RESEARCH

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

Mayla R. Boguslav[1*], Nourah M. Salem[1], Elizabeth K. White[2], Katherine J. Sullivan[1], Stephanie P. Araki[1], Michael Bada[1], Teri L. Hernandez[3], Sonia M. Leach[1,2] and Lawrence E. Hunter[1]

[*]Correspondence: mayla.boguslav@cuanschutz.edu
[1]Computational Bioscience Program University of Colorado Anschutz Medical Campus, E 17th Avenue, 80045 Aurora, CO, USA
Full list of author information is available at the end of the article

Abstract

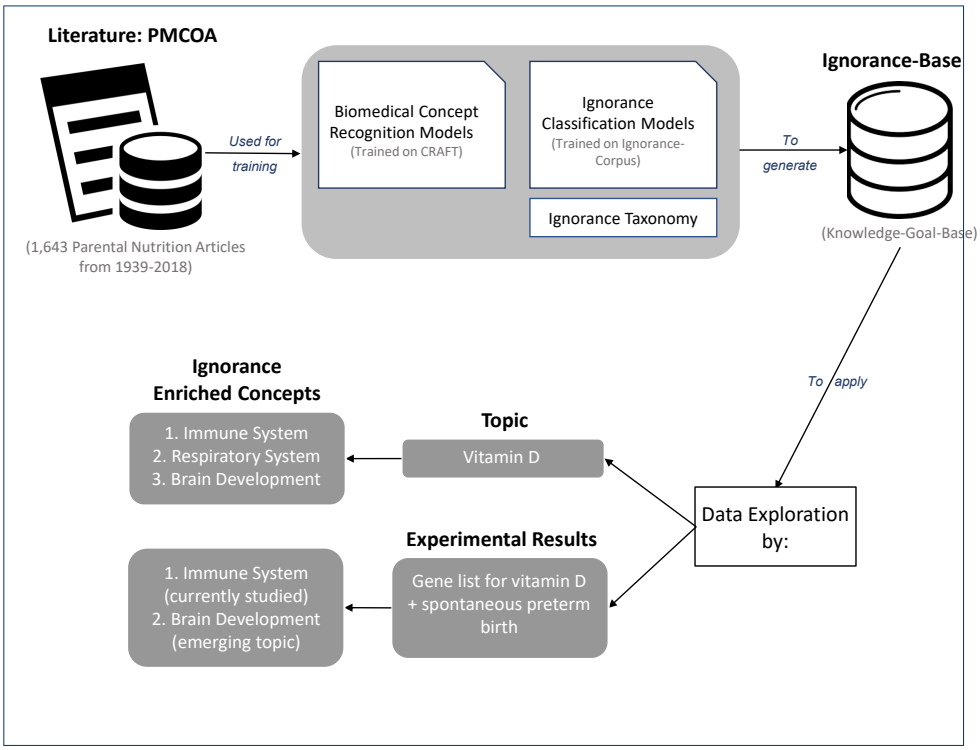
Background: Research progresses through the process of making the unknown unknowns into known unknowns, then into knowns. Many knowledge-bases exist to capture the knowns and can be used by graduate students to explore a topic or help researchers contextualize experimental results using concept enrichment methods. The unknowns surrounding these applications are equally as important as the knowns to continue to find the most pertinent questions and their answers. Prior work on known unknowns has sought to understand the phenomenon, annotate it, and automate its identification. However, no knowledge-bases exist to capture them and facilitate the widespread search for questions in order to find new avenues for exploration from a given topic or experimental results. The prenatal nutrition field especially could benefit from this due to the ethical and legal considerations in studying this underserved population.

Results: We present the first ignorance-base, a knowledge-base created from combining ignorance and biomedical concept classifiers over the prenatal nutrition literature to annotate the known unknowns or ignorance statements. Motivated by a graduate student's search for a thesis topic in vitamin D, we found three new interesting avenues for exploration (immune system, respiratory system, and brain development) that were buried among the many standard enriched concepts but were found by focusing on ignorance enriched concepts. We also used the ignorance-base to connect a vitamin D and spontaneous preterm birth gene list to ignorance statements and found an emerging topic of study (brain development) in relation to the gene list. This emerging topic also has an implied field (neuroscience) that could potentially help provide some 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 translational 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

Graphical abstract



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 emerge (known knowns) [1, 2, 3]. This work focuses on the known unknowns to help illuminate their entailed goals for scientific knowledge (i.e. actionable next step). Many **knowledge-bases** exist to capture the known knowns from domain experts, the scientific literature, and other data sources such as experimental results [4]. These Knowledge-bases are important based on their variety of applications [4], 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. For example, a graduate student interested in learning about the field of prenatal nutrition might consult a database of dietary supplements [5]. Or a researcher 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 [6].

Both of these examples explore the knowledge-bases of known facts, but equally important are the questions surrounding a topic or a set of experimental results. The goal of graduate school is to identify and provide at least some solutions for a question that is unanswered. But how do graduate students or scientists find the most pertinent question? There are many books [7, 8, 9, 10, 11] and articles [12, 13, 14, 15, 16] discussing the process of thesis development, including how to choose a topic, but to our knowledge there are no automated systems other than

an automated scheduling system for thesis and project presentations [17]. Similar to the representation of known knowns in knowledge-bases, efforts exist to capture the known unknowns through understanding the phenomenon [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54], creating taxonomies where a hierarchy of terms is linked by specified relationships [55, 29, 56, 25, 57, 58, 59, 60] and ontologies specifying relationships among controlled vocabularies [61, 62, 63, 64], annotating literature to create corpora [65, 66, 23, 21, 67, 68, 33, 69, 39, 49, 70, 71, 72], and automating identification of unknowns through classification tasks [73, 74, 25, 27, 28, 30, 31, 32, 34, 35, 36, 37, 38, 40, 43, 75, 44, 45, 46, 76, 48, 77, 50, 51, 78, 79, 52, 54, 24, 80, 81]. 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 [82, 83, 29, 84, 85, 38, 42, 86, 47, 22, 53, 87, 26]. However, none of these works have used these efforts to create a knowledge-base to capture the known unknowns formally. As such, no existing system can support exploring the questions surrounding a topic or set of experimental results.

Methods for standard functional enrichment analyses (contextualizing experimental results) use knowledge-bases and ontologies [88, 89, 90, 91, 92], and natural language processing (NLP) tools over the biomedical literature [93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103]. Some of this prior work not only aimed to characterize genes but also to help define new research areas (e.g., [93] as one of a few goals), generate new hypotheses (e.g., [104]), find information about genes of unknown function and fill gaps in knowledge (e.g., a preprint [102] employs 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. [105]) and some of these take uncertainty into account (e.g. [27, 106]). However, these works focus on confidence and relevance, respectively, rather than explicitly representing statements of known unknowns. Thus, to the best of our knowledge, there is no prior work that can facilitate the widespread search for questions to find new avenues for exploration from a given topic or set of experimental results.

Thus, the question is whether we can create and use a knowledge-base containing representations of known unknowns, formally defined as an **ignorance-base** (based on [1, 82]), to **explore by topic** and **explore by experimental results**, as done in the knowledge-bases. Identifying questions that need answers, allows us to then look to other fields and knowledge-bases to help answer them. 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 translational research through the continued illumination of and focus on the known unknowns and their respective goals for scientific knowledge.

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, prenatal nutrition is understudied and serves to benefit from the identification of questions that are well studied in other fields [107, 108, 109, 110]. Fetal development is a critical period and exposure to nutrition has a lifelong impact [111]. For example, the micronutrient vitamin D is very important for maternal and fetal health, affecting the immune and musculoskeletal systems, neurodevelopment, and hormones

[112, 113, 114, 115, 116] (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 [112]. If we can identify the known unknowns in prenatal nutrition, even just with regard to the role of vitamin D, then we can look to other fields to help us answer the questions raised, and thereby avoid the additional risks faced when studying pregnant women and their children. Thus, this work has the potential to facilitate new interdisciplinary interactions that could advance the study of this underserved population and potentially help accelerate translational research for mothers everywhere.

We present the ignorance-base as a formal representation of **statements of ignorance** using the example of the prenatal nutrition literature. We define statements of ignorance as statements of incomplete or missing knowledge categorized based on the entailed **knowledge goals** (i.e., the next actionable step based on the given unknown) [82]. We chose this formalization of known unknowns because of the knowledge goal focus in order to provide actionable next steps to the students, researchers, funders, or publishers. For example, "these <u>inconsistent</u> <u>observations</u> <u>point</u> to the <u>complicated</u> <u>role</u> of VITAMIN D in the IMMUNE modulation and disease process" (PMC4889866) is a statement of ignorance with the entailed knowledge goal 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. The ignorance statement and entailed knowledge goal are 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 <u>inconsistent</u> is an *indication of alternative research options or controversy of research*, the ignorance category, and <u>complicated</u> is an *indication of difficult research task*. We identify here and in prior work that most ignorance statements are implicit [82]. In order to also understand the biomedical subjects of these statements and be able to connect them to other knowledge-bases, we identified the biomedical concepts from the open biomedical ontologies (OBOs) (e.g. the all caps words in the example sentence) using concept recognition methods [117]. The ignorance-base combines the identification and classification of ignorance statements and biomedical concept recognition.

With the ignorance-base, we show it is possible to automatically find statements of ignorance related to a topic motivated by a graduate student's search for a thesis topic in vitamin D. Analogous to exploring by topic using knowledge-bases, we used the ignorance-base to find other keywords and knowledge goals for the graduate student by determining if other concepts were enriched in vitamin D ignorance statements compared to all statements (**ignorance enrichment**). We compared the standard approach of concept enrichment in a set of documents to ignorance enrichment and found that the concepts immune system, brain development, and respiratory system were enriched in vitamin D ignorance statements. In a similar vein, in order to help illuminate the types of knowledge goals to pursue and map out how they change over time, we identified ignorance categories that were over-represented in a subset of ignorance statements as compared to all ignorance

statements (**ignorance-category enrichment**). We argue how this helped narrow the graduate student's search to select specific ignorance categories to provide a filtered set of knowledge goals to explore as potential thesis topics. For example, the student could choose a topic that is a complete unknown or a topic where there are alternate existing hypotheses to explore. Further, we argue how this can help track emerging research areas or perform a longitudinal analysis of the evolution of research questions over time for funding agencies and publishers [118, 119, 120, 121, 122, 123, 124, 125, 126, 127].

We also show that it is possible to contextualize experimental results in terms of statements of ignorance to understand what questions may bear on them, motivated by a gene list connecting vitamin D and spontaneous preterm birth (sPTB) from the literature [6]. If vitamin D plays a role in preventing sPTB, it would affect mothers everywhere. In comparing our ignorance-base approach to the standard methods for the gene list of vitamin D and sPTB, we found enrichment of the concept immune system, as found by the original authors, as well as a novel putative relationship with the concept brain development.

The purpose of this work is to show initial feasibility of the ignorance-base and exploring it by topic and experimental results for expanded future work. (A list of the formal terms we have introduced here and their definitions are shown in Table 1.)

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	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

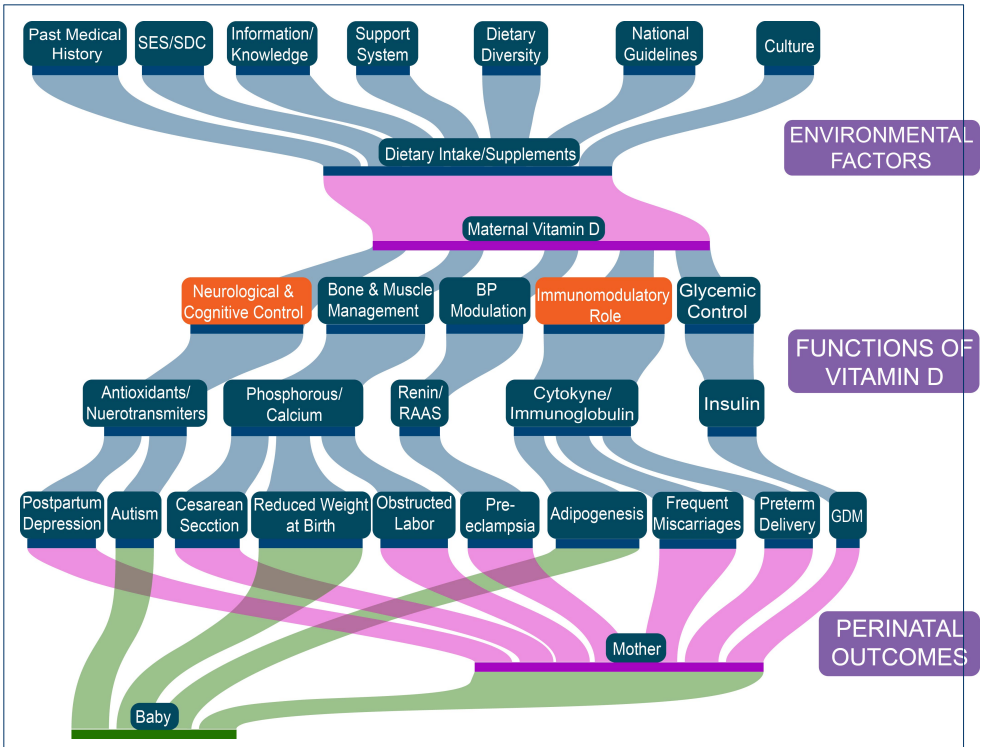


Figure 1: Relationship between society, maternal nutrition (vitamin D), and the effects on mother and offspring: a sankey diagram created based on Figure 3 from [112]. The orange color represents the findings from the exploration methods that the concepts related to brain development and immune system 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-bases exist to capture known unknowns from the literature to contextualize topics and experimental results in our collective scientific ignorance.	Many knowledge-bases exist to capture what is known. They are used to find information and contextualize experimental results. Known unknowns are recognized as important and studied under different names and focuses, including a focus on knowledge goals.	This study aims to build an ignorance-base and explore it by topic to help graduate students find thesis topics, and by experimental results to contextualize them in the known unknowns. We found avenues for research that are emerging and currently being studied.

Methods

We combine the best-performing ignorance classifiers (extending the work of [82] to create a corpus of 91 articles) with state-of-the-art biomedical concept classifiers [117] to create an ignorance-base that allows us to explore it by a topic and by experimental results. The ignorance-base can be queried by ignorance category, specific lexical cues, ontology concepts, or any combination of the three.

The rest of this section will be organized into the following subsections:

1 Creating the ignorance-base

2 Exploration by Topic
3 Exploration by Experimental Results

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 [128], allowing us more data and the ability to share it publicly. 1,643 prenatal nutrition articles (1939-2018) were gathered from querying PMCOA for 54 regular expressions determined in consultation with a prenatal nutrition expert, Teri L. Hernandez. All articles were provided in XML format and converted to text format parsing through the XML using a script in Java. All subsequent computation was implemented in Python 3, with its associated packages. The continued annotation effort used Knowtator [129] and Protege [130] as in previous work [82], 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 [117, 82].

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_NetworkxMultiDiGraph.gpickle), which semantically integrates eleven OBOs [131, 132]. A gene list (our motivating example of experimental results) was gathered from a PMCOA article (PMC6988958) [6]. We also used DAVID (a tool for functional annotation and enrichment analyses of gene lists) [133] as a standard approach for functional enrichment analysis to compare to ignorance enrichment.

Computation used a contemporary laptop and an NIH-funded shared supercomputing resource [134] 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 OBOs.

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.

Creating the ignorance-base

To create the ignorance-base, we expanded our corpus of ignorance statements to train and evaluate high quality ignorance classifiers and combined them with biomedical concept classifiers.

Expanding the ignorance corpus

To create high quality ignorance classifiers, we needed enough data to evaluate the classifiers on a held-out set of gold standard data. Thus, we compiled a corpus of 91

articles, annotating 31 new documents for ignorance lexical cues and categories over the 60 from our preliminary previous work [82] using the same methodologies 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 [129] with taxonomy lexical cues automatically highlighted and connected to the ignorance taxonomy (Table 2). 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 <u>however</u> 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 were added to the ignorance taxonomy. To capture the scope of each ignorance cue, we adapted the guidelines from BioScope [23] to highlight the whole sentence as the scope capturing all encompassed lexical cues due to difficulties with capturing only parts of the sentences. All annotations were reviewed and adjudicated together with Mayla R. Boguslav (M.R.B) as a separate adjudicator, to create gold-standard articles that achieved an inter-annotator agreement 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 [135, 136]. (See [82] 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 1-2 different articles and then adjudicated all with M.R.B. Since the classic IAA could not be calculated, we calculated an "F1 score" between the original annotation 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%, signifying that the annotator was at least 80% correct in their annotations as discussed with the adjudicator.

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 short comings 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 short comings 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 2 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 2) 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 disagree with each other OR contradictions.	to determine the correct option or a better option and if there are disagreements, or 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 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

Training and evaluating high quality ignorance classifiers

With these extra 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). 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). Any labelled sentences (scopes) or words (lexical cues) in the training set served as the positive samples to classify and any non-labelled sentences or words served as negative samples. Classification occurred at both the sentence- and word-levels, and created both binary and multi-classification problems. At the sentence level, the binary task was to determine if a sentence was a statement of ignorance, no matter the category. The multi-classification task mapped statements of ignorance to the specific ignorance categories of its lexical cues (13 categories plus *not ignorance*). Following our example above, the binary task would map it to a statement of ignorance and the multi-classifier would map the sentence to four different ignorance categories including, *alternative/controversy*, *incompletely understood*, *difficult task*, and *indefinite relationship*. At the word level, the goal was to determine which words in a sentence indicated a lexical cue, either regardless of the category (the binary task - identifying all underlined words), or for the specific taxonomy category

(the multi-classification task - identifying the ignorance categories attached to each word). For both the sentence- and word-level tasks, we created both one true multi-classifier and also 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 classifiers into an ensemble gave the full categorization for each sentence, since each classifier was predicting different categories to predict contradictions. 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 on the held-out test set of 26 articles for (1) the one binary task (ALL CATEGORIES BINARY), (2) the 13 separate binary tasks that together create an ensemble multi-classifier (each category individually), and (3) the one multi-classifier (ALL CATEGORIES COMBINED - the macro-average).

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

Each article was tokenized into sentences and then words used in the respective classification tasks. All models were chosen based on our prior work [82] and on an evaluation of many canonical models for concept recognition [117]. They found that the CRF [137] and BioBERT [138] performed the best for span detection [117]. For sentence classification, we began with the simple ANN from our prior work and added both BERT [139] and BioBERT [138], as they are state-of-the-art language models. The ANN consisted of a flattened layer followed by a couple of dense layers, with early stopping call backs and patience of 5 (more details can be found at [82]). For BioBERT, we used the vocabulary from BioBERT to train the base BERT model. For all models, we used a batch size of 16 and a learning rate of 1e-5. The epochs were tuned based on avoiding overfitting using truncating functions. We did not freeze the layers of the pre-trained BERT model and allowed the weights to keep updating during training to have better performance. For word classification, we represented the underlying data using BIO- tagging: a word at the beginning of a lexical cue is marked 'B'; a word inside a multi-word cue is marked 'I'; a word outside of a cue (i.e. not a word in the lexical cue) is marked 'O'. If the lexical cue contains 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 [117]).

CRF models were tuned with L1 and L2 regularization to avoid overfitting. For BioBERT, the NER baseline parameters performed quite well most likely because it is a similar task, and so we did not tune any other parameters [138]. For a more thorough discussion of all data preparation and representation, the performance metrics, and the classification algorithms please refer to our prior work [82, 117].

Combining ignorance and biomedical concept classifiers

To create the ignorance-base we combined the ignorance and biomedical concept classifiers over all 1,643 prenatal nutrition 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 [117] over all 1,643 articles to automatically identify biomedical concepts represented in ten OBOs, taking the best-performing models for each OBO (highest F1 score - harmonic mean between precision and recall). Most F1 scores ranged from 0.7-0.98 with the exception of PR at 0.53 (see Table 5 in [117]). 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 concepts (a few hundred of each). The ten OBOs used for our work were the same used to manually annotate the CRAFT Corpus [140, 141], a corpus of mouse articles:

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 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 focus 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.)

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 also mapped to UBERON: 0000955, MODULATION OF...IMMUNE SYSTEM mapped to GO:0002682, IMMUNE SYSTEM also mapped to UBERON:0002405, ANTENATAL mapped to GO:0007567, and VITAMIN D mapped to ChEBI:27300. This was all identified by the classifiers. Note that we also identified biomedical concepts in non-ignorance statements. The entailed knowledge goal is to explore the relationship between prenatal vitamin D deficiency and infant outcomes through the important role of vitamin D.

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 first combined all the literature data to connect sentences that have the same ignorance lexical cues, such as poorly understood, and then used PheKnowLator to combine all the sentences with the same 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 each other; these connections were used in exploration by experimental results. This network can be used to search for all sentences that include the lexical cue poorly understood, sentences with the ignorance category *unknown/novel*, sentences that include the biomedical concept VITAMIN D, or any combination of these features. Each sentence relates back to an article with its own metadata including the publication date, which was used to map how ignorance categories change over time for a topic. Note that all sentences in all articles were included whether or not they contain ignorance statements, allowing for the ignorance enrichment comparison to the background information.

Exploration by Topic

The goal of exploration by topic is motivated by the example of finding a thesis topic for a graduate student. To do so, we compared the standard approach to an ignorance approach. An input topic consists of a list of ontology concepts in PheKnowLator. In consultation with a prenatal nutrition specialist (T.L.H.), we mapped the graduate student's 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 all four search terms. For the standard 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.

With all the data, we conducted two analyses for both approaches and compared them: (1) an analysis of the most frequent concepts in the subset of sentences not including vitamin D itself (most frequent) and (2) an analysis of concepts enriched in the subset of sentences as opposed to the background sentences in our ignorance-base, again not including vitamin D itself (concept or ignorance enrichment). We

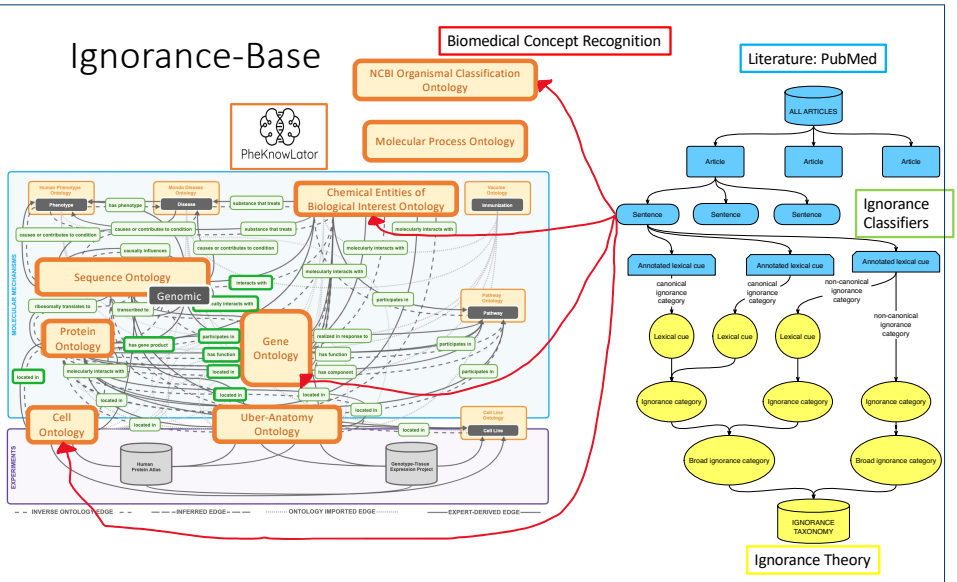


Figure 2: Network representation of the ignorance-base: The top right corner is the literature connecting the articles via tokenized sentences (in blue) to the ignorance taxonomy (in yellow) through the ignorance classifiers (the annotated lexical cues). The sentences also connect to the biomedical concepts on the left with PheKnowLator [131, 132] using the biomedical concept classifiers with the ontologies of interest in bold and larger font.

present both biomedical concept clouds and word clouds along with frequency tables to explore the most frequent terms. 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 [142, 88]. In comparing the approaches, 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 may be 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 are not interesting to us.

For the ignorance approach, we also showed how questions are asked (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 graduate student continued to deep dive into ignorance statements that included the 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 to determine the state of the research.

Exploration by Experimental Results

Connecting experimental results such as a gene list to ignorance statements can identify questions that may bear on the results, providing new avenues for explo-

ration, 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. The input topic was still an OBO concept list, but the extra pre-processing step was connecting the experimental results to OBO concepts in PheKnowLator [131, 132]. 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.

Using our motivating example, we mapped the vitamin D and sPTB gene list (Entrez genes) [6] 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) that connects the different ontologies together in PheKnowLator. For example, 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 contain these terms (our sentences of interest). Note that not all OBO concepts connected to a sentence. From here we performed all of the same analyses as we did for 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 3 for the exploration by experimental results pipeline.)

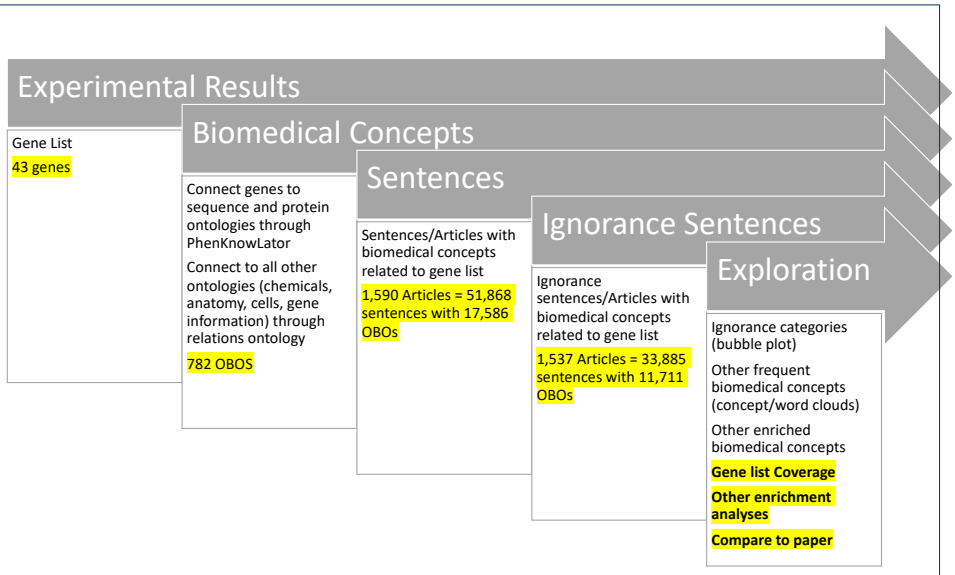


Figure 3: Exploration by Experimental Results (gene list) pipeline: The results are in yellow highlights for the example presented here. For exploration at the end of the pipeline, the three not highlighted are the same as exploration by topic and the three highlighted are the new additions based on a gene list.

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 OBO concepts 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 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 [133], 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 also to find new concepts that may be currently unknown to the functional annotation 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 and identified questions that may bear on it, providing new avenues for exploration from other fields. Yadama et. al. [6] focused on the immune system as their main findings, and we found a new path for the researchers to explore based on the fact that the concepts and knowledge goal statements were not mentioned or cited in the article.

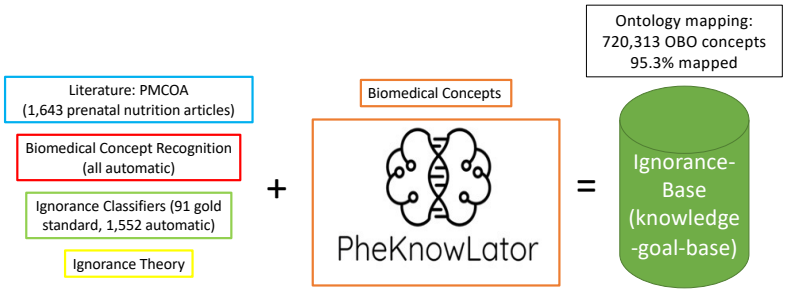
Results

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 it contained a wealth of data (see Figure 4) ripe for exploration 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). The power of combining these two types of classifiers is in the exploration methods.

Creating an expanded gold standard ignorance corpus (see Table 4) yielded high quality 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. These high quality classifiers, built on the 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 was near or above 80% for the classic annotation task [135, 136, 143] and for the split annotations, the annotators were correctly identifying statements of ignorance around 90% of the time. Further, 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.)



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

Figure 4: 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 PhenKnowLator [131, 132].

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 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 ignorance category and all categories combined.

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 Figure 5), 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

Table 6: Word Classification: the best model for word classification for each ignorance category and all categories combined. *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

title, with the most in the discussion and conclusion sections and the fewest in the abstract and results. Our collective scientific ignorance is represented throughout the literature.

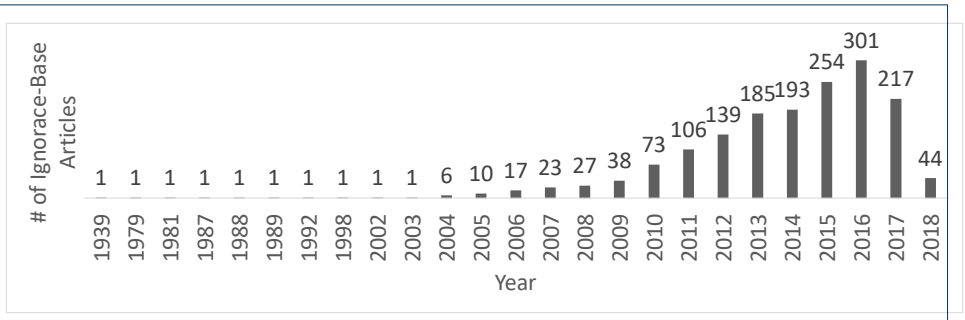


Figure 5: Article date distribution for the ignorance-base (1939-2018).

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 6 with proper definitions and example lexical cues in Table 2). (Note that we renamed our ignorance categories from our previous work [82] to 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 6) were 2,513 unique lexical cues both from related work and collected by 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 the 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. 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.

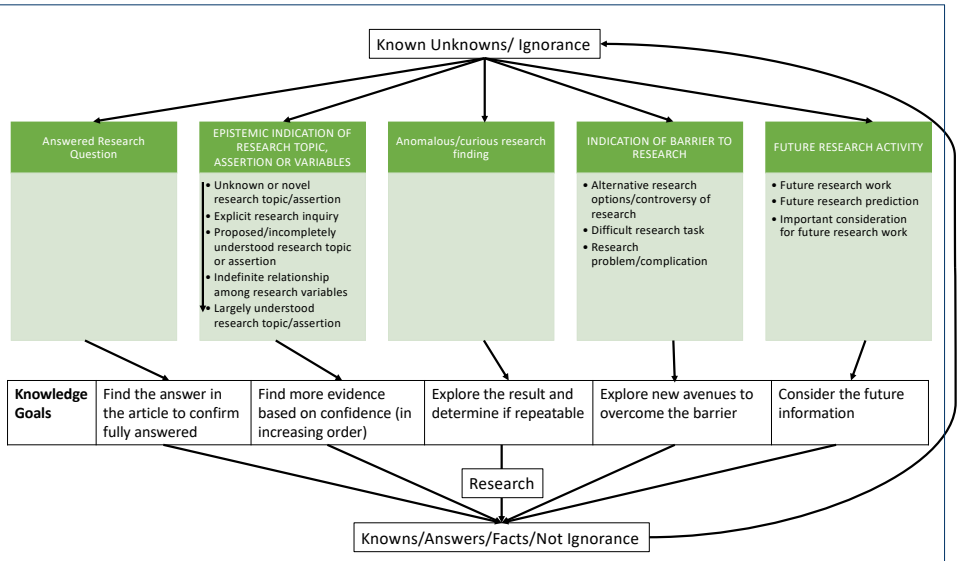


Figure 6: Ignorance taxonomy embedded in the research context: Starting from the top, research starts from known unknowns or ignorance. Our ignorance taxonomy is in green (an ignorance statement is an indication of each ignorance category) with knowledge goals underneath. Research is then conducted based on the knowledge goals to get answers; these then filter back to the known unknowns to identify the next research questions.

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). In terms of errors, there were 850 predicted unique OBO concepts with non-existent OBO identifiers as our biomedical concept classifiers predict identifiers character by character which allows for more errors (see our prior work for more details [117]). The other 1,432 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 ignorance categories, lexical cues, and OBO concepts.

Focusing on ignorance statements provides an alternative targeted exploration of a topic that is distinct from an unfiltered or standard approach

Focusing on ignorance statements provided the graduate student with new avenues of exploration (see Figure 7). We present results for both the standard approach (vi-

tamin D sentences only), the ignorance approach (vitamin D ignorance statements), and a comparison of the two.

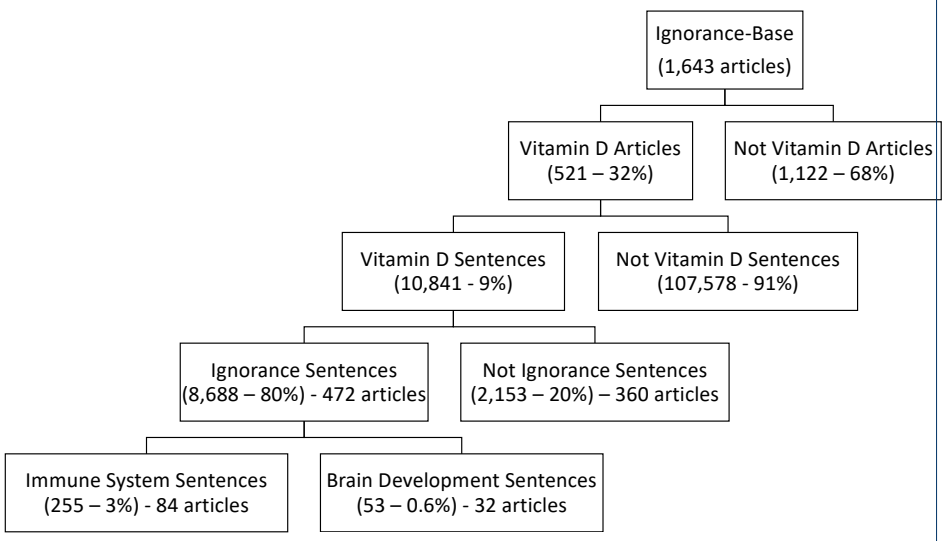


Figure 7: Exploring the ignorance-base by Vitamin D: Searching the ignorance-base for vitamin D yields many articles and sentences that can be explored using ignorance statements to find new research questions, including immune system and brain development.

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 7). 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. We explored this data to differentiate between knowns and unknowns.

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 8a). The first two aligned with the corpus theme of prenatal nutrition. VERSICONOL ACETATE is an intermediate in the biosynthesis of aflatoxin, which is a toxin produced by mold that may be toxic towards the vitamin D receptor in relation to rickets [144]. 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, supplementation, maternal, status, levels, and women were the most frequent terms and 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

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

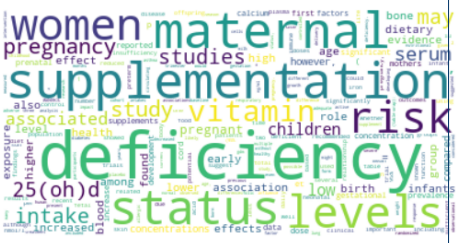
Word	Frequency
supplementation	14.34%
maternal	13.75%
status	13.31%
levels	12.36%
women	10.98%

(a) 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) Vitamin D sentences

Figure 8: Term frequency results: Frequent Biomedical Concepts and Words in (a) vitamin D ignorance statements and (b) 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.

captured meaningful information. Taking one of these terms or concepts provided avenues for the graduate student 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 not digestion. Thus, searching for "vitamin D" and "feeding behavior" in the literature, we found that vitamin intake during pregnancy in general affects both the metabolic system and food intake regulatory pathways in the offspring [145]. Therefore, frequent concepts that appear with a topic can provide keyword search terms for the graduate student.

Comparing the ignorance-approach to the standard approach highlighted terms that are more known unknowns vs. knowns. The most frequent concepts in the standard approach were the same as the ignorance approach (see Figure 8b), which

may signify that term frequency did not capture the difference in 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 non-ignorance approach, the term women disappeared (see Figure 8). This may signify that vitamin D deficiency is 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 is still unknown in general. The ignorance frequency term list provided an avenue for exploration, FEEDING BEHAVIOR. In addition, comparing this term list to the results from the standard approach may help distinguish between known unknowns and knowns, such as the word "deficiency".

Our ignorance enrichment method also 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. The ignorance approach found 11 ignorance enriched concepts, whereas the standard approach found 275, with an overlap of eight concepts (see Figure 9). 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 9). Further, in comparing the two different approaches, the ignorance approach found concepts that were more systems or broad categories, including immune system and brain development, compared to the standard approach which were more specific entities, such as blood serum and vitamin K.

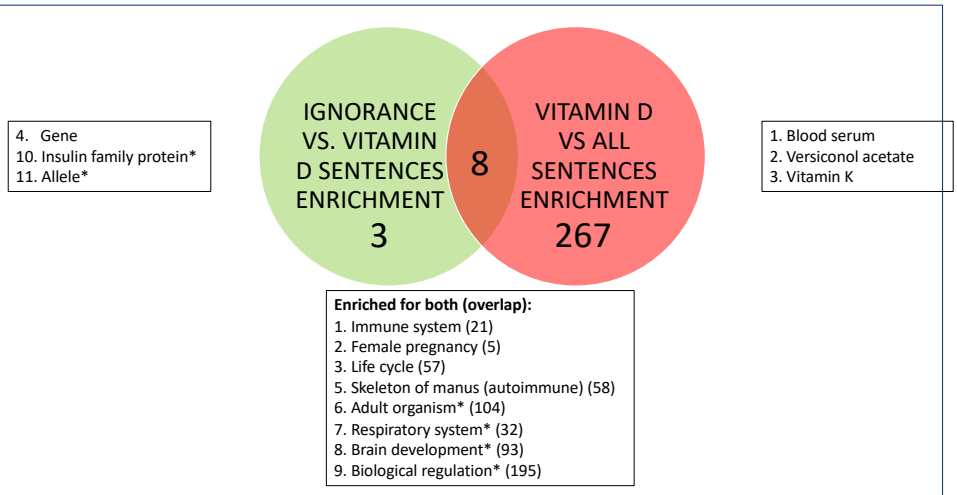


Figure 9: Comparison of standard and ignorance enrichment: A Venn diagram of biomedical concept enrichment between just vitamin D (pink) and ignorance vitamin D (green) sentences. Next to each bubble are concepts in their respective enrichment orders. The concepts in the middle are the overlap and the 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.

Focusing on any of the ignorance-enriched concepts, the ignorance approach provided the graduate student with a smaller list of targeted statements of knowledge goals to potentially pursue or spark ideas from. 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 and appear to be currently studied, with more room for future work. They 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 [146]. Deep diving into the concept IMMUNE SYSTEM in relation to VITAMIN D provided the graduate student with 255 ignorance statements with their entailed knowledge goals that spanned 84 articles (see Table 7 for the top eight articles with the most ignorance statements). Note that only one article had no ignorance statements that included VITAMIN D or IMMUNE SYSTEM. Thus, only using the ignorance statements themselves, we have already found a set of articles and sentences for the graduate student.

Table 7: 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

Understanding and tracing ignorance categories over time can both help the student 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 10), spanning all the thirteen categories of ignorance with ten enriched in VITAMIN D ignorance statements as compared to all ignorance statements (see the green highlights in Figure 10). For example, *unknown/novel* was enriched in this domain, pointing to what we do not know about the context of VITAMIN D in pregnancy and fetal development. Further, looking at the bubble plot for IMMUNE SYSTEM and VITAMIN D ignorance statements (see Figure 11) showed that *Unknown/novel* is spread out amongst the different articles. Now the graduate student can continue to narrow their topic not only with a biomedical concept, such as IMMUNE SYSTEM, but also with an ignorance category, such as *unknown/novel*.

Since the goal was to find the graduate student a thesis topic as an example of exploration by topic, we chose to dive deeper into the *unknown/novel* category for VITAMIN D and IMMUNE SYSTEM. Below are some ignorance sentences from this set with the biomedical concepts capitalized:

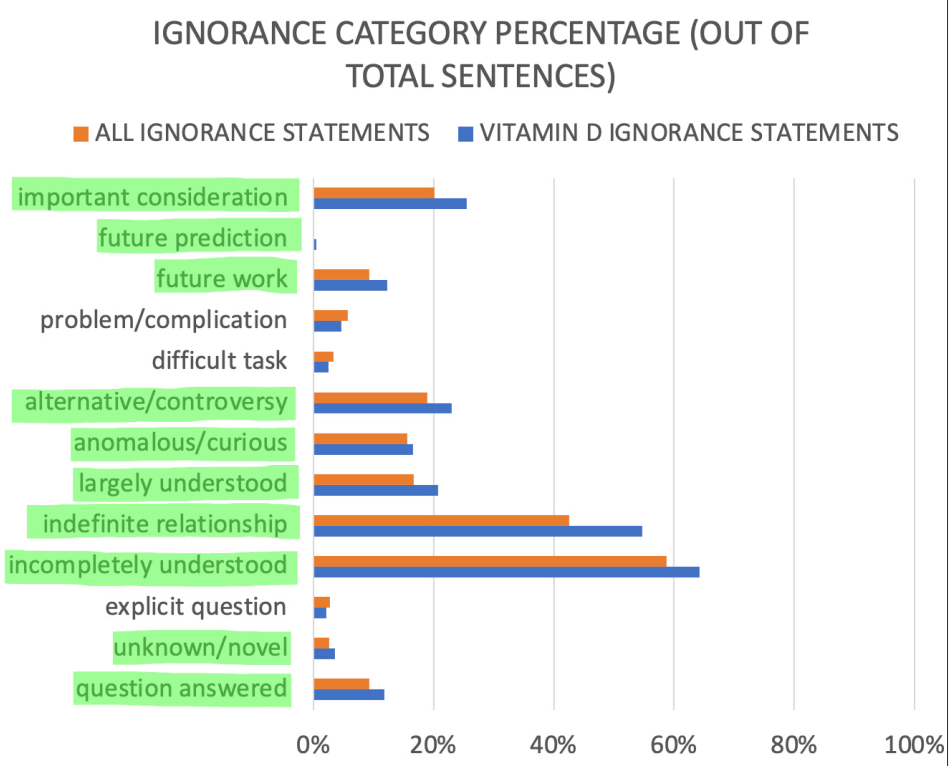


Figure 10: Ignorance-category enrichment: Ignorance vitamin D sentences compared to all ignorance sentences. The 10 categories highlighted in green were enriched.

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."
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."
3 "however, findings regarding the combined effects of PRENATAL and POSTNATAL VITAMIN D status on fs, two of the most critical periods for IMMUNE SYSTEM DEVELOPMENT (19,20), are unclear."
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."
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."

There were many *unknown/novel* ignorance sentences ripe for exploration. The overall thesis topic or knowledge goal that the graduate student could research is the need to explore the relationship between VITAMIN D and the IMMUNE

SYSTEM especially in pregnancy. 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 is the need to determine if VITAMIN D and BRAIN DEVELOPMENT are truly linked. Thus, from querying the ignorance-base for the topic VITAMIN D, the graduate student now has ignorance statements and knowledge goals connecting VITAMIN D to the IMMUNE SYSTEM and BRAIN DEVELOPMENT (see Figure 7).

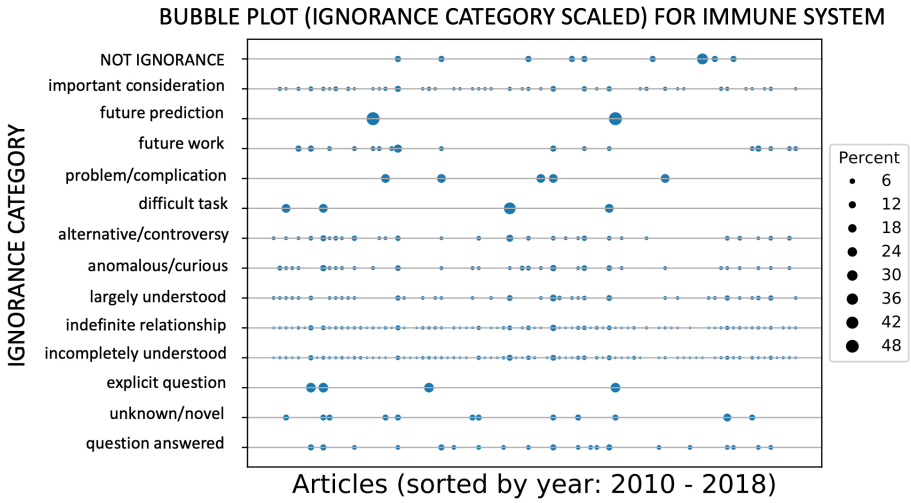


Figure 11: How ignorance changes over time: A bubble plot of vitamin D and immune system sentences (including non-ignorance sentences). The x-axis is the articles sorted by time. The y-axis is the ignorance categories. Each bubble represents the portion of sentences in each article in that ignorance 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.

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

Connecting a vitamin D and sPTB gene list from a paper [6] to ignorance statements found a new avenue for exploration, BRAIN DEVELOPMENT, that was not mentioned in the paper, and its implied field, neuroscience, could possibly help find answers. Following the exploration by experimental results pipeline (see Figure 3), the 43 genes mapped to 782 OBO concepts. These OBOs connected to 51,868 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 the sentences that 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.

With the gene list connected to so many concepts (782), combining gene list coverage and ignorance enrichment helped prioritize concepts (see Table 8). The highest covered OBO concept was PROTEIN CODING GENE (SO:0001217). In the top 25 most covered OBO concepts, some were enriched in all gene list sentences, but not in ignorance sentences (established information); some were not enriched for either (not interesting), and some were enriched for both (currently studied). (None were enriched only in ignorance and not in the gene list sentences - 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. Note that there were some IMMUNE SYSTEM and BRAIN related concepts. These concepts were ripe for exploration in relation to the gene list to find knowledge goals that may bear on them.

Table 8: Gene list coverage enrichment information for the top 25: NO INFORMATION means 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 INFORMATION	NO INFORMATION
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 INFORMATION	NO INFORMATION
GO:0045087	innate immune response	11	YES	YES*
GO:0070062	extracellular exosome	11	YES	NO
GO:0016021	bone marrow	10	NO INFORMATION	NO INFORMATION
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 INFORMATION	NO INFORMATION
CL:0000771	eosinophil	6	YES	NO

Combining other canonical enrichment methods with ignorance enrichment also helped prioritize the many OBO concepts produced by our gene list (see Figure 12 focusing only on the gene ontology). DAVID [147, 133] 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.

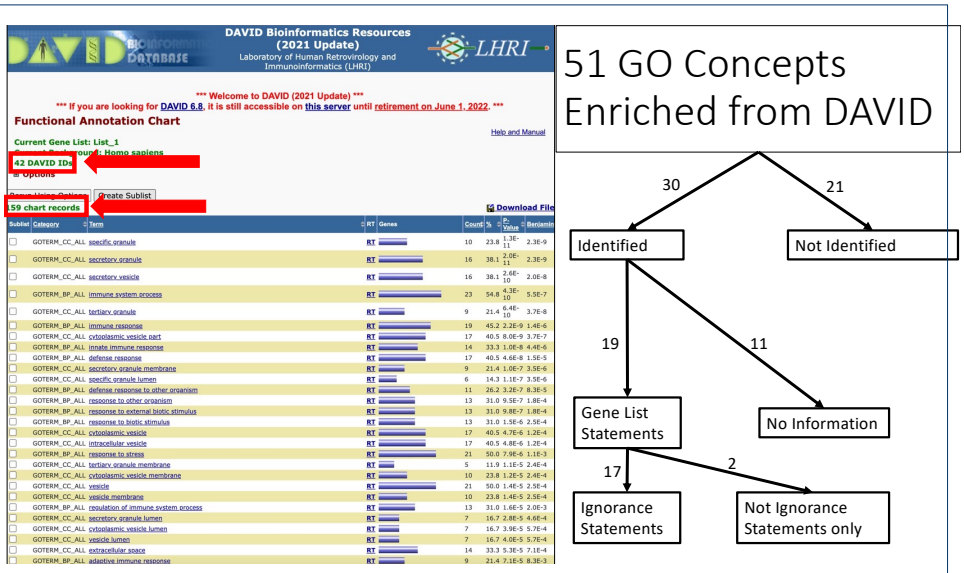


Figure 12: Enhancing canonical enrichment analysis using the ignorance-base: DAVID enrichment analysis for the gene ontology (GO) in relation to the ignorance-base. The DAVID initial analysis is on the left with 42 of the 43 genes found in DAVID mapping to 159 GO concepts. The right is a breakdown of where the 51 enriched GO concepts from DAVID fall within the ignorance-base.

When comparing ignorance enrichment to DAVID, we found more ignorance in general compared to established information. 3,173 concepts were enriched in ignorance with 159 in DAVID (see Figure 13). 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 knowledge. On the knowledge-side, most enriched concepts from DAVID pertained to the immune system which was the main finding of Yadama et. al. [6]. The two overlapping concepts, IMMUNE RESPONSE and INNATE IMMUNE RESPONSE, also pertain to immunity and the authors admit that there is still room for future work. These concepts are currently studied.

Looking at the ignorance enriched concepts only, we found new avenues to investigate that were not mentioned by the paper [6] that provided a new field to examine for answers. The top three ignorance-enriched GO concepts included FEMALE PREGNANCY, BIOLOGICAL REGULATION, and METABOLIC PROCESS (see Figure 12). Broadening this exploration beyond GO, there were 130 total ignorance enriched concepts for this gene list. (The * means that they were statistically significant with FDR but not family-wise error.) 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*. As mentioned above, Yadama et. al. [6] admit that there is still future work to be done on understanding

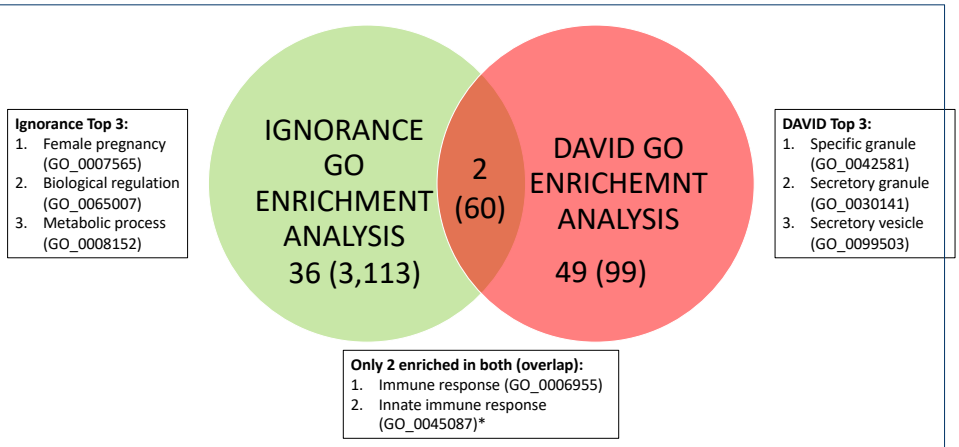


Figure 13: Comparison of DAVID and ignorance enrichment: A Venn diagram of gene ontology enrichment between DAVID (pink) and the ignorance-base (green). In parentheses are the total number of concepts found in each category without enrichment. Next to each bubble are the top three concepts for each enrichment method. The concepts in the middle are the overlap. *Statistically significant with FDR but not family-wise error.

the IMMUNE SYSTEM in relation to sPTB and VITAMIN D. 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. There were also no brain-related concepts in the DAVID analysis, signifying that this information is not yet established knowledge. This may be a new emerging topic that the authors could examine more in relation to vitamin D and spontaneous preterm birth and possibly look to the field of neuroscience for answers. We provide some sentences from five papers to explore below (the biomedical concepts of interest are capitalized):

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.4the importance of adequate dha 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." (PMC3607807)

4 "at present, preterm infants have a limit of viability (50% survival rate) of around 23–24 weeks ga 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 DEVELOPMENT9but plays a crucial role in CARNITINE SYNTHESIS,10although CARNITINE precursors may be more important.11is an important COFACTOR for more than 300 identified zinc metalloenzymes.12zinc insufficiency in late PREGNANCY disrupts NEURONAL REPLICATION and SYNAPTOGENSIS,13and maternal deficiency is associated with decreased dna, rna, and protein content of the f1 BRAIN.14zinc deficiency affects one in five world inhabitants.14ZINC supplementation reduces the risk of PRETERM BIRTH, though not sga.14D deficiency is under investigation for its role in protection against dm, cv, some ca, osteoporosis, and optimization of IMMUNE function.15VITAMIN D might be an important mediator in GUT HOMEOSTASIS and in signaling between microbiota and host.16the INTESTINAL microbiome in both newborns and LACTATING mothers influences infant and childhood FOOD allergy and eczema." (PMC4268639)

In the paper, Yadama et. al. focused on the mother's immune system, but it is possible that the 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 is 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. The ignorance-base has provided a novel avenue (and field), BRAIN DEVELOPMENT (neuroscience), along with specific knowledge goal statements that the authors can explore starting from their initial gene list.

Discussion

Focusing on ignorance statements through our ignorance-base and exploration methods can lead to new research avenues that could help accelerate translational research. The ignorance-base, containing articles spanning 1939-2018, has predicted areas of research that are currently studied and an emerging topic with a corresponding field that may prove fruitful to help find answers. The exploration by topic method applied to a graduate student interested in vitamin D, showed that vitamin D may play an important role in the immune system, respiratory system,

and brain development using ignorance-enrichment (see Figures 7- 11 and Table 7). Recent review articles [112, 113, 114, 115, 116] 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. One of the example sentences, "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) [148] shows that further research on the impact of vitamin D on infant outcomes is needed. The context of the sentence was also important: it is from a 2014 study in Rural Vietnam and appears in the abstract objective section. Because we can sort our ignorance statements by time and section, we showed that since 2014, more research has been conducted on this topic [114] as found in our ignorance-base. Further, we can track how research questions emerge using our ignorance taxonomy (see Figure 11). Ideally, we would create an ignorance-base over the entire body of scientific literature to truly help researchers, students, funders, and publishers capture the landscape of our collective scientific ignorance and how research questions evolve over time. Even this smaller-scale effort demonstrates that this is possible.

We further demonstrated that the ignorance-base and exploration by experimental results method can find an emerging topic (see Figures 3, 12, 13, and Table 8). Ignorance enrichment of the 43 genes in common between vitamin D and spontaneous preterm birth (sPTB) [6] found many concepts that relate to the brain and some that relate to the immune system (as found by [6]). This suggests 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) [149]. This sentence ties all the relevant concepts together: vitamin D may be causing in utero inflammation which is leading to preterm labor, and due to the preterm labor there is insufficient development of the gut-blood-brain barrier leading to potential neurotoxins for the fetus. This article is not cited by Yadama et. al. [6], but may be 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. 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?

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 ([112, 113, 114, 115, 116]) that discuss the impact of vitamin D on maternal and fetal health (see Figure 3 in [112] and our adapted Figure 1). All of these studies agree that there is a link 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" [112]. 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. Thus, we provided the ignorance-context to the genes in addition to the knowledge-context.

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 [131, 132]. We also utilized DAVID [133] as a comparison knowledge-base (see Figures 12 and 13) as well as other canonical methods (gene list coverage) to help prioritize the most relevant biomedical concepts (see Table 8). 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. All of these limitations could be addressed in future work. Further, there are many other knowledge-bases and methods that can also 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. The scale of the ignorance-base was small; ideally we would create an ignorance-base that includes all articles (or at least PMCOA to start). We focused only on the prenatal nutrition literature, and future work can determine if the ignorance taxonomy and methods generalize outside of it. In terms of exploration methods, concept enrichment provided more fruitful concepts (see Figures 9 and 10) than concept frequency (see Figure 8). Another avenue to explore would be co-occurrence terms. The creation of a tool 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.

Even with all these limitations, the current ignorance-base showed that 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. In addition, with the highlights, the annotators continued to find new lexical cues, which further extended the reach of the classifiers. More data can always be annotated from other fields besides prenatal nutrition. We used the gold standard ignorance corpus to create the ignorance classifiers and they performed quite well (see Tables 5 and 6) using the baseline hyperparameters. However, other parameter tunings and other algorithms, such as PubMedBERT [150], may yield improved results. For our biomedical concept classifiers, CRAFT [141] is a corpus of mouse articles, not prenatal nutrition, and it only includes 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 [151]), although all of them have their pros and cons.

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 and 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 along with two new exploration methods that can find statements of unknowns, controversies, and difficulties ripe for exploration. Just as the literature contains both knowledge and ignorance, so too can both knowledge-bases and ignorance-bases help the researcher navigate the literature to find the next important questions or knowledge gaps.

Overall, the ultimate goal of this work is to help students and researchers explore the literature through the questions and known unknowns. The scientific endeavor rests on our continuous ability to ask questions and push research farther as we learn more knowledge. The two methods presented here, exploration by topic and exploration by experimental results, provide two ways to help students, researchers, funders, and publishers explore the state of our scientific ignorance.

Conclusion

The ultimate goal is that students, researchers, funders, and publishers can use our ignorance-base and methods to help find the next important questions or knowledge gaps. These can possibly help accelerate translational research over purely manual methods. The exploration by topic method not only found new avenues for exploration for the graduate student interested in vitamin D using our novel method of ignorance enrichment (the immune system, respiratory system, and brain development), but also elucidated how questions are asked and how that changes

over time using our novel method of ignorance-category enrichment. Our exploration by experimental results found specific knowledge goals to pursue and found an emerging topic (brain development), with a corresponding field (neuroscience), that bear on a sPTB and vitamin D gene list. These questions 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 brings 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).

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.

Author details
[1]Computational Bioscience Program University of Colorado Anschutz Medical Campus, E 17th Avenue, 80045 Aurora, CO, USA. [2]Center for Genes, Environment and Health, National Jewish Health, Jackson Street, 80206 Denver, CO, USA. [3]College of Nursing, Department of Medicine/Division of Endocrinology, Metabolism, Diabetes, University of Colorado Anschutz Medical Campus, E 17th Avenue, 80045 Aurora, CO, USA.

References
1. Firestein, S.: Ignorance: How It Drives Science. OUP, USA (2012)
2. Kuhn, T.S.: The Structure of Scientific Revolutions, [2d ed., enl edn. International encyclopedia of unified science. Foundations of the unity of science, v. 2, no. 2. University of Chicago Press, Chicago (1970)
3. O'leary, Z.: The Essential Guide to Doing Research. Sage, Great Britain (2004)
4. Callahan, T.J., Tripodi, I.J., Pielke-Lombardo, H., Hunter, L.E.: Knowledge-based biomedical data science. Annual review of biomedical data science **3**, 23 (2020)
5. US Department of Health and Human Services, National Institutes of Health, Dietary Supplement Label Database (DSLD), Office of Dietary Supplements. https://dsld.nlm.nih.gov/dsld/
6. Yadama, A.P., Mirzakhani, H., McElrath, T.F., Litonjua, A.A., Weiss, S.T.: Transcriptome analysis of early pregnancy vitamin d status and spontaneous preterm birth. PLoS One **15**(1), 0227193 (2020)
7. Sternberg, D.: How to Complete and Survive a Doctoral Dissertation. St. Martin's Griffin, New York (1981)

8. Brause, R.S.: Writing Your Doctoral Dissertation: Invisible Rules for Success. Routledge, London (2012)
9. Burton, S., Steane, P.: Surviving Your Thesis. Routledge, London (2004)
10. Krathwohl, D.R., Smith, N.L.: How to Prepare a Dissertation Prosposal: Suggestions for Students in Education & the Social and Behavioral Sciences. Syracuse University Press, Syracuse, NY (2005)
11. Terrell, S.R.: Writing a Proposal for Your Dissertation: Guidelines and Examples. Guilford Publications, New York (2022)
12. Tanaka, M.L.: A thesis proposal development course for engineering graduate students. Journal of Biomechanical Engineering **142**(11) (2020)
13. Madsen, D.: Successful dissertations and theses: A guide to graduate student research from proposal to completion. (1983)
14. Lei, S.A.: Strategies for finding and selecting an ideal thesis or dissertation topic: A review of literature. College Student Journal **43**(4), 1324–1333 (2009)
15. Ségol, G.: Choosing a dissertation topic: Additional pointers. College Student Journal **48**(1), 108–113 (2014)
16. Eze, U.A., Adebayo, O., Nnodim, I.J., Adejo, A.C., Obazenu, L.O.: How to choose a dissertation topic. Nigerian Journal of Medicine **30**(2), 123–124 (2021)
17. Fiarni, C., Gunawan, A.S., Maharani, H., Kurniawan, H., *et al.*: Automated scheduling system for thesis and project presentation using forward chaining method with dynamic allocation resources. Procedia Computer Science **72**, 209–216 (2015)
18. Lakoff, G.: Hedges: A study in meaning criteria and the logic of fuzzy concepts. In: Contemporary Research in Philosophical Logic and Linguistic Semantics, pp. 221–271. Springer, Dordrecht (1975)
19. Hyland, K.: Hedging in Scientific Research Articles vol. 54. John Benjamins Publishing, Amsterdam/Philadelphia (1998)
20. Farkas, R., Vincze, V., Móra, G., Csirik, J., Szarvas, G.: 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, pp. 1–12 (2010)
21. Ganter, V., Strube, M.: Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 173–176 (2009)
22. Medlock, B., Briscoe, T.: Weakly supervised learning for hedge classification in scientific literature. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 992–999 (2007)
23. Vincze, V., Szarvas, G., Farkas, R., Móra, G., Csirik, J.: The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC bioinformatics **9**(11), 1–9 (2008)
24. Kilicoglu, H., Bergler, S.: Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. BMC bioinformatics **9**(S11), 10 (2008)
25. Kilicoglu, H., Rosemblat, G., Rindflesch, T.C.: Assigning factuality values to semantic relations extracted from biomedical research literature. PloS one **12**(7), 0179926 (2017)
26. Light, M., Qiu, X.Y., Srinivasan, P.: The language of bioscience: Facts, speculations, and statements in between. In: HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases, pp. 17–24 (2004)
27. Zerva, C., Batista-Navarro, R., Day, P., Ananiadou, S.: Using uncertainty to link and rank evidence from biomedical literature for model curation. Bioinformatics **33**(23), 3784–3792 (2017)
28. Al-Khawaldeh, F.T.: Hierarchical attention generative adversarial networks for biomedical texts uncertainty detection. Int J Adv Stud Comput Sci Eng **8**(6), 1–12 (2019)
29. Szarvas, G., Vincze, V., Farkas, R., Móra, G., Gurevych, I.: Cross-genre and cross-domain detection of semantic uncertainty. Computational Linguistics **38**(2), 335–367 (2012)
30. Velldal, E.: Detecting uncertainty in biomedical literature: A simple disambiguation approach using sparse random indexing. In: Semantic Mining in Biomedicine (2010)
31. Fujikawa, K., Seki, K., Uehara, K.: 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, pp. 67–74 (2012)
32. Ren, Y., Fei, H., Peng, Q.: 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), pp. 739–742 (2018). IEEE
33. Konstantinova, N., De Sousa, S.C.: Annotating negation and speculation: the case of the review domain. In: Proceedings of the Second Student Research Workshop Associated with RANLP 2011, pp. 139–144 (2011)
34. Zavala, R.R., Martinez, P.: The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: Comparative study. JMIR Medical Informatics **8**(12), 18953 (2020)
35. Apostolova, E., Tomuro, N., Demner-Fushman, D.: 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, pp. 283–287 (2011)
36. Qian, Z., Li, P., Zhu, Q., Zhou, G., Luo, Z., Luo, W.: Speculation and negation scope detection via convolutional neural networks. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 815–825 (2016)
37. Fei, H., Ren, Y., Ji, D.: Negation and speculation scope detection using recursive neural conditional random fields. Neurocomputing **374**, 22–29 (2020)
38. Szarvas, G.: Hedge classification in biomedical texts with a weakly supervised selection of keywords. In: Proceedings of Acl-08: HLT, pp. 281–289 (2008)
39. AL-Khawaldeh, F.T.: Speculation and negation annotation for arabic biomedical texts: Bioarabic corpus. World of Computer Science & Information Technology Journal (2016)
40. Khandelwal, A., Britto, B.K.: Multitask learning of negation and speculation using transformers. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pp. 79–87 (2020)
41. Al-Khawaldeh, F.T.: Speculation and negation detection for arabic biomedical texts. World of Computer

Science & Information Technology Journal **9**(3) (2019)
42. Cheng, K., Baldwin, T., Verspoor, K.: Automatic negation and speculation detection in veterinary clinical text. In: Proceedings of the Australasian Language Technology Association Workshop 2017, pp. 70–78 (2017)
43. Velldal, E.: Predicting speculation: a simple disambiguation approach to hedge detection in biomedical literature. Journal of Biomedical Semantics **2**(5), 1–14 (2011)
44. Agarwal, S., Yu, H.: Detecting hedge cues and their scope in biomedical text with conditional random fields. Journal of biomedical informatics **43**(6), 953–961 (2010)
45. Zhou, H., Li, X., Huang, D., Li, Z., Yang, Y.: 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, pp. 106–113 (2010)
46. Zhou, H., Deng, H., Huang, D., Zhu, M.: Hedge scope detection in biomedical texts: an effective dependency-based method. PloS one **10**(7), 0133715 (2015)
47. Medlock, B.: Exploring hedge identification in biomedical literature. Journal of biomedical informatics **41**(4), 636–654 (2008)
48. Ji, F., Qiu, X., Huang, X.-J.: Detecting hedge cues and their scopes with average perceptron. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task, pp. 32–39 (2010)
49. Hanauer, D.A., Liu, Y., Mei, Q., Manion, F.J., Balis, U.J., Zheng, K.: 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, p. 321 (2012). American Medical Informatics Association
50. Zhao, Q., Sun, C.-J., Liu, B., Cheng, Y.: Learning to detect hedges and their scope using crf. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task, pp. 100–105 (2010)
51. Clausen, D.: Hedgehunter: A system for hedge detection and uncertainty classification. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task, pp. 120–125 (2010)
52. Morante, R., Daelemans, W.: Learning the scope of hedge cues in biomedical texts. In: Proceedings of the BioNLP 2009 Workshop, pp. 28–36 (2009)
53. Liu, F., Zhou, P., Baccei, S.J., Masciocchi, M.J., Amornsiripanitch, N., Kiefe, C.I., Rosen, M.P.: Qualifying certainty in radiology reports through deep learning–based natural language processing. American Journal of Neuroradiology **42**(10), 1755–1761 (2021)
54. Verbeke, M., Frasconi, P., Van Asch, V., Morante, R., Daelemans, W., De Raedt, L.: Kernel-based logical and relational learning with klog for hedge cue detection. In: International Conference on Inductive Logic Programming, pp. 347–357 (2011). Springer
55. Mowery, D.L., Velupillai, S., Chapman, W.: 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, pp. 56–64 (2012)
56. Han, P.K., Klein, W.M., Arora, N.K.: Varieties of uncertainty in health care: a conceptual taxonomy. Medical Decision Making **31**(6), 828–838 (2011)
57. Pearl, J., Mackenzie, D.: The Book of Why: the New Science of Cause and Effect. Basic Books, ??? (2018)
58. Regan, H.M., Colyvan, M., Burgman, M.A.: A taxonomy and treatment of uncertainty for ecology and conservation biology. Ecological applications **12**(2), 618–628 (2002)
59. Smithson, M.: Ignorance and Uncertainty: Emerging Paradigms. Springer, ??? (2012)
60. Walker, V.R.: The siren songs of science: toward a taxonomy of scientific uncertainty for decisionmakers. Conn. L. Rev. **23**, 567 (1990)
61. Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M.H., Bug, B., Chibucos, M.C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., *et al.*: The ontology for biomedical investigations. PloS one **11**(4), 0154556 (2016)
62. Bastian, F.B., Chibucos, M.C., Gaudet, P., Giglio, M., Holliday, G.L., Huang, H., Lewis, S.E., Niknejad, A., Orchard, S., Poux, S., et al.: The confidence information ontology: a step towards a standard for asserting confidence in annotations. Database **2015** (2015)
63. Brush, M.H., Shefchek, K., Haendel, M.: Sepio: A semantic model for the integration and analysis of scientific evidence. In: ICBO/BioCreative (2016)
64. Chibucos, M.C., Mungall, C.J., Balakrishnan, R., Christie, K.R., Huntley, R.P., White, O., Blake, J.A., Lewis, S.E., Giglio, M.: Standardized description of scientific evidence using the evidence ontology (eco). Database **2014** (2014)
65. Rindflesch, T.C., Fiszman, M.: The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of biomedical informatics **36**(6), 462–477 (2003)
66. Thompson, P., Nawaz, R., McNaught, J., Ananiadou, S.: Enriching a biomedical event corpus with meta-knowledge annotation. BMC bioinformatics **12**(1), 1–18 (2011)
67. Bongelli, R., Canestrari, C., Riccioni, I., Zuczkowski, A., Buldorini, C., Pietrobon, R., Lavelli, A., Magnini, B.: A corpus of scientific biomedical texts spanning over 168 years annotated for uncertainty. In: LREC, vol. 12, pp. 2009–2014 (2012)
68. Díaz, N.P.C.: Detecting negated and uncertain information in biomedical and review texts. In: Proceedings of the Student Research Workshop Associated with RANLP 2013, pp. 45–50 (2013)
69. Konstantinova, N., De Sousa, S.C., Díaz, N.P.C., López, M.J.M., Taboada, M., Mitkov, R.: A review corpus annotated for negation, speculation and their scope. In: Lrec, pp. 3190–3195 (2012)
70. Zhou, H., Yang, H., Zhang, J., Kang, S., Huang, D.: The research and construction of chinese hedge corpus. Journal of Chinese Information Processing **29**(6), 83–89 (2015)
71. Sanchez, L.M., Vogel, C.: A hedging annotation scheme focused on epistemic phrases for informal language. In: Proceedings of the Workshop on Models for Modality Annotation (2015)
72. Jiménez-Zafra, S.M., Taulé, M., Martín-Valdivia, M.T., Urena-López, L.A., Martí, M.A.: Sfu review sp-neg: a spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns. Language

Resources and Evaluation **52**(2), 533–569 (2018)
73. Yang, H., De Roeck, A., Gervasi, V., Willis, A., Nuseibeh, B.: Speculative requirements: Automatic detection of uncertainty in natural language requirements. In: 2012 20th IEEE International Requirements Engineering Conference (RE), pp. 11–20 (2012). IEEE
74. Jean, P.-A., Harispe, S., Ranwez, S., Bellot, P., Montmain, J.: Uncertainty detection in natural language: A probabilistic model. In: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, pp. 1–10 (2016)
75. Sergeeva, E., Zhu, H., Tahmasebi, A., Szolovits, P.: 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), pp. 178–187 (2019)
76. Zhou, H., Huang, D., Li, X., YANG, Y.: Combining structured and flat features by a composite kernel to detect hedges scope in biological texts. Chinese Journal of Electronics **20**(3), 476–482 (2011)
77. Zhou, H., Li, X., Huang, D., Yang, Y., Ren, F.: Voting-based ensemble classifiers to detect hedges and their scopes in biomedical texts. IEICE TRANSACTIONS on Information and Systems **94**(10), 1989–1997 (2011)
78. Zhou, H., Xu, J., Yang, Y., Deng, H., Chen, L., Huang, D.: Chinese hedge scope detection based on structure and semantic information. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 204–215. Springer, ??? (2016)
79. Georgescul, M.: A hedgehop over a max-margin framework using hedge cues. In: Proceedings of the 14th International Conference on Computational Natural Language Learning: Shared Task, pp. 26–31 (2010)
80. Moncecchi, G.: Recognizing speculative language in research texts. PhD thesis, Université de Nanterre-Paris X; Universidad de la República-Proyecto de ... (2013)
81. Díaz, N.P.C.: Detección de la negación y la especulación en textos médicos y de. PhD thesis, Citeseer (2014)
82. Boguslav, M.R., Salem, N.M., White, E.K., Leach, S.M., Hunter, L.E.: Identifying and classifying goals for scientific knowledge. Bioinformatics Advances **1** (2021). doi:10.1093/bioadv/vbab012
83. Turner, M., Ive, J., Velupillai, S.: Linguistic uncertainty in clinical nlp: A taxonomy, dataset and approach. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 129–141 (2021). Springer
84. Dalloux, C., Claveau, V., Grabar, N.: Speculation and negation detection in french biomedical corpora. In: RANLP 2019-Recent Advances in Natural Language Processing, pp. 1–10 (2019)
85. Zou, B., Zhu, Q., Zhou, G.: 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), pp. 656–665 (2015)
86. Zhang, S., Kang, T., Zhang, X., Wen, D., Elhadad, N., Lei, J.: Speculation detection for chinese clinical notes: Impacts of word segmentation and embedding models. Journal of biomedical informatics **60**, 334–341 (2016)
87. Islam, J., Xiao, L., Mercer, R.E.: A lexicon-based approach for detecting hedges in informal text. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 3109–3113 (2020)
88. Tipney, H., Hunter, L.: An introduction to effective use of enrichment analysis software. Human genomics **4**(3), 1–5 (2010)
89. Shi Jing, L., Fathiah Muzaffar Shah, F., Saberi Mohamad, M., Moorthy, K., Deris, S., Zakaria, Z., Napis, S.: A review on bioinformatics enrichment analysis tools towards functional analysis of high throughput gene set data. Current Proteomics **12**(1), 14–27 (2015)
90. Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., DeLisi, C.: Gene set enrichment analysis: performance evaluation and usage guidelines. Briefings in bioinformatics **13**(3), 281–291 (2012)
91. Curtis, R.K., Orešič, M., Vidal-Puig, A.: Pathways to the analysis of microarray data. TRENDS in Biotechnology **23**(8), 429–435 (2005)
92. Wijesooriya, K., Jadaan, S.A., Perera, K.L., Kaur, T., Ziemann, M.: Urgent need for consistent standards in functional enrichment analysis. PLoS computational biology **18**(3), 1009935 (2022)
93. Rihm, B.H., Vidal, S., Nemurat, C., Vachenc, S., Mohr, S., Mazur, F., Houdry, P., Grandjean, F., Visvikis, S., Ducloy, J.: From transcriptomics to bibliomics. Medical Science Monitor **9**(8), 89–95 (2003)
94. VIDAL, S., DUCLOY, J., HOUDRY, P.: Mining medical data using multiple corpora interaction: the transcriptomics investigation server experiment. Systemics Cybernetics and Informatics (Actes du congrès à Houston, USA) **7**
95. Chen, J., Xu, H., Aronow, B.J., Jegga, A.G.: Improved human disease candidate gene prioritization using mouse phenotype. BMC bioinformatics **8**(1), 1–13 (2007)
96. Jourquin, J., Duncan, D., Shi, Z., Zhang, B.: Glad4u: deriving and prioritizing gene lists from pubmed literature. BMC genomics **13**(8), 1–12 (2012)
97. Jenssen, T.-K., Lægreid, A., Komorowski, J., Hovig, E.: A literature network of human genes for high-throughput analysis of gene expression. Nature genetics **28**(1), 21–28 (2001)
98. Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., Stoehr, P.: Ebimed—text crunching to gather facts for proteins from medline. Bioinformatics **23**(2), 237–244 (2007)
99. Fontaine, J.-F., Priller, F., Barbosa-Silva, A., Andrade-Navarro, M.A.: Genie: literature-based gene prioritization at multi genomic scale. Nucleic acids research **39**(suppl_2), 455–461 (2011)
100. Börnigen, D., Tranchevent, L.-C., Bonachela-Capdevila, F., Devriendt, K., De Moor, B., De Causmaecker, P., Moreau, Y.: An unbiased evaluation of gene prioritization tools. Bioinformatics **28**(23), 3081–3088 (2012)
101. Grimes, G., Wen, T., Mewissen, M., Baxter, R.M., Moodie, S., Beattie, J., Ghazal, P.: Pdq wizard: automated prioritization and characterization of gene and protein lists using biomedical literature. Bioinformatics **22**(16), 2055–2057 (2006)
102. Modlin, S.J., Gunasekaran, D., Zlotnicki, A.M., Elghraoui, A., Kuo, N., Chan, C.K., Valafar, F.: Resolving the hypotheticome: annotating m. tuberculosis gene function through bibliomic reconciliation and structural modeling. Preprint at https://doi. org/10.1101/358986 (2018)
103. Yang, L., Wang, B., Xia, G., Xia, Z., Xu, L.: Bibliomics-based selection of analgesics targets through google-pagerank-like algorithm. In: 2007 Second International Conference on Bio-Inspired Computing:

Theories and Applications, pp. 98–101 (2007). IEEE
104. Hettne, K.M., Weeber, M., Laine, M.L., Cate, H.t., Boyer, S., Kors, J.A., Loos, B.G.: 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), 1016–1024 (2007)
105. Miwa, M., Ohta, T., Rak, R., Rowley, A., Kell, D.B., Pyysalo, S., Ananiadou, S.: A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text. Bioinformatics **29**(13), 44–52 (2013)
106. Zerva, C.: Automatic Identification of Textual Uncertainty. The University of Manchester, United Kingdom (2019)
107. Holdcroft, A.: Gender bias in research: how does it affect evidence based medicine? SAGE Publications Sage UK: London, England (2007)
108. Slawson, N.: 'Women Have Been Woefully Neglected': Does Medical Science Have a Gender Problem? http://www.theguardian.com/education/2019/dec/18/women-have-been-woefully -neglected-does-medical-science-have-a-gender-problem
109. Mastroianni, A.C., Henry, L.M., Robinson, D., Bailey, T., Faden, R.R., Little, M.O., Lyerly, A.D.: Research with pregnant women: new insights on legal decision-making. Hastings Center Report **47**(3), 38–45 (2017)
110. Meadows, M.: Pregnancy and the drug dilemma. FDA Consumer magazine **35**(3), 16–20 (2001)
111. Hanson, M., , Gluckman, P.: Early developmental conditioning of later health and disease: physiology or pathophysiology? Physiological reviews (2014)
112. Arshad, R., Sameen, A., Murtaza, M.A., Sharif, H.R., Dawood, S., Ahmed, Z., Nemat, A., Manzoor, M.F.: Impact of vitamin d on maternal and fetal health: A review. Food Science & Nutrition (2022)
113. Tous, M., Villalobos, M., Iglesias, L., Fernández-Barrés, S., Arija, V.: Vitamin d status during pregnancy and offspring outcomes: a systematic review and meta-analysis of observational studies. European journal of clinical nutrition **74**(1), 36–53 (2020)
114. Todorova, M., Gerova, D., Galunska, B.: Vitamin d deficiency during pregnancy. Scripta Scientifica Medica **54**(1), 19–28 (2022)
115. Dror, D.: Vitamin d in pregnancy. In: Handbook of Vitamin D in Human Health, pp. 670–691. Wageningen Academic Publishers, Wageningen (2013)
116. KIRCA, A.Ş.: The effect of vitamin d deficiency in pregnancy on maternal results. Recent Studies in Health Sciences, 359–366 (2019)
117. Boguslav, M.R., Hailu, N.D., Bada, M., Baumgartner, W.A., Hunter, L.E.: Concept recognition as a machine translation problem. BMC bioinformatics **22**(1), 1–39 (2021)
118. Balili, C., Segev, A., Lee, U.: Tracking and predicting the evolution of research topics in scientific literature. In: 2017 IEEE International Conference on Big Data (big Data), pp. 1694–1697 (2017). IEEE
119. Balili, C., Lee, U., Segev, A., Kim, J., Ko, M.: Termball: tracking and predicting evolution types of research topics by using knowledge structures in scholarly big data. IEEE Access **8**, 108514–108529 (2020)
120. Sutherland, W.J., Fleishman, E., Mascia, M.B., Pretty, J., Rudd, M.A.: Methods for collaboratively identifying research priorities and emerging issues in science and policy. Methods in Ecology and Evolution **2**(3), 238–247 (2011)
121. Liu, N., Shapira, P., Yue, X.: Tracking developments in artificial intelligence research: Constructing and applying a new search strategy. Scientometrics **126**(4), 3153–3192 (2021)
122. Ostrowsky, J., Arpey, M., Moore, K., Osterholm, M., Friede, M., Gordon, J., Higgins, D., Molto-Lopez, J., Seals, J., Bresee, J.: Tracking progress in universal influenza vaccine development. Current opinion in virology **40**, 28–36 (2020)
123. Dinakar, B., Boguslav, M.R., Görg, C., Dinakarpandian, D.: Semantic changepoint detection for finding potentially novel research publications. In: BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, pp. 107–118 (2020). World Scientific
124. Boyack, K.W., Börner, K.: 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), 447–461 (2003)
125. Antonio, E., Alobo, M., Bayona, M.T., Marsh, K., Norton, A.: Funding and covid-19 research priorities-are the research needs for africa being met?[version 1; peer review: awaiting (2020)
126. Benner, J.S., Morrison, M.R., Karnes, E.K., Kocot, S.L., McClellan, M.: An evaluation of recent federal spending on comparative effectiveness research: priorities, gaps, and next steps. Health Affairs **29**(10), 1768–1776 (2010)
127. Wallis, S., Cole, D.C., Gaye, O., Mmbaga, B.T., Mwapasa, V., Tagbor, H., Bates, I.: Qualitative study to develop processes and tools for the assessment and tracking of african institutions' capacity for operational health research. BMJ open **7**(9), 016660 (2017)
128. Open Access Subset. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
129. Pielke-Lombardo, H.: Knowtator-2.0: A Text Annotation Plugin for Protege 5+. https://github.com/UCDenver-ccp/Knowtator-2.0
130. Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.A.: The protégé owl plugin: An open development environment for semantic web applications. In: International Semantic Web Conference, pp. 229–243 (2004). Springer
131. Callahan, T.J.: PheKnowLator Human Disease Knowledge Graphs – Class-based Knowledge Model with Inverse Relations and OWL-NETS Abstraction. doi:10.5281/zenodo.7029922. https://doi.org/10.5281/zenodo.7029922
132. Callahan, T.J.: 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. (2021). doi:10.5281/zenodo.5893789. https://doi.org/10.5281/zenodo.5893789
133. Sherman, B.T., Hao, M., Qiu, J., Jiao, X., Baseler, M.W., Lane, H.C., Imamichi, T., Chang, W.: David: a web

server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res **10** (2022)
134. Fiji User Guide. http://bficores.colorado.edu/biofrontiers-it/cluster-computing/fiji/fiji-user-guide-1
135. Hripcsak, G., Rothschild, A.S.: Agreement, the f-measure, and reliability in information retrieval. Journal of the American medical informatics association **12**(3), 296–298 (2005)
136. Dalianis, H.: Evaluation metrics and evaluation. In: Clinical Text Mining, pp. 45–53. Springer, Cham, Switzerland (2018)
137. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
138. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
139. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
140. Cohen, K.B., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M., Hunter, L.E.: The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain. In: Handbook of Linguistic Annotation, pp. 1379–1394. Springer, Dordrecht (2017)
141. Baumgartner Jr, W.A., Bada, M., Pyysalo, S., Ciosici, M.R., Hailu, N., Pielke-Lombardo, H., Regan, M., Hunter, L.: Craft shared tasks 2019 overview—integrated structure, semantics, and coreference. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pp. 174–184 (2019)
142. Lee, S., Lee, D.K.: What is the proper way to apply the multiple comparison test? Korean journal of anesthesiology **71**(5), 353–360 (2018)
143. Pustejovsky, J., Stubbs, A.: Natural Language Annotation for Machine Learning: A Guide to Corpus-building for Applications. O'Reilly Media, Inc., Sebastopol, CA (2012)
144. Costanzo, P., Santini, A., Fattore, L., Novellino, E., Ritieni, A.: Toxicity of aflatoxin b1 towards the vitamin d receptor (vdr). Food and Chemical Toxicology **76**, 77–79 (2015)
145. Sanchez-Hernandez, D., Anderson, G.H., Poon, A.N., Pannia, E., Cho, C.E., Huot, P.S., Kubant, R.: Maternal fat-soluble vitamins, brain development, and regulation of feeding behavior: an overview of research. Nutrition Research **36**(10), 1045–1054 (2016)
146. Ojo, O., Weldon, S.M., Thompson, T., Vargo, E.J.: 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), 1716 (2019)
147. Sherman, B.T., Huang, D.W., Tan, Q., Guo, Y., Bour, S., Liu, D., Stephens, R., Baseler, M.W., Lane, H.C., Lempicki, R.A.: David knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC bioinformatics **8**(1), 1–11 (2007)
148. Hanieh, S., Ha, T.T., Simpson, J.A., Thuy, T.T., Khuong, N.C., Thoang, D.D., Tran, T.D., Tuan, T., Fisher, J., Biggs, B.-A.: Maternal vitamin d status and infant outcomes in rural vietnam: a prospective cohort study. PloS one **9**(6), 99005 (2014)
149. Angelidou, A., Asadi, S., Alysandratos, K.-D., Karagkouni, A., Kourembanas, S., Theoharides, T.C.: Perinatal stress, brain inflammation and risk of autism-review and proposal. BMC pediatrics **12**(1), 1–12 (2012)
150. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
151. Wei, C.-H., Allot, A., Leaman, R., Lu, Z.: Pubtator central: automated concept annotation for biomedical full text articles. Nucleic acids research **47**(W1), 587–593 (2019)
152. of Medicine, N.L.: PubMed User Guide: Publication Types. https://pubmed.ncbi.nlm.nih.gov/help/publication-types

Additional Files

Supplementary File on Corpus Information

Details on the corpus specifically.

In annotating, K.J.S. and M.R.B. decided that one article (PMC4869271) was too difficult to annotate for this task because:

- The audience seemed to be public health workers whereas this task was more focused on researchers via scientific articles. This may be a true implementation paper.
- The article was full of quotes, making it very difficult to determine what was ignorance. Many of the quotes were also informal, meaning that it was unclear what was a quote or not. This further made the article difficult to follow.
- The article was most likely more of a technical report, public health education, government, and/or an interview rather than a scientific article based on the publication types from the NIH [152]. There did not appear to be much basic research in it.
- It was rather difficult to determine what was ignorance in the study itself versus the implementation of the policies being presented. The majority of the article seemed factual because it was mainly quotes from the public health workers.
- Neither K.J.S. nor M.R.B. were confident in their annotations of the article.

Thus, this article was excluded from the final corpus.

Supplementary File on Ignorance Classification

More detailed Tables and Figures on Ignorance classification at both the sentence- and word-levels.

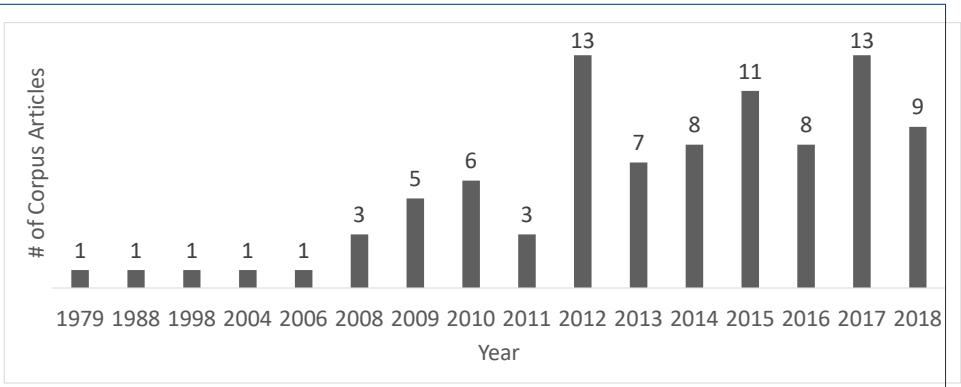


Figure 14: Article date distribution for the ignorance corpus (1979-2018).

Table 9: Annotation Statistics Per Ignorance Category: Total number of lexical cue annotations in all articles and statistics per ignorance category. SCOPE is the number of sentences that contain at least one ignorance lexical cue. Note that all categories except for ALL CATEGORIES and SCOPE (have 1) have zero minimum number of annotations. Note max = maximum.

Category	# total annotations	average # annotations per article	median # annotations per article	max # annotations
answered question	619	6.8	7	23
unknown/novel	328	3.6	2	20
explicit inquiry	177	1.95	1	27
incompletely understood	6471	71.11	67	329
indefinite relationship	3908	42.95	36	165
largely understood	1695	18.63	15	108
anomalous/curious	1012	11.12	9	39
alternative/controversy	1519	16.69	12	94
difficult task	320	3.52	2	25
problem/complication	532	5.85	3	29
future work	948	10.42	6	103
future prediction	366	4.02	2	52
important consideration	1786	19.63	9	199
ALL CATEGORIES	19681	216.27	200	1021
SCOPE	6700	73.63	69	262

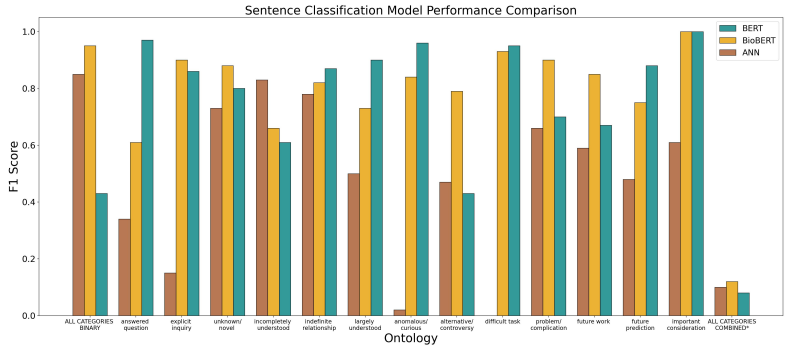


Figure 15: Sentence Classification summary: A bar plot summary of test F1 scores for sentence classification. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Table 10: Unique Annotation Statistics Per Ignorance Category: Total number of unique lexical cue annotations in all articles and statistics per ignorance category. We did not include SCOPE because the number of sentences is the same. Note that all categories except for ALL CATEGORIES (has 1) have zero minimum number of unique annotations. Note that the SCOPE is not in the unique table because we only capture the scope one time no matter how many lexical cue annotations occur within it. Note max = maximum.

Category	# total unique annotations	% unique annotations	avg # unique annotations per article	median # unique annotations per article	max # unique annotations
answered question	49	8%	4.32	5	14
unknown/novel	69	21%	2.53	2	9
explicit inquiry	16	9%	1.21	1	8
incompletely understood	368	6%	32.66	35	83
indefinite relationship	130	3%	15	15	45
largely understood	106	6%	7.59	7	25
anomalous/curious	63	6%	4.6	5	14
alternative/controversy	100	7%	7.57	7	24
difficult task	40	13%	2.3	2	11
problem/complication	41	8%	2.91	2	12
future work	105	11%	5.34	4	21
future prediction	14	4%	1.84	1	6
important consideration	101	6%	6.73	5	37
ALL CATEGORIES	1202	6%	94.59	104	273

Table 11: Annotation Counts per Section: Total number of annotations by section in all articles with section delineation and statistics per article. Note that every article contained a title and none of the titles had any ignorance annotations. Note avg = average, min = minimum, and max = maximum.

Section	# total articles	# total annotations	avg # annotations per article	median # annotations per article	min # annotations	max # annotations
abstract	71	123	1.73	1	0	29
introduction	86	2649	30.8	16	0	571
methods	61	2880	47.21	38	1	367
results	54	927	17.17	6.5	0	287
discussion	57	4906	86.07	55	3	330
conclusion	41	8196	199.9	150	2	990

Table 12: ANN sentence classification: Note that one sentence can map to more than one category and so they will not add up to the total binary. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.99	3354	0.85	2005
answered question	0.66	479	0.34	168
explicit inquiry	0.95	524	0.15	92
unknown/novel	0.87	347	0.73	63
incompletely understood	0.9	1582	0.83	225
indefinite relationship	0.86	1965	0.78	1072
largely understood	0.86	523	0.5	312
anomalous/curious	0.92	479	0.02	149
alternative/controversy	0.97	942	0.47	441
difficult task	0.88	342	0.0	93
problem/complication	0.68	324	0.66	202
future work	0.85	869	0.59	195
future prediction	0.93	178	0.48	55
important consideration	0.75	735	0.61	491
ALL CATEGORIES COMBINED*	0.17	3354	0.08	2005

Table 13: BERT sentence classification: Note that one sentence can map to more than one category and so they will not add up to the total binary. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.83	3354	0.43	2005
answered question	0.99	479	0.97	168
explicit inquiry	0.87	524	0.86	92
unknown/novel	0.89	347	0.8	63
incompletely understood	0.81	1582	0.61	225
indefinite relationship	0.91	1965	0.87	1072
largely understood	0.96	523	0.9	312
anomalous/curious	0.93	479	0.96	149
alternative/controversy	0.83	942	0.43	441
difficult task	0.92	342	0.95	93
problem/complication	0.92	324	0.7	202
future work	0.83	869	0.67	195
future prediction	0.9	178	0.88	55
important consideration	>0.99	735	>0.99	491
ALL CATEGORIES COMBINED*	0.23	3354	0.10	2005

Table 14: BioBERT sentence classification: Note that one sentence can map to more than one category and so they will not add up to the total binary.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.99	3354	0.95	2005
answered question	0.93	479	0.61	168
explicit inquiry	0.99	524	0.9	92
unknown/novel	0.88	347	0.88	63
incompletely understood	0.97	1582	0.66	225
indefinite relationship	0.88	1965	0.82	1072
largely understood	0.86	523	0.73	312
anomalous/curious	0.85	479	0.84	149
alternative/controversy	0.92	942	0.79	441
difficult task	0.91	342	0.93	93
problem/complication	0.82	324	0.9	202
future work	0.83	869	0.85	195
future prediction	0.88	178	0.75	55
important consideration	>0.99	735	>0.99	491
ALL CATEGORIES COMBINED*	0.22	3354	0.12	2005

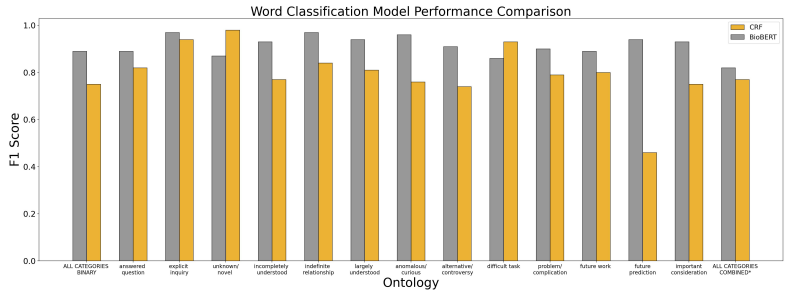


Figure 16: Word Classification summary: A bar plot summary of test F1 scores for word classification. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Table 15: CRF word classification. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.76	16748	0.75	7601
answered question	0.8	692	0.82	320
unknown/novel	0.97	401	0.98	155
explicit inquiry	0.94	136	0.94	43
incompletely understood	0.76	5764	0.77	2809
indefinite relationship	0.83	2847	0.84	1205
largely understood	0.79	1313	0.81	618
anomalous/curious	0.73	809	0.76	399
alternative/controversy	0.73	1262	0.74	598
difficult task	0.9	283	0.93	128
problem/complication	0.8	432	0.79	238
future work	0.82	991	0.8	391
future prediction	0.56	341	0.46	100
important consideration	0.75	1484	0.75	608
ALL CATEGORIES COMBINED*	0.77	13828	0.77	6239

Table 16: BioBERT word classification. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.9	16748	0.89	7601
answered question	0.92	692	0.89	320
unknown/novel	0.91	401	0.87	155
explicit inquiry	0.95	136	0.97	43
incompletely understood	0.94	5764	0.93	2809
indefinite relationship	0.98	2847	0.97	1205
largely understood	0.94	1313	0.94	618
anomalous/curious	0.91	809	0.96	399
alternative/controversy	0.9	1262	0.91	598
difficult task	0.9	283	0.86	128
problem/complication	0.88	432	0.9	238
future work	0.91	991	0.89	391
future prediction	0.97	341	0.94	100
important consideration	0.96	1484	0.93	608
ALL CATEGORIES COMBINED*	0.82	13828	0.82	6239