2022) 19–28.

[13] D. Dror, Vitamin d in pregnancy, in: Handbook of vitamin D in human health, Wageningen Academic Publishers, Wageningen, 2013, pp. 670–691.

[14] A. Ş. KIRCA, The effect of vitamin d deficiency in pregnancy on maternal results, Recent Studies in Health Sciences (2019) 359–366.

[15] M. L. Tanaka, A thesis proposal development course for engineering graduate students, Journal of Biomechanical Engineering 142 (11) (2020).

[16] K. W. Boyack, K. Börner, Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers, Journal of the American Society for Information Science and Technology 54 (5) (2003) 447–461.

[17] B. H. Rihm, S. Vidal, C. Nemurat, S. Vachenc, S. Mohr, F. Mazur, P. Houdry, F. Grandjean, S. Visvikis, J. Ducloy, From transcriptomics to bibliomics, Medical Science Monitor 9 (8) (2003) MT89–MT95.

[18] C. Balili, A. Segev, U. Lee, Tracking and predicting the evolution of research topics in scientific literature, in: 2017 IEEE international conference on big data (big data), IEEE, 2017, pp. 1694–1697.

[19] M. Faruqui, D. Das, Identifying well-formed natural language questions, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 798–803. doi:10.18653/v1/D18-1091. URL https://www.aclweb.org/anthology/D18-1091

[20] D. Lahav, J. S. Falcon, B. Kuehl, S. Johnson, S. Parasa, N. Shomron, D. H. Chau, D. Yang, E. Horvitz, D. S. Weld, et al., A search engine for discovery of scientific challenges and directions, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 11982–11990.

[21] O. Bodenreider, The unified medical language system (umls): integrat-

ing biomedical terminology, Nucleic acids research 32 (suppl_1) (2004) D267–D270.

[22] T. J. Callahan, I. J. Tripodi, H. Pielke-Lombardo, L. E. Hunter, Knowledge-based biomedical data science, Annual review of biomedical data science 3 (2020) 23.

[23] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al., The obo foundry: coordinated evolution of ontologies to support biomedical data integration, Nature biotechnology 25 (11) (2007) 1251–1255.

[24] R. Arp, B. Smith, A. D. Spear, Building ontologies with basic formal ontology, Mit Press, 2015.

[25] M. R. Boguslav, N. D. Hailu, M. Bada, W. A. Baumgartner, L. E. Hunter, Concept recognition as a machine translation problem, BMC bioinformatics 22 (1) (2021) 1–39.

[26] T. J. Callahan, PheKnowLator Human Disease Knowledge Graphs – Class-based Knowledge Model with Inverse Relations and OWL-NETS Abstraction (May 2021). doi:10.5281/zenodo.7029922.
URL `https://doi.org/10.5281/zenodo.7029922`

[27] T. J. Callahan, Phenotype Knowledge Translator: A FAIR Ecosystem for Representing Large-Scale Biomedical Knowledge, This is submission serves as a placeholder for a preprint that is being submitted to arXiv. As soon as a valid DOI has been produced, this submission will be updated with the preprint PDF, the DOI, and the submission authors. (Nov. 2021). doi:10.5281/zenodo.5893789.
URL `https://doi.org/10.5281/zenodo.5893789`

[28] C. Balili, U. Lee, A. Segev, J. Kim, M. Ko, Termball: tracking and pre-

dicting evolution types of research topics by using knowledge structures in scholarly big data, IEEE Access 8 (2020) 108514–108529.

[29] W. J. Sutherland, E. Fleishman, M. B. Mascia, J. Pretty, M. A. Rudd, Methods for collaboratively identifying research priorities and emerging issues in science and policy, Methods in Ecology and Evolution 2 (3) (2011) 238–247.

[30] N. Liu, P. Shapira, X. Yue, Tracking developments in artificial intelligence research: Constructing and applying a new search strategy, Scientometrics 126 (4) (2021) 3153–3192.

[31] J. Ostrowsky, M. Arpey, K. Moore, M. Osterholm, M. Friede, J. Gordon, D. Higgins, J. Molto-Lopez, J. Seals, J. Bresee, Tracking progress in universal influenza vaccine development, Current opinion in virology 40 (2020) 28–36.

[32] B. Dinakar, M. R. Boguslav, C. Görg, D. Dinakarpandian, Semantic changepoint detection for finding potentially novel research publications, in: BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, World Scientific, 2020, pp. 107–118.

[33] E. Antonio, M. Alobo, M. T. Bayona, K. Marsh, A. Norton, Funding and covid-19 research priorities-are the research needs for africa being met?[version 1; peer review: awaiting (2020).

[34] J. S. Benner, M. R. Morrison, E. K. Karnes, S. L. Kocot, M. McClellan, An evaluation of recent federal spending on comparative effectiveness research: priorities, gaps, and next steps, Health Affairs 29 (10) (2010) 1768–1776.

[35] S. Wallis, D. C. Cole, O. Gaye, B. T. Mmbaga, V. Mwapasa, H. Tagbor, I. Bates, Qualitative study to develop processes and tools for the

assessment and tracking of african institutions' capacity for operational health research, BMJ open 7 (9) (2017) e016660.

[36] A. P. Yadama, H. Mirzakhani, T. F. McElrath, A. A. Litonjua, S. T. Weiss, Transcriptome analysis of early pregnancy vitamin d status and spontaneous preterm birth, PLoS One 15 (1) (2020) e0227193.

[37] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, et al., The reactome pathway knowledgebase, Nucleic acids research 46 (D1) (2018) D649–D655.

[38] Us department of health and human services, national institutes of health, dietary supplement label database (dsld), office of dietary supplements (2022).
URL https://dsld.nlm.nih.gov/dsld/

[39] D. Sternberg, How to complete and survive a doctoral dissertation, St. Martin's Griffin, New York, 1981.

[40] R. S. Brause, Writing your doctoral dissertation: Invisible rules for success, Routledge, London, 2012.

[41] S. Burton, P. Steane, Surviving your thesis, Routledge, London, 2004.

[42] D. R. Krathwohl, N. L. Smith, How to prepare a dissertation prosposal: Suggestions for students in education & the social and behavioral sciences., Syracuse University Press, Syracuse, NY, 2005.

[43] S. R. Terrell, Writing a proposal for your dissertation: Guidelines and examples, Guilford Publications, New York, 2022.

[44] D. Madsen, Successful dissertations and theses: A guide to graduate student research from proposal to completion. (1983).

[45] S. A. Lei, Strategies for finding and selecting an ideal thesis or dissertation topic: A review of literature, College Student Journal 43 (4) (2009) 1324–1333.

[46] G. Ségol, Choosing a dissertation topic: Additional pointers, College Student Journal 48 (1) (2014) 108–113.

[47] U. A. Eze, O. Adebayo, I. J. Nnodim, A. C. Adejo, L. O. Obazenu, How to choose a dissertation topic, Nigerian Journal of Medicine 30 (2) (2021) 123–124.

[48] G. Lakoff, Hedges: A study in meaning criteria and the logic of fuzzy concepts, in: Contemporary research in philosophical logic and linguistic semantics, Springer, Dordrecht, 1975, pp. 221–271.

[49] K. Hyland, Hedging in scientific research articles, Vol. 54, John Benjamins Publishing, Amsterdam/Philadelphia, 1998.

[50] V. R. Walker, The siren songs of science: toward a taxonomy of scientific uncertainty for decisionmakers, Conn. L. Rev. 23 (1990) 567.

[51] M. Light, X. Y. Qiu, P. Srinivasan, The language of bioscience: Facts, speculations, and statements in between, in: HLT-NAACL 2004 workshop: linking biological literature, ontologies and databases, 2004, pp. 17–24.

[52] H. Kilicoglu, G. Rosemblat, T. C. Rindflesch, Assigning factuality values to semantic relations extracted from biomedical research literature, PloS one 12 (7) (2017) e0179926.

[53] M. Shardlow, R. Batista-Navarro, P. Thompson, R. Nawaz, J. McNaught, S. Ananiadou, Identification of research hypotheses and new knowledge from scientific literature, BMC medical informatics and decision making 18 (1) (2018) 46.

[54] R. Bongelli, I. Riccioni, R. Burro, A. Zuczkowski, Writers' uncertainty in scientific and popular biomedical articles. a comparative analysis of the british medical journal and discover magazine, Plos one 14 (9) (2019) e0221933.

[55] R. Farkas, V. Vincze, G. Móra, J. Csirik, G. Szarvas, The conll-2010 shared task: learning to detect hedges and their scope in natural language text, in: Proceedings of the fourteenth conference on computational natural language learning–Shared task, 2010, pp. 1–12.

[56] V. Ganter, M. Strube, Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009, pp. 173–176.

[57] B. Medlock, T. Briscoe, Weakly supervised learning for hedge classification in scientific literature, in: Proceedings of the 45th annual meeting of the association of computational linguistics, 2007, pp. 992–999.

[58] V. Vincze, G. Szarvas, R. Farkas, G. Móra, J. Csirik, The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes, BMC bioinformatics 9 (11) (2008) 1–9.

[59] H. Kilicoglu, S. Bergler, Recognizing speculative language in biomedical research articles: a linguistically motivated perspective, BMC bioinformatics 9 (S11) (2008) S10.

[60] C. Zerva, R. Batista-Navarro, P. Day, S. Ananiadou, Using uncertainty to link and rank evidence from biomedical literature for model curation, Bioinformatics 33 (23) (2017) 3784–3792.

[61] F. T. Al-Khawaldeh, Hierarchical attention generative adversarial networks for biomedical texts uncertainty detection, Int J Adv Stud Comput Sci Eng 8 (6) (2019) 1–12.

[62] G. Szarvas, V. Vincze, R. Farkas, G. Móra, I. Gurevych, Cross-genre and cross-domain detection of semantic uncertainty, Computational Linguistics 38 (2) (2012) 335–367.

[63] E. Velldal, Detecting uncertainty in biomedical literature: A simple disambiguation approach using sparse random indexing., in: Semantic Mining in Biomedicine, 2010.

[64] K. Fujikawa, K. Seki, K. Uehara, A hybrid approach to finding negated and uncertain expressions in biomedical documents, in: Proceedings of the 2nd international workshop on Managing interoperability and compleXity in health systems, 2012, pp. 67–74.

[65] Y. Ren, H. Fei, Q. Peng, Detecting the scope of negation and speculation in biomedical texts by using recursive neural network, in: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE, 2018, pp. 739–742.

[66] N. Konstantinova, S. C. De Sousa, Annotating negation and speculation: the case of the review domain, in: Proceedings of the Second Student Research Workshop associated with RANLP 2011, 2011, pp. 139–144.

[67] R. R. Zavala, P. Martinez, The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: Comparative study, JMIR Medical Informatics 8 (12) (2020) e18953.

[68] E. Apostolova, N. Tomuro, D. Demner-Fushman, Automatic extraction of lexico-syntactic patterns for detection of negation and speculation scopes, in: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 283–287.

[69] Z. Qian, P. Li, Q. Zhu, G. Zhou, Z. Luo, W. Luo, Speculation and negation scope detection via convolutional neural networks, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 815–825.

[70] H. Fei, Y. Ren, D. Ji, Negation and speculation scope detection using recursive neural conditional random fields, Neurocomputing 374 (2020) 22–29.

[71] G. Szarvas, Hedge classification in biomedical texts with a weakly supervised selection of keywords, in: Proceedings of acl-08: HLT, 2008, pp. 281–289.

[72] F. T. AL-Khawaldeh, Speculation and negation annotation for arabic biomedical texts: Bioarabic corpus, World of Computer Science & Information Technology Journal (2016).

[73] A. Khandelwal, B. K. Britto, Multitask learning of negation and speculation using transformers, in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, 2020, pp. 79–87.

[74] F. T. Al-Khawaldeh, Speculation and negation detection for arabic biomedical texts, World of Computer Science & Information Technology Journal 9 (3) (2019).

[75] K. Cheng, T. Baldwin, K. Verspoor, Automatic negation and speculation detection in veterinary clinical text, in: Proceedings of the Australasian Language Technology Association Workshop 2017, 2017, pp. 70–78.

[76] E. Velldal, Predicting speculation: a simple disambiguation approach to hedge detection in biomedical literature, Journal of Biomedical Semantics 2 (5) (2011) 1–14.

[77] S. Agarwal, H. Yu, Detecting hedge cues and their scope in biomedical text with conditional random fields, Journal of biomedical informatics 43 (6) (2010) 953–961.

[78] H. Zhou, X. Li, D. Huang, Z. Li, Y. Yang, Exploiting multi-features to detect hedges and their scope in biomedical texts, in: Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task, 2010, pp. 106–113.

[79] H. Zhou, H. Deng, D. Huang, M. Zhu, Hedge scope detection in biomedical texts: an effective dependency-based method, PloS one 10 (7) (2015) e0133715.

[80] B. Medlock, Exploring hedge identification in biomedical literature, Journal of biomedical informatics 41 (4) (2008) 636–654.

[81] F. Ji, X. Qiu, X.-J. Huang, Detecting hedge cues and their scopes with average perceptron, in: Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task, 2010, pp. 32–39.

[82] D. A. Hanauer, Y. Liu, Q. Mei, F. J. Manion, U. J. Balis, K. Zheng, Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients, in: AMIA Annual Symposium Proceedings, Vol. 2012, American Medical Informatics Association, 2012, p. 321.

[83] Q. Zhao, C.-J. Sun, B. Liu, Y. Cheng, Learning to detect hedges and their scope using crf, in: Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task, 2010, pp. 100–105.

[84] D. Clausen, Hedgehunter: A system for hedge detection and uncertainty classification, in: Proceedings of the Fourteenth Conference on

Computational Natural Language Learning–Shared Task, 2010, pp. 120–125.

[85] R. Morante, W. Daelemans, Learning the scope of hedge cues in biomedical texts, in: Proceedings of the BioNLP 2009 workshop, 2009, pp. 28–36.

[86] F. Liu, P. Zhou, S. J. Baccei, M. J. Masciocchi, N. Amornsiripanitch, C. I. Kiefe, M. P. Rosen, Qualifying certainty in radiology reports through deep learning–based natural language processing, American Journal of Neuroradiology 42 (10) (2021) 1755–1761.

[87] M. Verbeke, P. Frasconi, V. Van Asch, R. Morante, W. Daelemans, L. De Raedt, Kernel-based logical and relational learning with klog for hedge cue detection, in: International Conference on Inductive Logic Programming, Springer, 2011, pp. 347–357.

[88] D. L. Mowery, S. Velupillai, W. Chapman, Medical diagnosis lost in translation–analysis of uncertainty and negation expressions in english and swedish clinical texts, in: BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, 2012, pp. 56–64.

[89] P. K. Han, W. M. Klein, N. K. Arora, Varieties of uncertainty in health care: a conceptual taxonomy, Medical Decision Making 31 (6) (2011) 828–838.

[90] J. Pearl, D. Mackenzie, The book of why: the new science of cause and effect, Basic Books, 2018.

[91] H. M. Regan, M. Colyvan, M. A. Burgman, A taxonomy and treatment of uncertainty for ecology and conservation biology, Ecological applications 12 (2) (2002) 618–628.

[92] M. Smithson, Ignorance and uncertainty: Emerging paradigms, Springer Science & Business Media, 2012.

[93] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, et al., The ontology for biomedical investigations, PloS one 11 (4) (2016) e0154556.

[94] F. B. Bastian, M. C. Chibucos, P. Gaudet, M. Giglio, G. L. Holliday, H. Huang, S. E. Lewis, A. Niknejad, S. Orchard, S. Poux, et al., The confidence information ontology: a step towards a standard for asserting confidence in annotations, Database 2015 (2015).

[95] M. H. Brush, K. Shefchek, M. Haendel, Sepio: A semantic model for the integration and analysis of scientific evidence., in: ICBO/BioCreative, 2016.

[96] M. C. Chibucos, C. J. Mungall, R. Balakrishnan, K. R. Christie, R. P. Huntley, O. White, J. A. Blake, S. E. Lewis, M. Giglio, Standardized description of scientific evidence using the evidence ontology (eco), Database 2014 (2014).

[97] T. C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, Journal of biomedical informatics 36 (6) (2003) 462–477.

[98] P. Thompson, R. Nawaz, J. McNaught, S. Ananiadou, Enriching a biomedical event corpus with meta-knowledge annotation, BMC bioinformatics 12 (1) (2011) 1–18.

[99] R. Bongelli, C. Canestrari, I. Riccioni, A. Zuczkowski, C. Buldorini, R. Pietrobon, A. Lavelli, B. Magnini, A corpus of scientific biomedical texts spanning over 168 years annotated for uncertainty., in: LREC, Vol. 12, 2012, pp. 2009–2014.

[100] N. P. C. Díaz, Detecting negated and uncertain information in biomedical and review texts, in: Proceedings of the Student Research Workshop associated with RANLP 2013, 2013, pp. 45–50.

[101] N. Konstantinova, S. C. De Sousa, N. P. C. Díaz, M. J. M. López, M. Taboada, R. Mitkov, A review corpus annotated for negation, speculation and their scope., in: Lrec, 2012, pp. 3190–3195.

[102] H. Zhou, H. Yang, J. Zhang, S. Kang, D. Huang, The research and construction of chinese hedge corpus, Journal of Chinese Information Processing 29 (6) (2015) 83–89.

[103] L. M. Sanchez, C. Vogel, A hedging annotation scheme focused on epistemic phrases for informal language, in: Proceedings of the Workshop on Models for Modality Annotation, 2015.

[104] S. M. Jiménez-Zafra, M. Taulé, M. T. Martín-Valdivia, L. A. Urena-López, M. A. Martí, Sfu review sp-neg: a spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns, Language Resources and Evaluation 52 (2) (2018) 533–569.

[105] H. Yang, A. De Roeck, V. Gervasi, A. Willis, B. Nuseibeh, Speculative requirements: Automatic detection of uncertainty in natural language requirements, in: 2012 20th IEEE International Requirements Engineering Conference (RE), IEEE, 2012, pp. 11–20.

[106] P.-A. Jean, S. Harispe, S. Ranwez, P. Bellot, J. Montmain, Uncertainty detection in natural language: A probabilistic model, in: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, 2016, pp. 1–10.

[107] E. Sergeeva, H. Zhu, A. Tahmasebi, P. Szolovits, Neural token representations and negation and speculation scope detection in biomedical and general domain text, in: Proceedings of the Tenth International

Workshop on Health Text Mining and Information Analysis (LOUHI 2019), 2019, pp. 178–187.

[108] H. Zhou, D. Huang, X. Li, Y. YANG, Combining structured and flat features by a composite kernel to detect hedges scope in biological texts, Chinese Journal of Electronics 20 (3) (2011) 476–482.

[109] H. Zhou, X. Li, D. Huang, Y. Yang, F. Ren, Voting-based ensemble classifiers to detect hedges and their scopes in biomedical texts, IEICE TRANSACTIONS on Information and Systems 94 (10) (2011) 1989–1997.

[110] H. Zhou, J. Xu, Y. Yang, H. Deng, L. Chen, D. Huang, Chinese hedge scope detection based on structure and semantic information, in: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, Springer, 2016, pp. 204–215.

[111] M. Georgescul, A hedgehop over a max-margin framework using hedge cues, in: Proceedings of the 14th International Conference on Computational Natural Language Learning: Shared Task, 2010, pp. 26–31.

[112] G. Moncecchi, Recognizing speculative language in research texts, Ph.D. thesis, Université de Nanterre-Paris X; Universidad de la República-Proyecto de ... (2013).

[113] N. P. C. Díaz, Detección de la negación y la especulación en textos médicos y de, Ph.D. thesis, Citeseer (2014).

[114] M. Turner, J. Ive, S. Velupillai, Linguistic uncertainty in clinical nlp: A taxonomy, dataset and approach, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2021, pp. 129–141.

[115] C. Dalloux, V. Claveau, N. Grabar, Speculation and negation detection in french biomedical corpora, in: RANLP 2019-Recent Advances in Natural Language Processing, 2019, pp. 1–10.

[116] B. Zou, Q. Zhu, G. Zhou, Negation and speculation identification in chinese language, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 656–665.

[117] S. Zhang, T. Kang, X. Zhang, D. Wen, N. Elhadad, J. Lei, Speculation detection for chinese clinical notes: Impacts of word segmentation and embedding models, Journal of biomedical informatics 60 (2016) 334–341.

[118] J. Islam, L. Xiao, R. E. Mercer, A lexicon-based approach for detecting hedges in informal text, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 3109–3113.

[119] B. M. Konopka, Biomedical ontologies—a review, Biocybernetics and Biomedical Engineering 35 (2) (2015) 75–86.

[120] G. O. Consortium, The gene ontology resource: 20 years and still going strong, Nucleic acids research 47 (D1) (2019) D330–D338.

[121] S. Leonelli, Classificatory theory in data-intensive science: The case of open biomedical ontologies, International Studies in the Philosophy of Science 26 (1) (2012) 47–65.

[122] C. J. Mungall, M. Bada, T. Z. Berardini, J. Deegan, A. Ireland, M. A. Harris, D. P. Hill, J. Lomax, Cross-product extensions of the gene ontology, Journal of biomedical informatics 44 (1) (2011) 80–86.

[123] G. O. Consortium, Gene ontology annotations and resources, Nucleic acids research 41 (D1) (2012) D530–D535.

[124] D. L. Rubin, S. E. Lewis, C. J. Mungall, S. Misra, M. Westerfield, M. Ashburner, I. Sim, C. G. Chute, M.-A. Storey, B. Smith, et al., National center for biomedical ontology: advancing biomedicine through structured organization of scientific knowledge, Omics: a journal of integrative biology 10 (2) (2006) 185–198.

[125] H. Tipney, L. Hunter, An introduction to effective use of enrichment analysis software, Human genomics 4 (3) (2010) 1–5.

[126] L. Shi Jing, F. Fathiah Muzaffar Shah, M. Saberi Mohamad, K. Moorthy, S. Deris, Z. Zakaria, S. Napis, A review on bioinformatics enrichment analysis tools towards functional analysis of high throughput gene set data, Current Proteomics 12 (1) (2015) 14–27.

[127] J.-H. Hung, T.-H. Yang, Z. Hu, Z. Weng, C. DeLisi, Gene set enrichment analysis: performance evaluation and usage guidelines, Briefings in bioinformatics 13 (3) (2012) 281–291.

[128] R. K. Curtis, M. Orešič, A. Vidal-Puig, Pathways to the analysis of microarray data, TRENDS in Biotechnology 23 (8) (2005) 429–435.

[129] K. Wijesooriya, S. A. Jadaan, K. L. Perera, T. Kaur, M. Ziemann, Urgent need for consistent standards in functional enrichment analysis, PLoS computational biology 18 (3) (2022) e1009935.

[130] Q. Gan, M. Zhu, M. Li, T. Liang, Y. Cao, B. Zhou, Document visualization: an overview of current research, Wiley Interdisciplinary Reviews: Computational Statistics 6 (1) (2014) 19–36.

[131] S. VIDAL, J. DUCLOY, P. HOUDRY, Mining medical data using multiple corpora interaction: the transcriptomics investigation server experiment, Systemics Cybernetics and Informatics (Actes du congrès à Houston, USA) 7.

[132] J. Chen, H. Xu, B. J. Aronow, A. G. Jegga, Improved human disease candidate gene prioritization using mouse phenotype, BMC bioinformatics 8 (1) (2007) 1–13.

[133] J. Jourquin, D. Duncan, Z. Shi, B. Zhang, Glad4u: deriving and prioritizing gene lists from pubmed literature, BMC genomics 13 (8) (2012) 1–12.

[134] T.-K. Jenssen, A. Lægreid, J. Komorowski, E. Hovig, A literature network of human genes for high-throughput analysis of gene expression, Nature genetics 28 (1) (2001) 21–28.

[135] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, P. Stoehr, Ebimed—text crunching to gather facts for proteins from medline, Bioinformatics 23 (2) (2007) e237–e244.

[136] J.-F. Fontaine, F. Priller, A. Barbosa-Silva, M. A. Andrade-Navarro, Genie: literature-based gene prioritization at multi genomic scale, Nucleic acids research 39 (suppl_2) (2011) W455–W461.

[137] D. Börnigen, L.-C. Tranchevent, F. Bonachela-Capdevila, K. Devriendt, B. De Moor, P. De Causmaecker, Y. Moreau, An unbiased evaluation of gene prioritization tools, Bioinformatics 28 (23) (2012) 3081–3088.

[138] G. Grimes, T. Wen, M. Mewissen, R. M. Baxter, S. Moodie, J. Beattie, P. Ghazal, Pdq wizard: automated prioritization and characterization of gene and protein lists using biomedical literature, Bioinformatics 22 (16) (2006) 2055–2057.

[139] S. J. Modlin, D. Gunasekaran, A. M. Zlotnicki, A. Elghraoui, N. Kuo, C. K. Chan, F. Valafar, Resolving the hypotheticome: annotating m. tuberculosis gene function through bibliomic reconciliation and structural modeling, Preprint at https://doi. org/10.1101/358986 (2018).

[140] L. Yang, B. Wang, G. Xia, Z. Xia, L. Xu, Bibliomics-based selection of analgesics targets through google-pagerank-like algorithm, in: 2007 Second International Conference on Bio-Inspired Computing: Theories and Applications, IEEE, 2007, pp. 98–101.

[141] K. M. Hettne, M. Weeber, M. L. Laine, H. t. Cate, S. Boyer, J. A. Kors, B. G. Loos, Automatic mining of the literature to generate new hypotheses for the possible link between periodontitis and atherosclerosis: lipopolysaccharide as a case study, Journal of clinical periodontology 34 (12) (2007) 1016–1024.

[142] M. Miwa, T. Ohta, R. Rak, A. Rowley, D. B. Kell, S. Pyysalo, S. Ananiadou, A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text, Bioinformatics 29 (13) (2013) i44–i52.

[143] C. Zerva, Automatic identification of textual uncertainty, The University of Manchester, United Kingdom, 2019.

[144] B. T. Sherman, D. W. Huang, Q. Tan, Y. Guo, S. Bour, D. Liu, R. Stephens, M. W. Baseler, H. C. Lane, R. A. Lempicki, David knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis, BMC bioinformatics 8 (1) (2007) 1–11.

[145] B. T. Sherman, M. Hao, J. Qiu, X. Jiao, M. W. Baseler, H. C. Lane, T. Imamichi, W. Chang, David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update), Nucleic Acids Res 10 (2022).

[146] Open access subset (2018).
URL https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[147] H. Pielke-Lombardo, Knowtator-2.0: A text annotation plugin for protege 5+ (2018).
URL https://github.com/UCDenver-ccp/Knowtator-2.0

[148] H. Knublauch, R. W. Fergerson, N. F. Noy, M. A. Musen, The protégé owl plugin: An open development environment for semantic web applications, in: International semantic web conference, Springer, 2004, pp. 229–243.

[149] Fiji user guide.
URL https://bit.colorado.edu/biofrontiers-computing/fiji/fiji-user-guide/

[150] G. Hripcsak, A. S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, Journal of the American medical informatics association 12 (3) (2005) 296–298.

[151] H. Dalianis, Evaluation metrics and evaluation, in: Clinical text mining, Springer, Cham, Switzerland, 2018, pp. 45–53.

[152] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).

[153] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[154] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[155] M. Korobov, sklearn-crfsuite.
URL https://sklearn-crfsuite.readthedocs.io/en/latest/index.html

[156] N. Collier, H. S. Park, N. Ogata, Y. Tateisi, C. Nobata, T. Ohta, T. Sekimizu, H. Imai, K. Ibushi, J. Tsujii, The genia project: corpus-based knowledge acquisition and information extraction from genome research papers, in: Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999, pp. 271–272.

[157] K. B. Cohen, K. Verspoor, K. Fort, C. Funk, M. Bada, M. Palmer, L. E. Hunter, The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain, in: Handbook of Linguistic Annotation, Springer, Dordrecht, 2017, pp. 1379–1394.

[158] W. A. Baumgartner Jr, M. Bada, S. Pyysalo, M. R. Ciosici, N. Hailu, H. Pielke-Lombardo, M. Regan, L. Hunter, Craft shared tasks 2019 overview—integrated structure, semantics, and coreference, in: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 174–184.

[159] S. Lee, D. K. Lee, What is the proper way to apply the multiple comparison test?, Korean journal of anesthesiology 71 (5) (2018) 353–360.

[160] J. Pustejovsky, A. Stubbs, Natural Language Annotation for Machine Learning: A guide to corpus-building for applications, O'Reilly Media, Inc., Sebastopol, CA, 2012.

[161] P. Costanzo, A. Santini, L. Fattore, E. Novellino, A. Ritieni, Toxicity of aflatoxin b1 towards the vitamin d receptor (vdr), Food and Chemical Toxicology 76 (2015) 77–79.

[162] D. Sanchez-Hernandez, G. H. Anderson, A. N. Poon, E. Pannia, C. E. Cho, P. S. Huot, R. Kubant, Maternal fat-soluble vitamins, brain development, and regulation of feeding behavior: an overview of research, Nutrition Research 36 (10) (2016) 1045–1054.

[163] N. C. Harvey, C. Holroyd, G. Ntani, K. Javaid, P. Cooper, R. Moon, Z. Cole, T. Tinati, K. Godfrey, E. Dennison, et al., Vitamin d supplementation in pregnancy: a systematic review., Health technology assessment (Winchester, England) 18 (45) (2014) 1–190.

[164] G. Saggese, F. Vierucci, F. Prodam, F. Cardinale, I. Cetin, E. Chiappini, G. L. de'Angelis, M. Massari, E. Miraglia Del Giudice, M. Miraglia Del Giudice, et al., Vitamin d in pediatric age: consensus of the italian pediatric society and the italian society of preventive and social pediatrics, jointly with the italian federation of pediatricians, Italian journal of pediatrics 44 (2018) 1–40.

[165] R. Saraf, S. M. Morton, C. A. Camargo Jr, C. C. Grant, Global summary of maternal and newborn vitamin d status–a systematic review, Maternal & child nutrition 12 (4) (2016) 647–668.

[166] O. Ojo, S. M. Weldon, T. Thompson, E. J. Vargo, The effect of vitamin d supplementation on glycaemic control in women with gestational diabetes mellitus: a systematic review and meta-analysis of randomised controlled trials, International journal of environmental research and public health 16 (10) (2019) 1716.

[167] S. Hanieh, T. T. Ha, J. A. Simpson, T. T. Thuy, N. C. Khuong, D. D. Thoang, T. D. Tran, T. Tuan, J. Fisher, B.-A. Biggs, Maternal vitamin d status and infant outcomes in rural vietnam: a prospective cohort study, PloS one 9 (6) (2014) e99005.

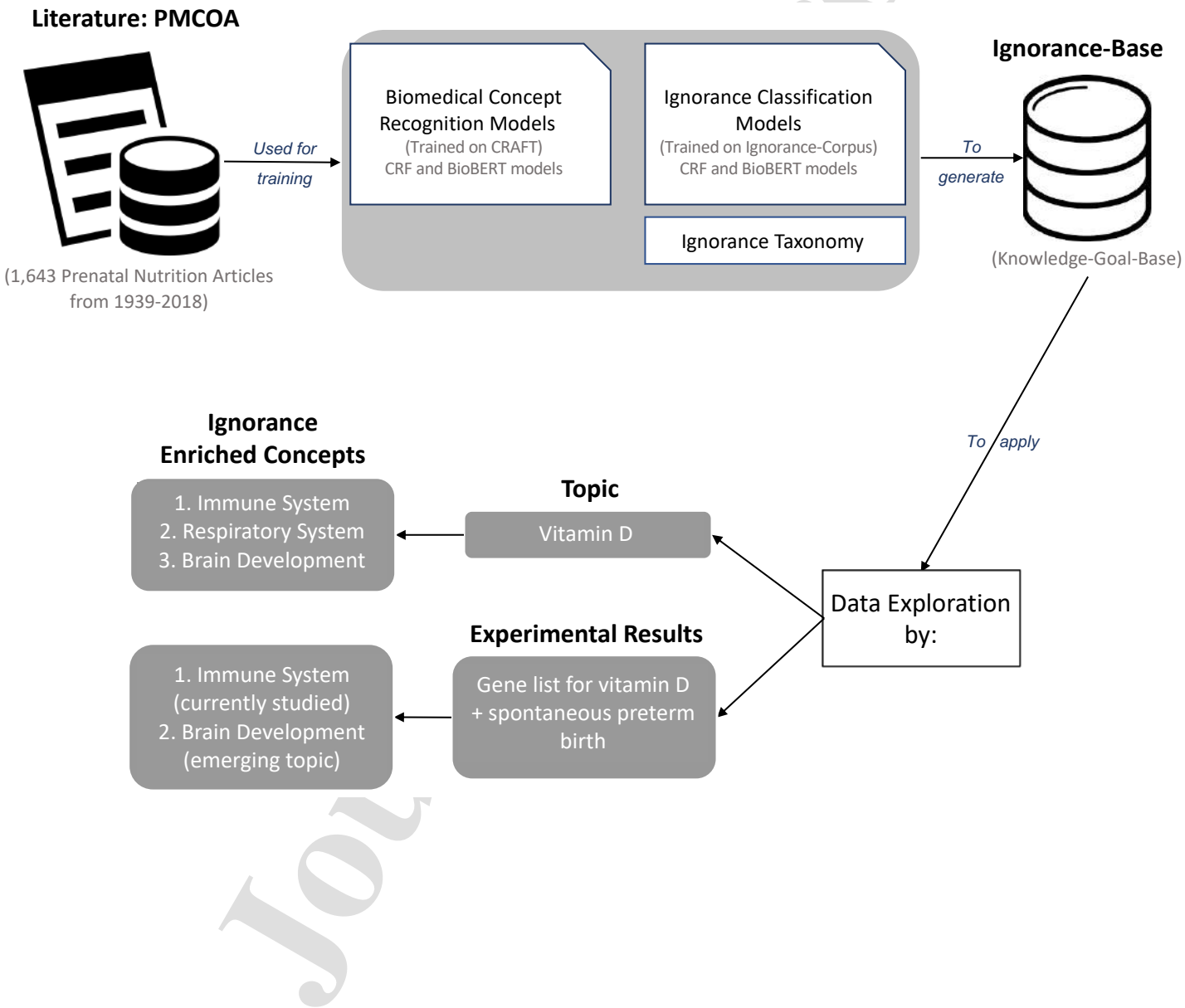
[168] A. Angelidou, S. Asadi, K.-D. Alysandratos, A. Karagkouni, S. Kourembanas, T. C. Theoharides, Perinatal stress, brain inflammation and risk of autism-review and proposal, BMC pediatrics 12 (1) (2012) 1–12.

[169] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for

biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (1) (2021) 1–23.

[170] C.-H. Wei, A. Allot, R. Leaman, Z. Lu, Pubtator central: automated concept annotation for biomedical full text articles, Nucleic acids research 47 (W1) (2019) W587–W593.

[171] N. L. of Medicine, Pubmed user guide: Publication types (2021). URL https://pubmed.ncbi.nlm.nih.gov/help/publication-types



Authors' contributions

Mayla R. Boguslav: Conceptualization, Methodology, Data Curation, Software, Validation, Writing - Original Draft.
Nourah M. Salem: Methodology, Software, Validation, Writing - Review & Editing.
Elizabeth K. White: Data Curation, Validation, Writing - Review & Editing.
Katherine J. Sullivan: Data Curation, Validation, Writing - Review & Editing.
Stephanie P. Araki: Data Curation, Validation.
Michael Bada: Data Curation, Writing - Review & Editing.
Teri L. Hernandez: Supervision, Writing - Review & Editing.
Sonia M. Leach: Supervision, Writing - Review & Editing.
Lawrence E. Hunter: Supervision, Writing - Review & Editing.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: