Published in final edited form as:

J Biomed Inform. 2022 July ; 131: 104120. doi:10.1016/j.jbi.2022.104120.

Discovering novel drug-supplement interactions using SuppKG generated from the biomedical literature

Dalton Schutte[a,b], Jake Vasilakes[a,b,e], Anu Bompelli[b], Yuqi Zhou[a,b], Marcelo Fiszman[f], Hua Xu[g], Halil Kilicoglu[d], Jeffrey R. Bishop[c], Terrence Adam[a,b], Rui Zhang[a,b,*]

[a]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA

[b]Department of Pharmaceutical Care & Health Systems, University of Minnesota, Minneapolis, MN, USA

[c]Department of Experimental and Clinical Pharmacy, University of Minnesota, Minneapolis, MN, USA

[d]School of Information Sciences, University of Illinois, Champaign, IL, USA

[e]National Centre for Text Mining, School of Computer Science, The University of Manchester, Manchester, United Kingdom

[f]NITES - Núcleo de Inovação e Tecnologia Em Saúde, Pontifical Catholic University of Rio de Janeiro, Brazil

[g]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Objective: Develop a novel methodology to create a comprehensive knowledge graph (SuppKG) to represent a domain with limited coverage in the Unified Medical Language System (UMLS), specifically dietary supplement (DS) information for discovering drug-supplement interactions

[*]Corresponding author at: 8-100 Phillips-Wangensteen Building, 516 Delaware Street SE, Minneapolis, MN 55455, University of Minnesota. zhan1386@umn.edu (R. Zhang).

Author Contributions.

DS, JV and RZ conceived the study design and wrote the initial draft of the manuscript. DS and JV carried out the experiments and produced the original draft of the manuscript. AB and YZ annotated the PubMed abstracts. TA and JB conducted the qualitative evaluations of literature-based discovery. HK, MF and HX were consulted on specifics of the experimental design. All authors contributed to the production of the final manuscript.

Declaration of Competing Interest

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

Competing Interests.

Dr. Hua Xu and The University of Texas Health Science Center at Houston have financial related research interest in Melax Technologies Inc.

CRediT authorship contribution statement

Dalton Schutte: Methodology, Software, Writing – original draft. **Jake Vasilakes: . Anu Bompelli:** Data curation, Validation. **Yuqi Zhou:** Data curation, Validation. **Marcelo Fiszman:** Validation, Writing – review & editing. **Hua Xu:** Writing – review & editing. **Halil Kilicoglu:** Methodology, Data curation, Writing – review & editing. **Jeffrey R. Bishop:** Validation, Writing – review & editing. **Terrence Adam:** Validation, Writing – review & editing. **Rui Zhang:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

(DSI), by leveraging biomedical natural language processing (NLP) technologies and a DS domain terminology.

Materials and Methods: We created SemRepDS (an extension of an NLP tool, SemRep), capable of extracting semantic relations from abstracts by leveraging a DS-specific terminology (iDISK) containing 28,884 DS terms not found in the UMLS. PubMed abstracts were processed using SemRepDS to generate semantic relations, which were then filtered using a PubMedBERT model to remove incorrect relations before generating SuppKG. Two discovery pathways were applied to SuppKG to identify potential DSIs, which are then compared with an existing DSI database and also evaluated by medical professionals for mechanistic plausibility.

Results: SemRepDS returned 158.5% more DS entities and 206.9% more DS relations than SemRep. The fine-tuned PubMedBERT model (significantly outperformed other machine learning and BERT models) obtained an F1 score of 0.8605 and removed 43.86% of semantic relations, improving the precision of the relations by 26.4% over pre-filtering. SuppKG consists of 56,635 nodes and 595,222 directed edges with 2,928 DS-specific nodes and 164,738 edges. Manual review of findings identified 182 of 250 (72.8%) proposed DS-Gene-Drug and 77 of 100 (77%) proposed DS-Gene1-Function-Gene2-Drug pathways to be mechanistically plausible.

Discussion: With added DS terminology to the UMLS, SemRepDS has the capability to find more DS-specific semantic relationships from PubMed than SemRep. The utility of the resulting SuppKG was demonstrated using discovery patterns to find novel DSIs.

Conclusion: For the domain with limited coverage in the traditional terminology (e.g., UMLS), we demonstrated an approach to leverage domain terminology and improve existing NLP tools to generate a more comprehensive knowledge graph for the downstream task. Even this study focuses on DSI, the method may be adapted to other domains.

Keywords

Dietary supplements; Drug supplement interactions; Knowledge discovery; Natural language processing; Text mining

1. Introduction

The 1994 Dietary Supplement Health and Education Act (DSHEA) defines a dietary supplement (DS) as “a product intended to supplement the diet that contains one or more of the following dietary ingredients: a vitamin, a mineral, an herb or other botanical, an amino acid, a dietary substance to supplement the diet or a concentrate, metabolite, constituent, extract, or combination of any ingredients thereof” [1]. Data from the 2017–2018 National Health and Nutrition Examination Survey (NHANES) found that 57.6% of U.S. adults aged over 20 used some sort of DS. The data also shows that across the age groups for both men and women, prevalence of DS use increases with age [2].

The DSHEA also classifies DSs as a category of food and thus do not require pre-market approval by the FDA like pharmaceutical drugs. Furthermore, many DS are non-patentable so there is less incentive to conduct research to clarify their interactions with pharmaceutical drugs [3]. This leads to the effects of many DS not being adequately understood. With this comes the risk that an individual may experience an interaction between a DS they are taking and a pharmaceutical substance. Some systematic reviews have investigated the literature for interactions between drugs and DSs [4-9]. However, they note the limited literature studying drug-supplement interactions (DSI), the quality of the reviewed studies, the sample size in the reviewed studies, etc. as being limitations. The literature suggests that there is still insufficient knowledge surrounding DS and their potential interactions with other substances, adverse or otherwise.

A couple challenges were identified to better understand DSIs. First, while there is some research regarding DS in general, there is limited research into DS interactions with pharmaceuticals, although there are many studies on DS. So, we will use literature-based discovery methods to discover DSIs. The absence of a dataset suited towards our purposes and the infeasibility of creating an annotated dataset is a major obstacle to using state of the art machine learning models for named entity recognition (NER) and relation extraction tasks. Second, the UMLS was found to have a lack of representation of DS terminology [10]. While there exist various biomedical tools, such as MetaMap and SemRep (see details in section 2.1), which are benchmark tools to extract entities and relations from text, they rely on the UMLS to perform these tasks. As such, a lack of representation of DS in the UMLS means that extracted DS entities and relations involving DS will also be under-represented in the output of these tools. This under-representation affects the ability of methods, such as literature-based discovery, to find potentially meaningful information regarding DS and DSIs.

To address these challenges, we propose an approach for generating a comprehensive DS knowledge graph (SuppKG) from the literature by leveraging a DS specific terminology (iDISK) and improving an NLP tool and later demonstrate its usage for a literature-based discovery (LBD) task on DSI discovery. This approach can potentially be adapted for other under-represented domains. Specifically, in this study, we contribute in the following aspects. First, we created an extended version of SemRep, called SemRepDS, by integrating DS terms from a DS specific knowledge base with the UMLS to enable better recognition of DS terms in biomedical literature. We use SemRep for NER and relation extraction because SemRep does not require large amounts of annotated training data and showed strong performance in a recent evaluation [11]. Relation extraction is an NLP task that attempts to identify a subject, object, and the relationship between them from text. Applying relation extraction to a large corpus of text, such as PubMed abstracts, allows for relationships to be found where they were previously unknown. SemRep is such an existing tool to extract relationships from literature (details in section 2.1). However, SemRep is limited when it comes to DS because SemRep relies on the UMLS, which has been found to have a lack of DS representation [10]. Therefore, we will extend the UMLS using iDISK to expand SemRep's coverage of DS entities and semantic relations. This method allows for expanding SemRep's coverage while not requiring extensive, time-consuming annotation efforts. Second, we created a comprehensive DS knowledge graph, SuppKG, containing all identified relationships involving DS concepts by the expanded SemRepDS. To improve the quality of the semantic relations in SuppKG, we developed an accuracy classifier to identify which semantic relations are incorrect. Finally, we demonstrate the usage of SuppKG on the LBD task of discovering novel DSIs using discovery patterns. Of the 350 potential DSIs we discovered, 325 were not found in well-known knowledge bases. Through evaluation by domain experts, we identified 76 DSIs are mechanistically plausible.

2. Background and related work

2.1. SemRep and SemMedDB

MetaMap [12,13] links entity mentions in biomedical text to concepts in the UMLS Metathesaurus. It returns a mapping decision for each identified phrase in a body of text. SemRep is a rule-based tool for extracting semantic relations from biomedical text using MetaMap to map text to concepts in the UMLS [14]. A semantic relation is comprised of three parts, the subject concept, a predicate, the object concept. An example of a semantic relation is:

“Effects of Asian sand dust, Arizona sand dust, amorphous silica and aluminum oxide on allergic inflammation in the murine lung.”

- Subject: DC0002374 - Alumina
- Predicate: EFFECTS
- Object: C0021375 – Inflammation, allergic
- Relation: Alumina EFFECTS Inflammation, Allergic

In the example above, the subject and object are identified by MetaMap in the UMLS Metathesaurus. These concepts are used by SemRep along with a set of trigger rules and limited by the Semantic Network to produce semantic relations. Henceforth, we will use “SemRep identified X DS” with the implication being that MetaMap did the entity extraction, but the results from the MetaMap extraction were included in the SemRep output. The DC0002374 is a DCUI (a letter D was added before a concept unifier identifier [CUI] to represent it as a DS concept) used to identify DS concepts in iDISK. See our prior work [15] for additional details.

Applying SemRep to the entire collection of PubMed citations produces a large database of relations, entities, and source sentences called SemMedDB [16]. While an extension of SemMedDB has been produced before (SemMedDB UTH [17]) this method involved extending SemRep by modifying the Metathesaurus directly to undo the suppression of the drugs in the National Cancer Institute Thesaurus. Prior work has demonstrated that SemRep can be extended to the domains of disaster information management [18] and medical informatics [19] using various knowledge engineering techniques.

2.2. Literature based discovery

Literature Based Discovery (LBD) is a method for generating hypotheses by finding implicit relationships in the research literature [20]. Two categories of LBD are open and closed LBD. Open LBD involves providing a term and the task is to find connections between two concepts that may not be directly related. Closed LBD takes two terms and finds concepts shared between them [21]. Discovery patterns exploit the explicit relationships between

UMLS concepts by imposing conditions on what pathways are valid, potential relationships [20].

Previous work on Drug-Drug Interactions (DDIs) discovery has used a variety of data sources and techniques. One study identified combinations of Drug-Gene relationships associated with known DDIs using Medline abstracts and a random forest classifier [22]. Another used electronic medical records (EMR) to narrow a set of potential DDIs generated from *in vitro* studies to a collection of drug pairs that were found to have higher risk ratios for myopathy [23]. SemMedDB [24] has also been used in the past to discover potential DDIs using discovery patterns based on semantic types of the UMLS concepts extracted from PubMed abstracts in combination with patient EMR data [25].

In a recent study that performed LBD for drug repurposing [26], a BERT [27] model was fine-tuned on annotated semantic relations to identify which relations were correctly implied by their source sentence. The fine-tuned BERT model was used to remove incorrect relations from SemMedDB before open LBD was performed for drug repurposing.

2.3. The integrated dietary supplement knowledge base (iDISK)

The Integrated Dietary Supplement Knowledge base (iDISK) is a data model that contains terminology of DS ingredients. The DS information was gathered from four well-trusted DS resources: Natural Medicines Comprehensive Database, Memorial Sloan Kettering Cancer Center, Dietary Supplement Label Database, and the Natural Health Products Database and compiled into a data model with the assistance of domain experts [15].

Like the UMLS data model, the core iDISK data elements are atoms, concepts, concept attributes, relationships, and relationship attributes. Included in the iDISK data model are links to other controlled vocabularies to allow for wider use and integration with existing biomedical knowledge representations such as the UMLS. iDISK contains 144,654 unique concepts, of which 4,208 are concepts for DS ingredients and 137,568 concepts for DS products.

2.4. Drug supplement interactions

There are limited studies that use LBD to discover DSIs. Our group previously discovered interactions between cancer drugs and DS through LBD [28]. We leveraged SemMedDB to identify both known and unknown DSIs through expert validation. Recently, the Allen AI Institute conducted DSI identification using articles from Semantic Scholar and RoBERTa [29,30]. Supp.AI mined the literature to find published DSIs. The Supp.AI lexicon of DS is based on data mined from Medline articles that have UMLS CUIs. Our prior work has shown that the UMLS coverage for DS terminology and synonyms is incomplete [10] which imposes an inherent limitation on the terminology Supp.AI can use. Comparing terms available in iDISK [15] and in Supp.AI, we found that 87.3% of the concepts in Supp.AI were contained in iDISK but only 43.5% of the concepts in iDISK were contained in Supp.AI. In addition, Supp.AI is designed to extract DSI mentioned in the literature but not to discover novel DSI, which is the focus of this study.

3. Materials and methods

Fig. 1 provides a visual summary of the entire process. First, we used the terminology and relations contained in iDISK to extend the domain of MetaMap in SemRep, which resulted in the SemRepDS system. Second, we collected abstracts from PubMed that were processed by SemRepDS to extract semantic relations to construct a comprehensive DS knowledge graph, SuppKG. Finally, we demonstrate the usage of SuppKG using discovery patterns to discover novel DSIs.

3.1. Extend SemRep

The first phase in our process was creating SemRepDS. We have two steps described as follows.

3.1.1. Integrating iDISK into the UMLS Metathesaurus—Since SemRep relies on the UMLS Metathesaurus and the Metathesaurus' coverage of DS is limited [10], we added iDISK terminology to the UMLS Metathesaurus. To achieve this, the MetaMap Datafile Builder[1] pipeline was used to generate the necessary data files required for MetaMap [12,13], which were then mapped to the DS extended Metathesaurus. The extended Metathesaurus was then linked to SemRep for relation extraction, henceforth, SemRepDS. The users can choose to use the 2006AA or 2018AA UMLS versions for SemRep, and we chose to use the 2006AA version since SemRep is optimized for this version and concept ambiguity that is present in later UMLS versions [11]. Additionally, a recent evaluation of SemRep used the 2006AA version and obtained strong performance on various corpora [11].

We extend the UMLS 2006AA Metathesaurus with DS ingredient concepts from iDISK as follows. For each iDISK concept with no existing links to any UMLS concept, we created a new concept entry in the UMLS MRCONSO.RRF file. This involved 1) defining a new concept unique identifier (CUI) for the DS concept, and 2) writing all the concept's atoms to MRCONSO.RRF. The UMLS defines preferred terms for each concept, which specify the canonical name for the concept, and which are determined according to a source vocabulary ranking defined in MRRANK.RRF. iDISK also specifies preferred terms for each concept according to a source ranking, and we designated atoms as "preferred" accordingly for all new DS concepts added to the UMLS. In our prior study [10], we found the overlap between iDISK terms and UMLS terms. To address this, concepts were merged wherever synonymy was found to reduce the likelihood of duplicate concepts. Adding iDISK concepts that *do* have existing links to the UMLS is a similar process, but we used the CUI of the linked UMLS concept instead of creating a new one, and we kept the existing UMLS preferred term rather than using the iDISK preferred term. For all iDISK concepts added, we ensured that they were assigned the "Pharmacologic Substance (phsu)" semantic type, as this semantic type is used by SemRep to determine which concepts are potential candidates for our target semantic relations, INTERACTS_WITH, INHIBITS, AUGMENTS, etc. The semantic types for each concept were written to MRSTY.RRF. We also added entries for the iDISK source vocabularies to MRSAB.RRF and MRRANK.RRF. Finally, the RRF files

[1] https://metamap.nlm.nih.gov/DataFileBuilder.shtml.

were fed into the MetaMap Data File Builder, which indexed the files and generated a concept database that can be used by MetaMap.

3.1.2. SemRepDS—

The files resulting from running the MetaMap Data File Builder, as described above, were then used to replace the files that come packaged with SemRep. This is done by replacing the corresponding files in the SemRep installation. By using the new files, SemRep will have access to the standard UMLS as well as the specialized DS vocabulary contained within iDISK. We refer to SemRep using these expanded files as SemRepDS. We further evaluate the SemRepDS in section 3.2.3.

3.2. Build SuppKG

The next phase was to use SemRepDS and a collection of abstracts from articles on PubMed to extract semantic relations that were used to build a specialized knowledge graph, SuppKG. This required us to, first, collect PubMed abstracts for SemRepDS to extract semantic relations from, second, filter out potentially incorrect semantic relations, third, evaluate the SemRepDS output pre- and post- filtering, and fourth, load the filtered output into a graph data structure.

3.2.1. Gather and process PubMed abstracts—

To build a knowledge graph containing DS information, we queried PubMed for abstracts using the Entrez API. Due to resource constraints, we could not simply use all of the available PubMed abstracts. Instead, we used the terms contained in iDISK as search terms to collect PMIDs. We restricted results to English abstracts and used the “Human Subjects” filter in PubMed to restrict the results further to studies conducted in humans.

After the queries using all of the terms in iDISK were run, we had collected 608,725 unique PMIDs. We used the Entrez API to collect the abstracts associated with each PMID. Due to requirements for MetaMap, all abstracts were pre-processed to remove any non-ASCII characters. Where necessary, Greek letters were replaced with their English written form (i.e., β was replaced with “beta”). The abstracts were batched into multiple files to allow for data parallelization.

To compare the ability of SemRepDS for extracting DS entities and semantic relations, we processed the abstracts separately using both base SemRep v1.8 and SemRepDS (mentioned in 3.1.2). By comparing the two, we hoped to show that more information related to DS can be found and used to create a more comprehensive knowledge graph to discover DSIs.

3.2.2. Filtering and ranking semantic relations with diverse BERT models—

To improve the quality of the final knowledge graph, relations were filtered for correctness using a BERT [27] model that had been fine-tuned for the binary classification task using 6000 annotated relations (inter-annotator agreement = 0.842) [31], developed previously [26]. The annotated relations were sampled from the December 2016 release of SemMedDB and were annotated by two health informaticians for use as a gold standard. The relations are split equally between substance interactions and clinical medicine relations.

To further include semantic relations with DS mentions, we randomly collected 300 abstracts containing 492 relations with DS mentions for annotation by two informatics graduate students with backgrounds in pharmacy and pharmaceutical sciences (AB and YZ). Each relation was labeled as correct or incorrect, with a “correct” relation indicating that the extracted relation was verified as included in the source sentence. Among the combined annotated dataset, 67.02% (4,251/6,492) were labeled as correct relations.

The combined dataset was split 70/20/10 for training, development, and test sets, respectively. The train and development sets were sampled so that they were balanced evenly between correct and incorrect labelled relations. The test set was sampled to match the original distribution of correct/incorrect relations.

We chose six BERT variants and traditional machine learning models to evaluate for the accuracy classifier. This included the base uncased BERT to use as a baseline and five biomedical domain-specific BERT models that had attained SotA results on various biomedical NLP tasks: PubMedBERT abstracts only, PubMedBERT abstracts and full text [32], BioBERT [33], BlueBERT [34], and BioClinicalBERT [35]. We also chose to use traditional machine learning models (i.e., logistic regression, random forest, gradient boosting, and support vector machine) to compare against the BERT models.

All models were evaluated using permutation testing [36]. Fifty random shuffles of the labels were generated to use for the permutation testing. We used permutation testing since it approximates variations in the test data and provides a better sense of generalization to new data.

Hyperparameters for training the BERT models were: five epochs, learning rate of 0.0003, weight decay of 0.1, gradient clipping was used with a max of 1, batch size was 16, and 200 warmup steps were used at the beginning of each training cycle. Optimization was performed using Adam [37] with decoupled weight decay and a cosine learning rate decay.

Based on the results from both the permutation testing and multiple training cycles, we found the PubMedBERT abstract only model to obtain the highest F_1 score in both tests. The PubMedBERT model was then finetuned and used as the accuracy classifier for the semantic relations.

3.2.3. Evaluating SemRepDS—

We evaluated the entity extraction and the semantic relation extraction for SemRep and SemRepDS. The entity extraction, or named entity recognition (NER), was evaluated using a previously labeled dataset [10].

To evaluate the accuracy of the semantic relation extraction, we used SemRepDS to process the same 300 abstracts (described in section 3.2.2) followed by human evaluation to identify any potential systematics errors. Since the relations from the 300 abstracts were used to train the PubMedBERT accuracy classifier, we also sampled 50 additional abstracts containing 224 relations before filtering. The same informatics students (AB and YZ) assessed the correctness of these relations based on the source sentences. We then report the precision of semantic relations before and after filtering.

3.2.4. Creating SemMedDB-DS and SuppKG—

We used the fine-tuned PubMedBERT model to filter the output from SemRepDS (see section 3.2.1). A semantic relation was removed if the PubMEDBERT model returned a likelihood score below 0.5. The semantic relations that remained after filtering we refer to as SemMedDB-DS.

This collection of semantic relations can be represented using a graph-structured data model where the subjects and objects form the set of graph nodes, *N*, and the predicates form the set of directed edges, *E*. We thus created a knowledge graph, called SuppKG, using only the filtered set of relations from SemRepDS. Equivalently, SuppKG was created by loading the semantic relations contained in SemMedDB-DS into a graph data structure. We will make both SemMedDB-DS and SuppKG available on our GitHub (https://github.com/zhang-informatics/SemRep_DS/tree/main/SuppKG).

The potential usage for SuppKG includes discovering potential DSIs, repurposing DS for treatments, discovering adverse events associated with DS, providing interpretation for mechanism of action for DS with diseases, etc. In this study, we will demonstrate this usage for discovering potential DSIs.

3.3. Use case for the SuppKG: LBD for DSI discovery:

The final phase of our process was to demonstrate the utility of SuppKG, and thus the utility of extending SemRep with a specialized vocabulary. We did this by performing literature-based discovery (LBD) using discovery patterns on SuppKG. The pathways found using the discovery patterns were then evaluated by experts with pharmaceutical and clinical backgrounds for mechanistic plausibility.

3.3.1. Discovery of DS-Drug interactions (DSI)—

A discovery pattern is a form of link prediction that uses a template, or pattern, of a series of concepts that suggests a direct relationship between the initial and terminal concepts [20]. The idea being that a pattern defined by experts, when applied to a knowledge graph, should return a link between the initial and terminal concepts where a direct link was not previously known to exist. In the context of our work, a discovery pattern is a series of concepts found in the collection of concepts identified by SemRepDS.

Two discovery patterns were used for interaction discovery: DS-Gene-Drug (DsGD) and DS-GeneA-Biological_Function-GeneB-Drug (DsGFGD). We used these discovery patterns since they were found to have produced meaningful drug-drug interactions similar to our prior study [25]. The DsGD pathway means that the pattern would return potential interactions between a DS concept and a pharmaceutical drug concept that both acted on a shared gene concept. To ensure the novelty of the found pathways, we filtered out any that had direct links between the DS and drug concepts in SuppKG.

For each pathway, the interactions were ranked in descending order based on the sum of each semantic relations’ accuracy confidence scores assigned by the classifier.

3.3.2. Evaluation of discovered DSIs—

The evaluation of the top 250 DsGD DSI pathways and 100 DsGFGD DSI pathways was conducted by two of the authors (TA and JB), who are pharmacists with pharmaceutical sciences and clinical backgrounds. A protocol was developed to operationalize a review and evaluation process and ensure similar resources were used by both experts. A sample size of 350 pathways was used in the mechanistic evaluation. Based on Yamane's sample size formula [44], which assumes a 95% confidence interval and maximum variance, the resulting level of precision is 0.0534.

The first step in evaluating a pathway was to identify all terms. A term is a particular instance of a concept in the UMLS, for example, "cardiovascular stroke" and "heart attack" are both specific instances of the concept "myocardial infarction". This was accomplished by confirming that the text-to-concept mapping was correct then checking the UMLS, PubMed, and, as a final effort, an internet search to identify the terms. The term to concept mapping required verification since MetaMap does not always perfectly map terms to the proper concept.

The second step was to evaluate the relationships themselves by confirming the relations were correctly extracted from the abstract, checking the associated paper(s), and, finally, an internet search. If after these steps, they felt that the DSI suggested by the pathway was correctly extracted from the associated abstracts and if each relation in the pathway represented a logical biochemical/microbiological connection, they rated the DSI as 'plausible'. Otherwise, the DSI was determined to be 'implausible'. We then report the percentage of plausible DSIs and discuss some examples.

As an example, for the process above, consider the sentence and the extracted semantic relation below.

"While calcium has been shown to reduce the risk of pre-eclampsia and maternal mortality, calcium, phosphorus, potassium, magnesium and manganese can have negative impacts on organoleptic properties, so many products tested have not included these nutrients or have done so in a limited way."

Semantic Relation: manganese prevents pre-eclampsia.

SemRepDS correctly identified "manganese" and "pre-eclampsia" from the text, so the first step does not rule the pathway out as being incorrect because we have proper term to concept mappings. In the second step, however, we see that the relation was not correctly extracted. Due to the structure of the sentence, most likely, SemRepDS erroneously determined that manganese prevents pre-eclampsia in the preceding clause. However, manganese is being discussed as having negative impacts on organoleptic properties, thus rendering the semantic relation false. So, if this semantic relation were part of a proposed DSI pathway, the entire pathway would be deemed "implausible" regardless of the plausibility of any other semantic relations in the proposed pathway. Not all the semantic relations are as straightforward as the example above and many required our experts to review the source abstract or do additional research to confirm a particular relation.

To check if our DSI approach identified known DSIs, we compared our list against the Natural Medicine Interaction Checker database [38]. Natural medicine is expert-curated monographs for natural products, which include drug-supplement interactions.

4. Results

4.1. Comparison between SemRep and SemRepDS

Each semantic relation was paired with its source sentence in a database for comparison and further processing. In this study, Table 1 demonstrates that SemRepDS increased the number of DS entities identified by 158.52% after adding the iDISK terminology with UMLS. The expanded DS entities further improved the identification of additional 148,308 (206.93%) relations with at least one DS entity.

The gold standard evaluation set we used [10] contained only DS entities, so non-DS entities extracted by SemRepDS were not counted in the evaluation. Note that SemRepDS use the entities as the output of the MetaMap with the updated terms (details in section 3.1.1). The performance of SemRepDS on DS NER task is comparable with other non-DS entities. For an understanding of SemRep’s performance for NER on non-DS entities, please refer to a prior evaluation paper on SemRep [39] and Table 2 below.

4.2. Evaluation of semantic relations generated by SemRepDS

Among the 300 abstracts included based on random sampling, the precision of the pre-filtering SemRepDS output was found to be 0.67. This is comparable to a recent evaluation of SemRep that found a precision of 0.69 [11]. We compared several machine learning-based and BERT-based models for classifying the correctness of semantic predications (detailed results shown in Supplementary Table 1). PubMed BERT (the best performing classifier) was then used to filter generated semantic relations down to 2,710,240 (59.94%) For the sample of 50 random abstracts, before filtering the precision was 0.72 and after filtering increased to 0.91.

4.3. Statistics of DS knowledge graph - SuppKG

Our combined contains graph contains 130,763 nodes with 1,434,007 directed edges. SuppKG alone contains 56,635 nodes with 595,222 directed edges. SuppKG will be made available as a NetworkX graph object stored in a pickled file and a json file[2].(See Table 3).

Of the 2.7 million semantic relations, the most common predicates that remained after filtering the SemRepDS output with PubMedBERT were TREATS and COEXISTS_WITH (19.4% and 18.9%, respectively). The primary predicates of interest for our discovery pathways, TREATS, INTERACTS_WITH, INHIBITS, and STIMULATES, comprised a total of 30.92% of all predicates in the filtered semantic relationships.

4.4. Comparison to existing knowledge graphs

See Table 4.

[2] https://github.com/zhang-informatics/SemRep_DS/tree/main/SuppKG.

4.5. Mechanistic evaluation

The expert review found that 72.8% (182/250) of the DsGD relations and 77.0% (77/100) of the DsGFGD relations were mechanistically correct. The mechanistic correctness of a relation does not necessarily imply clinical utility and are considered ‘clinically plausible’.

4.6. Comparison of existing DSI knowledge base

Among 100 DSI list, five DSIs found with the DGD pathway were found in the Natural Medicines [39] database and none of the DSIs found with the DGFGD pathway were found in the database. The known interactions are:

flaxseed INHIBITS tnf receptor ligands STIMULATES plasminogen activator urokinase.

curcumin INHIBITS interleukin 1, beta STIMULATES plasminogen activator urokinase.

allicin INHIBITS inerleukin 1, beta STIMULATES plasminogen activator urokinase.

activin INHIBITS inerleukin 1, beta STIMULATES plasminogen activator urokinase.

vanadium STIMULATES atp STIMULATES vasopressin.

5. Discussion

DSIs are recognized and have potential risks for patients, however interactions between DS and pharmaceutical drugs are neither widely understood nor well identified in the biomedical literature. One of such barriers is the incomplete representation of DS terminology in the current biomedical terminologies. [10]. Until now, there was no prior study that integrated a new, specialized vocabulary (i.e., iDISK in this study) with the UMLS and used the resulting extended UMLS for relation extraction. We successfully extended the UMLS with a DS-specific vocabulary and used the resulting extended UMLS for discovering substance interactions from the literature. We also successfully used a fine-tuned PubMedBERT model to filter incorrect semantic relations. As such, we demonstrated the utility of vocabulary extension by applying discovery pathways to SuppKG to find novel DSIs, most of which were determined to be mechanistically plausible.

5.1. SemRepDS and SemMedDB-DS

We chose to extend SemRep, rather than trying to train a machine learning model for relation extraction from scratch for several reasons. First, we do not have access to large amounts of training data and manual annotation was deemed infeasible. Second, SemRep is a strong baseline (obtaining a precision of 0.9 on the CDR corpus [11]), which builds on general linguistic rules, which provide better explainability and transparency to the relation extraction process. Third, SemRep can be extended to work with additional terminologies. While SemRep generally yields lower recall, this is mitigated, to some extent, because we are processing data at the literature-scale. If a relation is missing in one article, it may be extracted in another. Furthermore, we improve the quality of the extraction semantic relations (i.e., increase the precision) by using an accuracy classifier to allow us to filter out potentially incorrect relations. Extending SemRep requires extending the

UMLS Metathesaurus, which MetaMap relies on for entity extraction. A prior evaluation of MetaMap found that the F_1 scores ranged from 0.37 to 0.67 on various corpora [39].

In our prior work, we found that iDISK can enrich UMLS to represent DS and can further improve performance on a NER task [10]. In this study, we further integrated iDISK terms with UMLS through MetaMap Data File Builder, which was demonstrated to increase the recall of recognizing DS terms and DS-related relations. This results in more relations generated by SemRepDS than SemRep. Compared with SemRep, SemRepDS can extract entities and relations with specific DS mentions, demonstrating SemRepDS' ability to extract data that would otherwise be absent. Thus, the additional relations extracted with SemRepDS are unique and not found in SemMedDB.

There are 49,571 additional relations that were extracted with SemRepDS which gives us a richer knowledge graph to work with. Furthermore, 512,201 (25.55%) of the relations extracted by SemRepDS contain at least one DS mention. Using a knowledge graph constructed from SemRepDS output contains more relations as well as relations with specialized terms that will facilitate DSI discovery.

Perhaps more importantly, we have demonstrated that specialized terminology can successfully be integrated with the UMLS Metathesaurus using the Data File Builder and the resulting extended Metathesaurus used to identify additional entities and relations. Extending the UMLS with a specialized terminology can be done with other domains beyond DS [18,19]. Additionally, the use-cases for an extended UMLS are not limited to DSI discovery. The extended UMLS was used for other types of literature-based discovery, information extraction, and other informatics tasks.

Because SemRep-DS still contains the original CUIs found in the UMLS, it is possible to integrate SemMedDB-DS and SuppKG with SemMedDB, and any graphs derived from it, by mapping nodes based on matching CUIs. Additionally, because of how DCUIs were assigned in iDISK, it is possible to map DS concepts with DCUIs to UMLS concepts with CUIs. SuppKG could be combined with SemMedDB to likely improve the results of DSI discovery since SemMedDB is formed by performing entity extraction on the entirety of Medline and would likely contain some relations that would open up new pathways not present in SuppKG alone.

5.2. Semantic relation correctness classifiers

A sample of 50 abstracts had a pre-filtering precision of 0.72 and a post-filtering precision of 0.91. This is a substantial improvement over the raw output from SemRepDS. While the resulting knowledge graph is smaller, the likelihood of any path being comprised of true relations is higher. The potential DSIs identified by our pathways are likely to be of a higher quality than if we used the same pathways on the pre-filtered graph.

It is somewhat surprising that the PubMedBERT model trained on abstracts only outperformed the one trained on abstracts and full-text articles. However, it was observed by the authors in [33] that the abstract-only out-performed the abstract + full-text model on some tasks. Since our dataset is derived from PubMed abstracts, the inclusion of full-text articles likely resulted in a shift away from the target distribution that hindered the full-text model compared to the abstract only model, which contains many of the abstracts included in our sample. An abstract contains more distilled statements regarding the exact nature of a study and its findings. Thus, the abstract-only model would have a more unadulterated representation of the relationships that SemRepDS is extracting. The PubMedBERT model used for filtering obtained an F_1 score of 0.87 which suggests that the model was reasonably successful in identifying correct relations.

5.3. Link prediction for proposed DSIs

Below are some examples of relations found with our method that were deemed to be of potential clinical interest by our experts. The examples are all from the DS-Gene-Drug pathway. (see Table 5).

5.4. Error analysis

We conducted error analysis on false positives during evaluation. Errors stem from how terms are mapped to concepts when MetaMap tries to map text to a concept in our extended UMLS. An example of a mapping error due to MetaMap we observed is ‘diet’ mapping to ‘Bill Henderson Protocol’, a proposed diet to help fight cancer. Another involved contextual differences in the meaning of a word. For example, “contracted” identified in one context based on relationships with contraction of gel media in a laboratory experiment versus contraction of blood vessels.

Another source of error is with SemRepDS. The extraction of semantic relationships by SemRepDS has a precision of only 0.67. While filtering improved this to 0.91, there is still a non-trivial proportion of incorrect relations contained in SemMedDB-DS. As such, any of the proposed pathways that contained an incorrect relation would be rendered invalid based on our evaluation criteria.

The most common types of errors found in the sample of filtered relations were either SemRepDS missing the negation of a predicate or the improper attribution of an entity to a relation in a different clause. We include some illustrative examples of pathways that were incorrect due to a semantic error or extraction error. For example, “manganese prevents pre-eclampsia” extracted from “While calcium has been shown to reduce the risk of pre-eclampsia and maternal mortality, calcium, phosphorus, potassium, magnesium and manganese can have negative impacts on organoleptic properties, so many products tested have not included these nutrients or have done so in a limited way.” Here, SemRepDS incorrectly associated manganese with “reduce the risk of” from the first clause containing “pre-eclampsia”.

5.5. Limitations and future work

One major limitation of this work is a consequence of the performance of SemRepDS for NER. The precision, recall, and F_1 scores are 0.4458, 0.4752, and 0.4600, respectively, which means there is a non-trivial amount of error that can be introduced in the information extraction step. This can propagate to downstream tasks and affect the DSIs being extracted from SuppKG. The errors will be incorrect term extractions or missing terms entirely which can result in erroneous links in SuppKG and incompleteness, respectively. We take steps to mitigate this by using a trained biomedical BERT model to filter out incorrect semantic predications. The BERT model is trained using the semantic triple and the source sentence, as such, if a term present in the triple is not in the source sentence, then there is a chance it would be scored low enough by the BERT model to not be included in SuppKG. However, we do not know the extent to which this occurs. This limitation could be partially addressed by using more sophisticated relation extraction tools or models rather than SemRepDS. Additionally, more work could be done to filter out incorrect semantic predications before building SuppKG.

Available DS concepts and terms are inherently limited by those included in sources for iDISK, which imposes a limitation on the terms SemRepDS can identify. Thus, there might be abstracts that contain DS mentions and relations but cannot be identified by SemRepDS. This poses a limitation on all downstream tasks, particularly any tasks using the final knowledge graph. SuppKG will be restricted to the concepts contained in the UMLS and iDISK and thus any LBD performed using SuppKG will have similar restrictions. As such, any potential DSIs involving DS concepts not contained in iDISK or the UMLS cannot be discovered using our discovery patterns. We acknowledge this as a limitation of this specific implementation of the work and future work could include expansion of the DS terminology by incorporating other sources to mitigate the impact of this limitation. However, even with these limitations, we still believe that the methodology at large is successful at expanding SemRep with DS concepts, that would otherwise be lacking, to create a more comprehensive knowledge graph than one made without integrating iDISK.

Another limitation is that the DSI discovery was performed on a graph generated by a subset of PubMed abstracts rather than all abstracts. This may have resulted in the unintentional exclusion of some relations and potential DSIs from our final list. This was due to computational limitations as processing the entirety of PubMed with SemRep takes around one month. There is additional work that can be done to improve the ability of SemRepDS to properly extract relations when negation is involved. We also plan to explore a larger knowledge graph that uses the entirety of SemMedBD and SuppKG.

We see that the BERT filtering significantly improved the precision of the relations, but there is still room for incorrect relations to have been included in SuppKG. This means that there are still potentially incorrect relations that were used in some of the pathways found with our patterns. The BERT filtering model substantially improved the precision of the semantic relations but there is room (e.g., increase the annotated data set, etc.) to further improve the quality of SuppKG with additional or refined filtering methods.

While the pathways we used produced novel interactions not found in the literature, there are other methods we would like to use. There are limitations to the use of rules based DSI discovery, namely that the pathways need to be decided on by experts and that the pathways are not precise since they don’t use information contained in the knowledge graph aside from semantic types, resulting in a large volume of meaningless interactions. We will explore machine learning based methods such as standard embedding models (e.g. TransE), graph neural networks and transformer-based models which are not as interpretable as discovery patterns since they only give the initial and terminal nodes rather than a full pathway [26].

The evaluation of only 50 DSIs from each of the discovery patterns is also a limitation of our study. Due to the very labor-intensive nature of the determining if a pathway is plausible (see section 3.3.2), we were limited to using only 100 DSIs. In future work, especially if considering more sophisticated learning-based approaches, a larger sample would ideally be used if there is the person-power available to review the pathways.

6. Conclusion

The UMLS Metathesaurus contains limited data specific to the growing DS domain. In this study, we demonstrate successful augmentation of SemRep with a DS specific terminology (iDISK) by using the Data File Builder to expand DS representation in the UMLS Metathesaurus. The resulting comprehensive DS knowledge graph, SuppKG, was improved by training models to remove incorrect relations, thus reducing downstream error propagation. We also demonstrated its usage by discovering several novel DSIs not found in the literature and some that have potential clinical interest. The proposed approach can be also adapted to other domains with limited coverage in the UMLS.

Acknowledgements

This work was supported by the National Institutes of Health's National Center for Complementary and Integrative Health and the Office of Dietary Supplements grant number R01AT009457 (to RZ). The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

References

[1]. Dietary Supplement Health and Education Act of 1994. 1994.

[2]. Mishra S, Stierman B, Gahche JJ, et al. , Dietary supplement use among adults: united states, 2017–2018, NCHS Data Brief (2021) 1–8.

[3]. Brown AC, An overview of herb and dietary supplement efficacy, safety and government regulations in the United States with suggested improvements. Part 1 of 5 series, Food Chem. Toxicol 107 (2017) 449–471, 10.1016/j.fct.2016.11.001. [PubMed: 27818322]

[4]. Alsanad SM, Williamson EM, Howard RL, Cancer patients at risk of herb/food supplement-drug interactions: a Systematic Review: cancer patients herb/food supplement-drug interactions, Phytother. Res 28 (12) (2014) 1749–1755, 10.1002/ptr.5213. [PubMed: 25158128]

[5]. Jalloh MA, Gregory PJ, Hein D, Risoldi Cochrane Z, Rodriguez A, Dietary supplement interactions with antiretrovirals: a systematic review, Int J STD AIDS 28 (1) (2017) 4–15, 10.1177/0956462416671087. [PubMed: 27655839]

[6]. Robien K, Oppeneer SJ, Kelly JA, Hamilton-Reeves JM, Drug-vitamin D interactions: a systematic review of the literature, Nutr. Clin. Pract 28 (2) (2013) 194–208, 10.1177/0884533612467824. [PubMed: 23307906]

[7]. Violi F, YH G. Lip P Pignatelli D Pastori, Interaction between dietary vitamin K intake and anticoagulation by vitamin K antagonists: is it really true? A Systematic Rev. Med. (Baltimore) 95 (10) (2016) e2895, 10.1097/MD.0000000000002895.

[8]. Fong SYK, Gao Q, Zuo Z, Interaction of carbamazepine with herbs, dietary supplements, and food: a systematic review, Evid. Based Complement Alternat. Med 2013 (2013) 1–15, 10.1155/2013/898261.

[9]. Romoli M, Perucca E, Sen A, Pyridoxine supplementation for levetiracetam-related neuropsychiatric adverse events: a systematic review, Epilepsy Behav. 103 (2020) 106861, 10.1016/j.yebeh.2019.106861. [PubMed: 31917143]

[10]. Vasilakes J, Bompelli A, Bishop JR, Adam TJ, Bodenreider O, Zhang R, Assessing the enrichment of dietary supplement coverage in the unified medical language system, J. Am. Med. Inform. Assoc 27 (10) (2020) 1547–1555, 10.1093/jamia/ocaa128. [PubMed: 32940692]

[11]. Kilicoglu H, Rosemblat G, Fiszman M, Shin D, Broad-coverage biomedical relation extraction with SemRep, BMC Bioinformatics 21 (1) (2020), 10.1186/s12859-020-3517-7.

[12]. Aronson AR, Lang F-M, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc 17 (3) (2010) 229–236. [PubMed: 20442139]

[13]. Aronson AR, Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program, In: Proceedings of AMIA Symposium. (2001) 17–21.

[14]. Rindflesch TC, Fiszman M, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, J. Biomed. Inform 36 (6) (2003) 462–477, 10.1016/j.jbi.2003.11.003. [PubMed: 14759819]

[15]. Rizvi RF, Vasilakes J, Adam TJ, Melton GB, Bishop JR, Bian J, Tao C, Zhang R, iDISK: the integrated dietary supplements knowledge base, J. Am. Med. Inform. Assoc 27 (4) (2020) 539–548, 10.1093/jamia/ocz216. [PubMed: 32068839]

[16]. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC, SemMedDB: a PubMed-scale repository of biomedical semantic predications, Bioinformatics 28 (23) (2012) 3158–3160. [PubMed: 23044550]

[17]. Fathiamini S, Johnson AM, Zeng J, Araya A, Holla V, Bailey AM, Litzenburger BC, Sanchez NS, Khotskaya Y, Xu H, Meric-Bernstam F, Bernstam EV, Cohen T, Automated identification of molecular effects of drugs (AIMED), J. Am. Med. Inform. Assoc 23 (4) (2016) 758–765, 10.1093/jamia/ocw030. [PubMed: 27107438]

[18]. Keselman A, Rosemblat G, Kilicoglu H, Fiszman M, Jin H, Shin D, Rindflesch TC, Adapting semantic natural language processing technology to address information overload in influenza epidemic management, J. Am. Soc. Inf. Sci. Technol 61 (12) (2010) 2531–2543.

[19]. Rosemblat G, Shin D, Kilicoglu H, Sneiderman C, Rindflesch TC, A methodology for extending domain coverage in SemRep, J. Biomed. Inform 46 (6) (2013) 1099–1107. [PubMed: 23973273]

[20]. Hristovski D, Friedman C, Rindflesch TC, et al. , Exploiting semantic relations for literature-based discovery, AMIA Annu. Symp. Proc (2006) 349–353. [PubMed: 17238361]

[21]. Henry S, McInnes BT, Literature based discovery: models, methods, and trends, J. Biomed. Inform 74 (2017) 20–32. [PubMed: 28838802]

[22]. Percha B, Garten Y, Altman RB, Discovery and explanation of drug-drug interactions via text mining, Pac Symp. Biocomput. Pac. Symp. Biocomput (2012) 410–421. [PubMed: 22174296]

[23]. Duke JD, Han X.u., Wang Z, Subhadarshini A, Karnik SD, Li X, Hall SD, Jin Y, Callaghan JT, Overhage MJ, Flockhart DA, Strother RM, Quinney SK, Li L, Butte AJ, Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions, PLoS Comput. Biol 8 (8) (2012) e1002614, 10.1371/journal.pcbi.1002614. [PubMed: 22912565]

[24]. Kilicoglu H, Fiszman M, Rodriguez A, et al. Semantic MEDLINE: A Web Application to Manage the Results of PubMed Searches. In: Salakoski T, Schuhmann DR, Pyysalo S, eds. Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008). 2008. 69–76.

[25]. Zhang R, Cairelli MJ, Fiszman M, Rosemblat G, Kilicoglu H, Rindflesch TC, Pakhomov SV, Melton GB, Using semantic predications to uncover drug-drug interactions in clinical data, J. Biomed. Inform 49 (2014) 134–147, 10.1016/j.jbi.2014.01.004. [PubMed: 24448204]

[26]. Zhang R, Hristovski D, Schutte D, Kastrin A, Fiszman M, Kilicoglu H, Drug repurposing for COVID-19 via knowledge graph completion, J. Biomed. Inform 115 (2021) 103696, 10.1016/j.jbi.2021.103696. [PubMed: 33571675]

[27]. Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT (1). 2019.

[28]. Zhang R, Adam TJ, Simon G, et al. , Mining biomedical literature to explore interactions between cancer drugs and dietary supplements, AMIA Summits Transl. Sci. Proc 2015 (2015) 69. [PubMed: 26306241]

[29]. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv190711692 Cs Published Online First: 26 July 2019.http://arxiv.org/abs/1907.11692 (accessed 8 Apr 2021).

[30]. Wang LL, Tafjord O, Cohan A, et al. SUPP.AI: Finding Evidence for Supplement-Drug Interactions. ArXiv190908135 Cs Published Online First: 6 July 2020.http://arxiv.org/abs/1909.08135 (accessed 8 Apr 2021).

[31]. Vasilakes J, Rizvi R, Melton GB, Pakhomov S, Zhang R, Evaluating active learning methods for annotating semantic predications, JAMIA Open 1 (2) (2018) 275–282. [PubMed: 30740594]

[32]. Gu Y, Tinn R, Cheng H, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ArXiv200715779 Cs Published Online First: 20 August 2020.http://arxiv.org/abs/2007.15779 (accessed 25 Sep 2020).

[33]. Lee J, Yoon W, Kim S, et al. , BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240. [PubMed: 31501885]

[34]. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019. 58–65.

[35]. Alsentzer E, Murphy J, Boag W, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019. 72–8.

[36]. Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. In: 2009 Ninth IEEE International Conference on Data Mining. Miami Beach, FL, USA: IEEE 2009. 908–13. doi:10.1109/ICDM.2009.108.

[37]. Kingma DP, Ba J. Adam: A method for stochastic optimization. ArXiv Prepr ArXiv14126980 2014.

[38]. Natural Medicines Interaction Checker. https://naturalmedicines.therapeuticresearch.com/tools/interaction-checker.aspx.

[39]. Demner-Fushman D, Rogers WJ, Aronson AR, MetaMap Lite: an evaluation of a new Java implementation of MetaMap, J. Am. Med. Inform. Assoc 24 (4) (2017) 841–844. [PubMed: 28130331]

[40]. Ioannidis V, Song X, Manchanda S, et al. DRKG - Drug Repurposing Knowledge Graph for COVID-19. Published Online First: 2020.https://github.com/gnn4dr/DRKG/.

[41]. Breit A, Ott S, Agibetov A, Samwald M, Lu Z, OpenBioLink: a benchmarking framework for large-scale biomedical link prediction, Bioinformatics 36 (13) (2020) 4097–4098, 10.1093/bioinformatics/btaa274. [PubMed: 32339214]

[42]. Santos A, Colaço AR, Nielsen AB, et al. , Clinical knowledge graph integrates proteomics data into clinical decision-making, Bioinformatics (2020), 10.1101/2020.05.09.084897.

[43]. Zheng S, Rao J, Song Y, Zhang J, Xiao X, Fang EF, Yang Y, Niu Z, PharmKG: a dedicated knowledge graph benchmark for bomedical data mining, Brief Bioinform. 22 (4) (2021), 10.1093/bib/bbaa344.

[44]. Yamane T, Statistics, an introductory analysis, 2nd Ed., Harper and Row, New York, 1967.

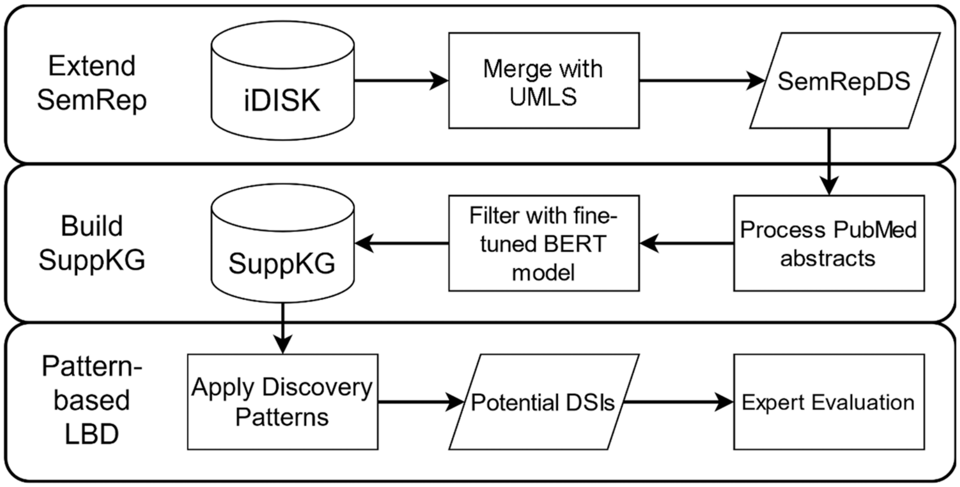

Fig. 1.

Overview of the methodology.

Table 1

Comparison of the output from SemRep and SemRepDS.

	SemRep	SemRepDS	Difference
DS Entities Mentions	539,863	1,395,653	855,790 (+158.52%)
Relations with at least one DS Entity	71,669	219,977	148,308 (+206.93%)

Table 2

Performance of SemRepDS on a DS NER task.

	Precision	Recall	F_1 Score
SemRepDS	0.4458	0.4752	0.4600

Table 3

Distribution of the predicates after filtering the combined relations with PubMedBERT.

Predicate	Count	Predicate	Count
TREATS	525,719 (19.40%)	ISA	24,383 (0.90%)
COEXISTS_WITH	511,108 (18.86%)	PREDISPOSES	21,236 (0.78%)
PROCESS_OF	257,484 (9.50%)	COMPARED_WITH	19,295 (0.71%)
CAUSES	235,599 (8.69%)	ADMINISTERED_TO	18,940 (0.70%)
INTERACTS_WITH	213,407 (7.87%)	METHOD_OF	15,203 (0.56%)
AFFECTS	209,644 (7.74%)	DIAGNOSES	6,531 (0.24%)
LOCATION_OF	169,137 (6.24%)	MEASURES	4,562 (0.17%)
PART_OF	131,192 (4.84%)	PRECEDES	3,132 (0.12%)
ASSOCIATED_WITH	120,297 (4.44%)	COMPLICATES	1,846 (0.07%)
USES	96,211 (3.55%)	HIGHER_THAN	1,817 (0.07%)
INHIBITS	59,524 (2.20%)	OCCURS_IN	1,535 (0.06%)
AUGMENTS	55,167 (2.04%)	MANIFESTATION_OF	1,199 (0.04%)
DISRUPTS	45,017 (1.66%)	CONVERTS_TO	1,055 (0.04%)
PRODUCES	41,402 (1.53%)	SAME_AS	156 (0.01%)
STIMULATES	39,332 (1.45%)	LOWER_THAN	110 (0.00%)
PREVENTS	39,104 (1.44%)		
TOTAL		**2,710,240**	

Table 4

High-level comparison of other biomedical knowledge graphs and SuppKG.

Knowledge Graph	Nodes	Relations	Relation Types	Purpose	Source Data
DRKG[40]	97 K	5.9 M	107	Drug Repurposing	Drugbank, GNBR, Hetionnet, STRING, IntAct, DGldb
OpenBioLink (directed, high quality) [41]	185 K	9.3 M	28	Benchmarking	Bgee, STITCH, CTD, HPO, DrugCentral, SIDER, GO, DO, HPO, UBERON, DisGeNet, STRING, UniProt, NCBI, PubChem, REACTOME, KEGG
Clinical KG[42]	16 M	220 M	57	Omic based research	See: https://github.com/MannLabs/CKG
Pharm KG[43]	188 k	1.1 M	29	Benchmarking	DrugBank, TTD, OMIM, PharmGKB, GNNBR
SuppKG	**56,635**	**595,222**	**31**	**Dietary Supplement Research**	**UMLS, iDISK, PubMed**

Table 5

Examples of discovered DSIs and expert evaluations.

Identified drug-supplement Interaction	Identified Pathway	DS-Gene Source Sentences	Gene-Drug Source Sentences	Clinical Context
Curcumin And Oxytocin	Curcumin *INHIBITS* Interleukin 1-beta *PRODUCES* Oxytocin	“Additionally, the curcumin could effectively decrease mRNA expression of proinflammatory cytokines (IL-1beta, IL-6, and TNF-alpha) and suppress NF-kappaB activation.” PMID: 30,551,030	“Thus, systemic IL-1beta acts centrally to increase oxytocin secretion.” PMID: 16,553,786	Curcumin is a yellow chemical produced by turmeric (Curcuma longa). It is sold as a herbal supplement and a culinary spice. Oxytocin is a medication used to contract the uterus to increase the speed of labor and to stop bleeding following delivery. We found that Interleukin-1-beta influences oxytocin signaling and production. Therefore, co-administration of both may reduce the effect of oxytocin on labor induction
Melatonin And Reboxetine	Melatonin *INTERACTS_WITH* Acetylcholine receptor *INHIBITS* Reboxetine	“Melatonin, the pineal hormone produced during the dark phase of the light–dark cycle, modulates neuronal acetylcholine receptors located presynaptically on nerve terminals of the rat vas deferens.” PMID: 10,454,762	“In conclusion, (−)-reboxetine non-competitively inhibits muscle AChRs by binding to the TCP luminal site and by inducing receptor desensitization (maybe by interacting with non-luminal sites), a mechanism that is shared by tricyclic antidepressants.”, PMID: 23,917,086	Melatonin is a hormone primarily released by the pineal gland at night and has long been used as a dietary supplement for the short-term treatment of insomnia. Reboxetine is a norepinephrine reuptake inhibitor used for the treatment of major depression. The clinical action of reboxetine may be partly produced by its inhibitory action on acetylcholine receptors, and melatonin may act as an acetylcholine receptors inducer. Therefore, when taken together, melatonin might interfere with the antidepressant effect of reboxetine.
Garlic And Cisplatin	Garlic *INHIBITS* Interleukin 1, beta *INTERACTS_WITH* Cisplatin	“Allicin inhibited interleukin-1beta (IL-1beta) induced overproduction of nitric oxide, inducible nitric oxide synthase, prostaglandin E2, and cyclooxygenase-2, as well as pro-inflammatory cytokines tumor necrosis factor alpha and interleukin-6 in chondrocytes in a dose-dependent manner.” PMID: 30,160,278	“Gomisin N and gamma-schizandrin also decreased the transcription of interleukin 1beta and inflammatory chemokines.” PMID: 23,085,209	Cisplatin is a chemotherapy medication used to treat a number of cancers. Garlic administration results in a dose-dependent reduction of IL-6, and cisplatin increased IL-6 secretion and cellular migration and proliferation. Therefore, concomitant administration of garlic and cisplatin may interfere in cancer treatment.
Buckwheat honey And Heparin	Buckwheat honey *INTERACTS_WITH* Chemoattractant cytokine *INTERACTS_WITH* Heparin	“The Effect of Manuka Honey on dHL-60 Cytokine, Chemokine, and Matrix-Degrading Enzyme Release under Inflammatory Conditions.” PMID: 31,245,627	“We investigated the possibility that specificity exists in the recognition of particular heparin/heparan sulfate structures by chemokines, by studying the binding of four members of the chemokine superfamily to heparin and heparan sulfate.” PMID: 7,922,353	Heparin is the most common clinically used anticoagulant and also has anti-inflammatory properties through inhibition of proinflammatory cytokines. Honey is used clinically to treat cough and allergies and has also been shown to inhibit the expression of proinflammatory cytokines. Therefore, honey and heparin may have a synergistic anti-inflammatory effect.