SUPP.AI: Finding Evidence for Supplement-Drug Interactions

Lucy Lu Wang, Oyvind Tafjord, Arman Cohan, Sarthak Jain, Sam Skjonsberg, Carissa Schoenick, Nick Botner, Waleed Ammar
Allen Institute for AI
Seattle, WA 98103
{lucyw, oyvindt, armanc, sarthakj, sams, carissas, nickb, waleeda}@allenai.org

Abstract

Dietary supplements are used by a large portion of the population, but information on their pharmacologic interactions is incomplete. To address this challenge, we present SUPP.AI, an application for browsing evidence of supplement-drug interactions (SDIs) extracted from the biomedical literature. We train a model to automatically extract supplement information and identify such interactions from the scientific literature. To address the lack of labeled data for SDI identification, we use labels of the closely related task of identifying drug-drug interactions (DDIs) for supervision. We fine-tune the contextualized word representations of the RoBERTa language model using labeled DDI data, and apply the fine-tuned model to identify supplement interactions. We extract 195k evidence sentences from 22M articles (P=0.82, R=0.58, F1=0.68) for 60k interactions. We create the SUPP.AI application for users to search evidence sentences extracted by our model. SUPP.AI is an attempt to close the information gap on dietary supplements by making up-to-date evidence on SDIs more discoverable for researchers, clinicians, and consumers.

1 Introduction

More than half of US adults use dietary supplements (Kantor et al., 2016). Supplements include vitamins, minerals, enzymes, and other herbal and animal products. Supplements and pharmaceutical drugs, when taken together, can cause adverse interactions (Sprouse and van Breemen, 2016; Asher et al., 2017; Ronis et al., 2018). Some studies describe the prevalence of supplement-drug interactions (SDIs) in the hospital setting (Levy et al., 2016, 2017a,b) or among groups such as patients with cancer (Alsanad et al., 2014), cardiac disease (Karny-Rahkovich et al., 2015), HIV/AIDS (Jalloh et al., 2017), or Alzheimer's disease (Spence et al., 2017). However, these studies largely rely on manual curation of the literature, and are slow and expensive to produce and update. It is also difficult to aggregate their results, and researchers, clinicians, and consumers can lack appropriate up-to-date information to make informed decisions about supplement use.

A resource that provides experimental evidence for SDIs could serve as a good intermediary tool, allowing experts to quickly access information and translate it for healthcare providers and consumers. Such a tool could ease the bottleneck of manual curation by directing researcher attention to the most pertinent and novel interactions appearing in recent trials and case reports. Our goal is to create such a resource using state-of-the-art methods in NLP and IE, and allow users to better identify appropriate uses of supplements as well as risks for SDIs.

Automated approaches have been used to extract drug-drug interactions (DDIs) from literature and other documents (Tari et al., 2010; Percha et al., 2011; Segura-Bedmar et al., 2011; Kim et al., 2014; Zhang et al., 2016; Noor et al., 2017; Lim et al., 2018), complementing broadly-used but primarily manual methods (Grizzle et al., 2019). We expand upon this work to automatically extract evidence for SDIs, as well as supplement-supplement interactions (SSIs), from a large corpus of 22M biomedical and clinical texts derived from Semantic Scholar.[1] We leverage labeled datasets for DDI identification for supervision, and train a model that transfers to the related task of identifying supplement interactions. We surface the resulting evidence on SUPP.AI for browsing and search.

To summarize, our contributions are:

1. A model for identifying SDI/SSI evidence
2. A dataset of 195k evidence sentences supporting supplement interactions, publicly accessi-

[1] https://www.semanticscholar.org/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 362–371

ble for download or via a web API, and

3. SUPP.AI, an application for browsing and searching the extracted evidence.

2 Supplement interaction browser

Information on supplement interactions have immediate implications on public health, which can only be realized by making the data easily accessible to any interested researcher, clinician or consumer. We note that many medical providers in developing countries do not have subscriptions to clinical databases such as TRC[2] and UpToDate,[3] and may lack an easy way to identify possible supplement interactions before prescribing drugs to their patients. To fill this gap, we develop SUPP.AI (available at https://supp.ai/), an application for browsing evidence of supplement interactions extracted from clinical and biomedical literature. SUPP.AI allows users to:

- Search for supplements or drugs,
- Search through potential interactions,
- Browse evidence sentences with supplement and drug entities highlighted,
- Navigate links to source papers

We design SUPP.AI to be a rapid way for users to access and search extracted SDI and SSI evidence. Our goal for this application is to provide a high quality, broadly-sourced, up-to-date, and easily accessible platform for searching through SDI and SSI evidence, while providing sufficient information for users to judge the quality of each piece of evidence. In Section 3, we describe the NLP pipeline used to extract evidence from scientific papers. Below, we describe the user interface and data features of SUPP.AI.

2.1 User interface

Besides the main search page seen by users when they first navigate to the site, SUPP.AI consists of two other types of pages: entity and interaction pages. Entity pages provide information about one supplement or drug, and a list of potential interacting entities, sorted by quantity of evidence. We provide information such as synonyms, drug trade names, and definitions about each entity upon hover over or expansion. Interaction pages display all discovered pieces of evidence supporting an interaction between a pair of entities. The evidence is sorted by additional features extracted from source papers, such as the level of evidence and recency, discussed in Section 2.2.

Figure 1 shows the interface, with results for the ginkgo supplement. Results on the entity page (*left*) list 140 possible interactions to entities such as Warfarin and Nitric Oxide. When a result is selected, the interaction page is displayed (*right*), showing evidence sentences supporting the interaction along with metadata and links to each source paper. Spans linked to supplement and drug entities in evidence sentences are highlighted. To see more context or detail about the interaction, the user can navigate to the source paper to continue reading.

2.2 Supporting data for search

We extract additional paper metadata as a way to judge evidence quality. From Semantic Scholar, we retrieve the paper title, authors, publication venue, and year of publication. Medical Subject Headings (MeSH) tags associated with each paper are used to determine whether its results are derived from clinical trials, case reports, or animal studies. We also attempt to identify the retraction status of each paper, again using MeSH tags. Evidence sentences are ordered and presented based on associated paper metadata, prioritizing non-retracted studies, clinical trials, human studies, and recency (year of publication).

Using the RxNorm relationship *has_tradename* via the Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004), we derive trade names associated with drug ingredients, e.g. *Prozac* and *Sarafem* are trade names of the ingredient *fluoxetine*. Trade drugs are associated with active drug ingredients and indexed for search. Users can query a trade name rather than an active ingredient and be directed to the relevant interactions.

2.3 Data & API

Data on the site are periodically updated as new papers are incorporated into the Semantic Scholar corpus. Snapshots of the data are available for download at https://api.semanticscholar.org/supp/. Live data on the site, which is updated more frequently, can be accessed through our search API, documented at https://supp.ai/docs/api. Additionally, we provide training data, evaluation data, and the curated drug/supplement identifier lists (discussed in Section 3) used to

[2] https://naturalmedicines.therapeuticresearch.com/
[3] https://www.uptodate.com/

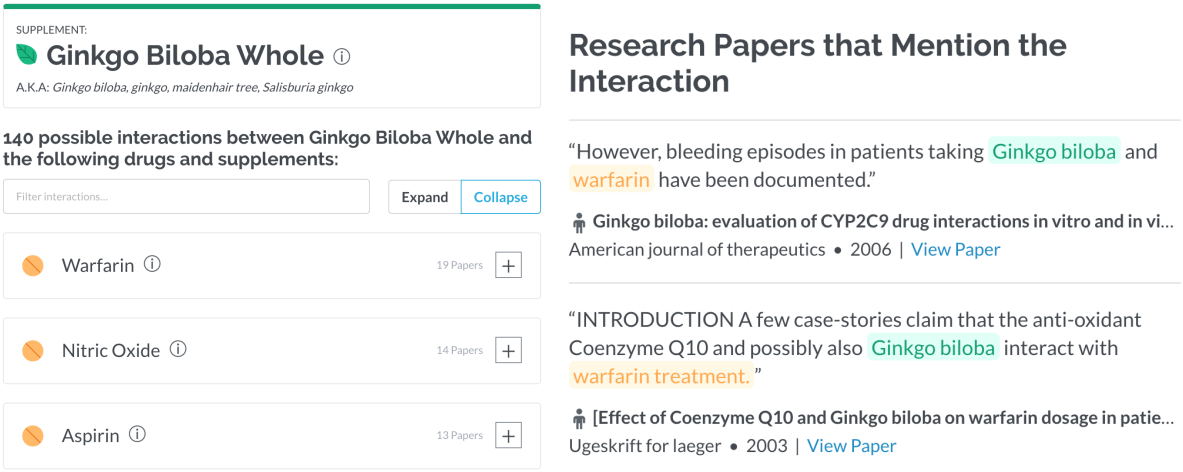

Figure 1: Top results for interactions with Ginkgo (*left*), and top evidence sentences for the SDI between Ginkgo and Warfarin (*right*). Source paper metadata are given below each evidence sentence.

produce the dataset of interactions at https://github.com/allenai/sdi-detection. We encourage others to reuse our data and model to improve information availability around supplement interactions and safety.

3 Methods

An overview of our NLP pipeline is given in Figure 2. We first retrieve Medline-indexed articles using the Semantic Scholar API,[4] and pre-process the text to generate candidate evidence sentences (Section 3.1). We then use our DDI-detection model, a neural network classifier based on BERT (Devlin et al., 2018) and fine-tuned on labeled DDI data from Ayvaz et al. (2015) (Section 3.2), to classify sentences for the existence of an interaction. Sentences classified as positive by our model are collated and surfaced on SUPP.AI (Section 2).

3.1 Generating candidate evidence

Approximately 22M Medline-indexed articles are downloaded using the Semantic Scholar API. The scispaCy library (Neumann et al., 2019) is used to perform sentence tokenization, NER, and entity linking over all paper abstracts. Entity mentions are linked to Concept Unique Identifiers (CUIs) from the UMLS Metathesaurus. An example sentence from Vaes and Hendeles (2000) is shown with linked entity mentions:

Hemorrhage (C0019080) and tendencies were noted in four cases (C0868928) with ginkgo (C0330205) use and in three cases (C0868928) with garlic (C0017102); in none of these cases (C0868928) were patients (C0030705) receiving warfarin (C0043031).

Of these linked entities, we preserve entities on a list of curated supplements and drugs (entities in blue). We generate these curated lists in a semi-automatic fashion, by querying the children of UMLS supplement and drug classes and performing fuzzy name matching to known supplements or drugs crawled from the web. We also perform clustering of similar entities to reduce redundancy in the final dataset, e.g., combining several variants of Vitamin D together into a single entity. Details on identifier curation and clustering are given in Appendix A.

We retain all sentences containing at least two entity mentions. For each sentence, we generate candidate evidence as each combination of two entity spans from that sentence.

3.2 DDI-detection model

We train a DDI-detection model to predict whether a given candidate sentence provides evidence of an interaction between two drug entities. Our DDI-detection model uses pre-trained BERT models (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) to encode input sequences. These models have been shown to be effective at domain transfer, and are able to achieve high performance using small amounts of task-specific annotated data. In particular, we use the large version of the pre-trained RoBERTa model, a further-optimized BERT model, that has approximately 340M parameters (Liu et al., 2019).

[4] https://api.semanticscholar.org/

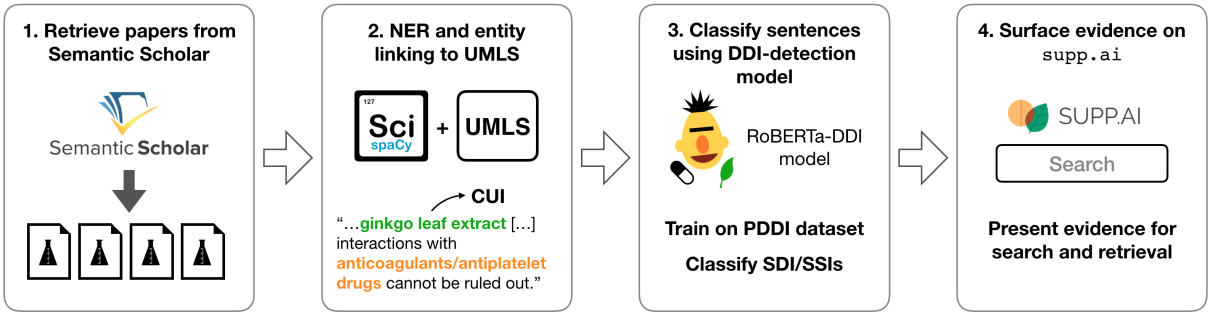

Figure 2: Pipeline for identifying sentences containing evidence of SDIs and SSIs.

We fine-tune the pre-trained embeddings of the RoBERTa language model using labeled data for DDI classification, and we call the resulting model RoBERTa-DDI.

Input layer: The input layer consists of the sequence of byte-pair encoding word pieces (Radford et al., 2019) in a sentence. We replace entity mention spans with the special tokens `[Arg1]` and `[Arg2]`. This helps generalization by preventing the model from memorizing entity pairs with positive interactions in the training set. For example:

```
[CLS] Combination [Arg1] may also
decrease the plasma concentration of
[Arg2].  [SEP]
```

where `[Arg1]` and `[Arg2]` replace the spans "hormonal contraceptives" and "acetaminophen" respectively. We add special tokens `[CLS]` and `[SEP]` at the beginning and end of each sentence to leverage their representations learned in pre-training. At prediction time, candidate sentences are masked similarly and fed to the trained model.

Model architecture: As the name implies, RoBERTa-DDI uses the pre-trained RoBERTa representations (Liu et al., 2019) to encode input sequences. We refer readers to Liu et al. (2019), Devlin et al. (2018), and Vaswani et al. (2017) for more details on BERT and transformer architecture. For the RoBERTa-DDI model, we add a dropout layer followed by one feedforward (output) layer with a softmax non-linearity, which takes the representation of the `[CLS]` token at the top transformer layer as input and outputs probabilities for labels $\{0, 1\}$, where 1 indicates an interaction.

Model training: Due to similarities between DDIs and SDIs/SSIs, we hypothesize that a classifier trained to identify DDI evidence should perform well in identifying SDI and SSI evidence. We therefore take advantage of existing labeled data for categorizing DDIs to fine-tune the model. We use pre-trained weights distributed by the authors of Liu et al. (2019), and further fine-tune the model parameters (as well as parameters of the output layer) using labeled DDI data from the Merged-PDDI dataset (Ayvaz et al., 2015).

In particular, we use training data from the DDI-2013 (Segura-Bedmar et al., 2013) and NLM-DailyMed (Stan et al., 2014) datasets, as they are relatively large and contain evidence sentences with annotated drug mention spans. The DDI-2013 dataset consists of sentences extracted from DrugBank and Medline; the NLM-DailyMed dataset draws sentences from cardiovascular drug product labels retrieved from DailyMed. Both datasets contain multi-class labels for different types of interactions. We distinguish between detection, a binary classification problem where the goal is to determine whether an interaction exists or not, and multi-class classification, where the goal is to determine the type of interaction. In this work, we focus on detection, but provide results for a variant of our model trained on classification that obtains SOTA performance compared to prior work.

For detection, we collapse labels corresponding to all interaction types (e.g., mechanism, advise, effect, etc.) into binary labels of 0 and 1, where 0 means no interaction, and 1 means an interaction of some type exists. Collapsing the positive labels is necessary for training one DDI-detection model on both the DDI-2013 and NLM-DailyMed datasets, since the two datasets are annotated with inconsistent interaction types. We preserve the train/test splits used in Ayvaz et al. (2015), and create a development set from the training set for iteration on model design and tuning.

A sentence from the training data can contain multiple drug entities. For training, we generate pairwise combinations of drug mention spans in each sentence. We note that many sentences are

seen multiple times by our model with different labeled spans. Due to combinatorial explosion, and to prevent our model from learning excessively from a few instances containing lots of entity mentions, we restrict the training data to sentences containing less than or equal to 100 pairwise entity combinations. Table 1 shows the resulting data splits for the two datasets.

Dataset	Train	Dev.	Test	Label=1
DDI-2013	18362	2069	5688	17.2%
NLM-DailyMed	11372	1255	927	22.7%

Table 1: DDI training data split.

Our training hyperparameters follow those presented by Liu et al. (2019) (learning rate = 1e-5; 4 epochs). No additional hyperparameter tuning is performed.

4 Results & evaluation

Of the 22M articles we retrieve, around 4.6M abstracts contain candidate sentences. After initial filtering, 33.0M candidate sentences containing supplement entity mentions are classified by RoBERTa-DDI. Around 625k (1.9%) of these sentences are classified as positive for an interaction. We perform entity normalization across positive sentences based on CUI clusters, and perform additional ad hoc filtering of evidence to eliminate incorrectly detected spans resulting from poor NER and linking, such as the span "retina" linking to Vitamin A (C0040845). The resulting 195k sentences contain mentions of 2044 unique supplements and 2772 unique drugs, and provide evidence sentences for 60k interactions sourced from 133k papers.

Comparisons of model variants on DDI classification and detection (including SOTA results on both tasks) are given in Appendix B. To evaluate the transferability of DDI detection to the related task of SDI/SSI detection, we use a test set consisting of 500 sentences annotated for the presence or absence of a supplement interaction. To obtain a balanced test set despite the rare presence of a positive interaction, we sample half the instances from the set of sentences labeled as positive by a previous variant of our model based on fine-tuning BERT-large, and the other half from those labeled as negative. After manual annotation, 40% of the sampled instances were positive for an interaction. Annotation was performed by two authors without seeing model predictions, with an inter-annotator agreement of 94%. This test set was used for final evaluation, and never for model development or tuning. Table 2 shows the performance of RoBERTa-DDI on the DDI and supplement test sets. Performance on the SDI test set has precision 0.82, recall 0.58, and F1-score 0.68. Although there is performance degradation during transfer, the precision of detection remains high at 0.82.

Decrease in recall can be attributed to a larger percentage of positive instances in the SDI test set (roughly 40%, compared to 20% in the DDI training data). Another factor is the presence of incorrectly labeled entity spans in the supplements test set due to NER/linking errors. To better understand this second source of errors, we attempt to evaluate the performance of the scispaCy entity linker. Processing each sentence from the two DDI training sets using scispaCy, we determine that only 80% of drug entities from DDI-2013 and 76% from NLM-DailyMed are recognized and linked. The likelihood of supplement entities being successfully linked is likely lower, due to sparse training data for supplement NER and linking. These numbers provide an estimate of the global ceiling on recall for our model. In future work, we aim to explore ways to improve NER and linking and assess their impact on the results of SDI detection. SDI/SSI sentences in our output set can also be labeled by biomedical expert annotators and used to further tune the model for SDI/SSI detection.

Evaluation set	Prec.	Rec.	F1
Drugs (DDI-2013)	0.90	0.87	0.88
Drugs (NLM-DailyMed)	0.83	0.85	0.84
Supplements-500	0.82	0.58	0.68

Table 2: The RoBERTa-DDI model (trained on drug-drug interaction labels) is evaluated on two DDI evaluation sets (first two rows) and our supplement interaction evaluation set (last row).

5 Discussion

Information describing the safety and efficacy of dietary supplements can be difficult to find. The inability to locate evidence of SDIs can challenge clinician ability to advise patients and cause risks for consumers of dietary supplements. It is our hope that extracting evidence for SDIs/SSIs from a large corpus of scientific literature and making the evidence available through an easily accessible search interface can offset some of these risks.

This work demonstrates how NLP techniques can be extraordinarily useful for extracting information and relationships specific to an application domain in healthcare. Re-purposing existing labeled data from related domains (that would be expensive to generate in a new domain) can be a way to derive maximum utility from curation efforts. Continuing, we look to investigate fine-grained interaction types, and provide better classification of the level of evidence provided by each sentence or document towards a particular SDI or SSI. We also aim to leverage similar techniques for identifying evidence of indications, contraindications, and side effects of dietary supplements from the biomedical and clinical literature, and make these discoverable on SUPP.AI.

5.1 Related Work

Consumer-facing websites such as the NIH Office of Dietary Supplements[5] or WebMD[6] provide facts about common supplements, but this information can be incomplete and may not support researcher or clinician needs. TRC Natural Medicines[7] and UpToDate[8], two dedicated clinical resources, contain high-quality, curated evidence, but may not be broadly accessible due to their subscription format. Drug databases like DrugBank (Wishart et al., 2018), RxNorm (Nelson et al., 2011), and the National Drug File Reference Terminology (NDFRT) (Simonaitis and Schadow, 2010) contain only partial coverage of supplement terminology (Manohar et al., 2015b), and primarily focus on aggregating drug information.

Several prior studies have experimented with extracting safety information of supplements and supplement interactions from various forms of text. Zhang et al. (2015) employ machine learning techniques to filter supplement interaction relationships in SemMedDB, a database of relationships extracted from Medline articles. Jiang et al. (2017) develop a model for identifying adverse effects related to dietary supplements as reported by consumers on Twitter, and discover 191 adverse effects pertaining to 4 dietary supplements. Fan et al. (2016) and Fan and Zhang (2018) analyze unstructured clinical notes to predict whether a patient started, continued or discontinued a dietary supplement, which can be useful as a building block for identifying adverse effects in clinical notes (as attempted by the same authors in Fan et al. (2017) for the drug warfarin). Wang et al. (2017) proposes using topic models to analyze the adverse effects of dietary supplements as mentioned in the Dietary Supplement Label Database, and finds that Latent Dirichlet Allocation models (Blei et al., 2003) can be used to group dietary supplements with similar adverse effects based on their labels. As far as we know, there are no other studies investigating the task of sentence-level identification of SDI/SSI evidence from the scientific literature. No previous work has investigated the utility of using labeled DDI data for transfer learning to SDI/SSI identification.

5.2 Limitations

There are several limitations of this work. First, we distinguish between supplements and drugs. Both supplements and drugs are pharmacologic entities, with their separate classification more attributable to marketing and social pressures rather than functional differences. However, due to this somewhat arbitrary distinction, supplement entities are not well represented in databases of pharmaceutical entities, and less information is publicly available on their interactions. We also use UMLS CUIs as a way of identifying supplement and drug entities. The lack of a standardized terminology to describe dietary supplements is discussed in Manohar et al. (2015a) and Wang et al. (2016), which estimate UMLS coverage of these terms to be between 14-54%. This limitation prevents us from identifying many supplement entities. Lastly, our dependence on NLP-pipeline tools sets a performance ceiling due to unsolved problems in NER and linking. Although scispaCy is performant and detects a large number of relevant entities, our evaluations show that many supplement and drug entities are missed. A system such as MetaMapLite (Demner-Fushman et al., 2017) has higher recall, but performance is slow and there are practical challenges to using it to process large numbers of documents.

Conclusion

Insufficient regulation in the supplement space introduces dangers for the many users of these supplements. Claims of interactions are difficult to validate without links to source evidence. We create an NLP pipeline to detect SDI/SSI evidence from scientific literature, leveraging UMLS identifiers, scispaCy for NER and entity linking, BERT-based

[5] https://ods.od.nih.gov/
[6] https://www.webmd.com/vitamins/index
[7] https://naturalmedicines.therapeuticresearch.com/
[8] https://www.uptodate.com/

language models for classification, and labeled data from a related domain for training. We use this pipeline to extract evidence from 22M biomedical and clinical articles with high precision. The extracted SDI/SSI evidence are made search-able through a public web interface, `SUPP.AI`, where we integrate additional metadata about source papers to help users make decisions about the reliability of evidence. Our dataset and web interface can be leveraged by researchers, clinicians, and curious individuals to increase understanding about supplement interactions. We hope to encourage additional research to improve the safety and benefits of dietary supplements for their consumers.

Acknowledgments

We would like to thank Oren Etzioni for his indispensable feedback and support of this project. We thank Amandalynne Paullada for contributing to an earlier prototype, and we thank Asma Ben Abacha, Pieter Cohen, Taha Kass-Hout, Beth Ranker, Lia Schmitz, Heidi Tafjord, and our users for helpful comments on improving `SUPP.AI`.

References

Saud M. Alsanad, Elizabeth M. Williamson, and Rachel L. Howard. 2014. Cancer patients at risk of herb/food supplement-drug interactions: a systematic review. *Phytotherapy research: PTR*, 28(12):1749–1755.

Gary N. Asher, Amanda H. Corbett, and Roy L. Hawke. 2017. Common Herbal Dietary Supplement-Drug Interactions. *American Family Physician*, 96(2):101–107.

Serkan Ayvaz, John Horn, Oktie Hassanzadeh, Qian Zhu, Johann Stan, Nicholas P. Tatonetti, Santiago Vilar, Mathias Brochhausen, Matthias Samwald, Majid Rastegar-Mojarad, Michel Dumontier, and Richard D. Boyce. 2015. Toward a complete dataset of drug-drug interaction information from publicly available sources. *Journal of Biomedical Informatics*, 55:206–217.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.

Geeticka Chauhan, Matthew B. A. McDermott, and Peter Szolovits. 2019. Reflex: Flexible framework for relation extraction in multiple domains. In *BioNLP@ACL*.

Dina Demner-Fushman, Willie J. Rogers, and Alan R. Aronson. 2017. Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association : JAMIA*, 24:841–844.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Yadan Fan, Terrence Adam, Reed McEwan, Serguei V. S. Pakhomov, Genevieve B. Melton, and Rui Zhang. 2017. Detecting signals of interactions between warfarin and dietary supplements in electronic health records. In *MedInfo*.

Yadan Fan, Lu He, and Rui Zhang. 2016. Classification of use status for dietary supplements in clinical notes. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1054–1061.

Yadan Fan and Rui Zhang. 2018. Using natural language processing methods to classify use status of dietary supplements in clinical notes. In *BMC Medical Informatics and Decision Making*.

Amy J. Grizzle, John Horn, Carol Collins, Jodi Schneider, Daniel C. Malone, Britney Stottlemyer, and Richard David Boyce. 2019. Identifying Common Methods Used by Drug Interaction Experts for Finding Evidence About Potential Drug-Drug Interactions: Web-Based Survey. *Journal of Medical Internet Research*, 21(1):e11182.

Mohamed A. Jalloh, Philip J. Gregory, Darren Hein, Zara Risoldi Cochrane, and Aleah Rodriguez. 2017. Dietary supplement interactions with antiretrovirals: a systematic review. *International journal of STD & AIDS*, 28(1):4–15.

Keyuan Jiang, Yongbing Tang, G. Elliott Cook, and Michael M. Madden. 2017. Discovering potential effects of dietary supplements from twitter data. In *Proceedings of the 2017 International Conference on Digital Health*, pages 119–126.

Elizabeth D. Kantor, Colin D. Rehm, Mengmeng Du, Emily White, and Edward L. Giovannucci. 2016. Trends in Dietary Supplement Use Among US Adults From 1999-2012. *JAMA*, 316(14):1464–1474.

Orith Karny-Rahkovich, Alex Blatt, Gabby Atalya Elbaz-Greener, Tomer Ziv-Baran, Ahuva Golik, and Matityahu Berkovitch. 2015. Dietary supplement consumption among cardiac patients admitted to internal medicine and cardiac wards. *Cardiology Journal*, 22(5):510–518.

Sun Kim, Haibin Liu, Lana Yeganova, and W. John Wilbur. 2014. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. *Journal of biomedical informatics*, 55:23–30.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Ilana Levy, Samuel Attias, Eran Ben Arye, Lee Goldstein, and Elad Schiff. 2016. Interactions between dietary supplements in hospitalized patients. *Internal and Emergency Medicine*, 11(7):917–927.

Ilana Levy, Samuel Attias, Eran Ben-Arye, Lee Goldstein, and Elad Schiff. 2017a. Adverse events associated with interactions with dietary and herbal supplements among inpatients. *British Journal of Clinical Pharmacology*, 83(4):836–845.

Ilana Levy, Samuel Attias, Eran Ben-Arye, Lee Goldstein, and Elad Schiff. 2017b. Potential drug interactions with dietary and herbal supplements during hospitalization. *Internal and Emergency Medicine*, 12(3):301–310.

Sangrak Lim, Kyubum Lee, and Jaewoo Kang. 2018. Drug drug interaction extraction from the literature using a recursive neural network. volume 13, page e0190926.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Nivedha Manohar, Terrance J. Adam, Serguei V. Pakhomov, Genevieve B. Melton, and Rui Zhang. 2015a. Evaluation of Herbal and Dietary Supplement Resource Term Coverage. *Studies in Health Technology and Informatics*, 216:785–789.

Nivedha Manohar, Terrence Adam, Serguei V. S. Pakhomov, Genevieve B. Melton, and Rui Zhang. 2015b. Evaluation of herbal and dietary supplement resource term coverage. *Studies in health technology and informatics*, 216:785–9.

Stuart J. Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. 2011. Normalized names for clinical drugs: Rxnorm at 6 years. *Journal of the American Medical Informatics Association : JAMIA*, 18:441–8.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *ArXiv*, abs/1902.07669.

Adeeb Noor, Abdullah Assiri, Serkan Ayvaz, Connor Clark, and Michel Dumontier. 2017. Drug-drug interaction discovery and demystification using Semantic Web technologies. *Journal of the American Medical Informatics Association: JAMIA*, 24(3):556–564.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. *ArXiv*, abs/1906.05474.

Bethany Percha, Yael Garten, and Russ B. Altman. 2011. Discovery and explanation of drug-drug interactions via text mining. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 410–21.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Martin J. J. Ronis, Kim B. Pedersen, and James Watt. 2018. Adverse Effects of Nutraceuticals and Dietary Supplements. *Annual Review of Pharmacology and Toxicology*, 58:583–601.

Sunil Kumar Sahu and Ashish Anand. 2017. Drug-drug interaction extraction from biomedical text using long short term memory network. *Journal of biomedical informatics*, 86:15–24.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. Semeval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *SemEval@NAACL-HLT*.

Isabel Segura-Bedmar, Paloma Martínez, and César de Pablo-Sánchez. 2011. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC bioinformatics*, 12 Suppl 2:S1.

Linas Simonaitis and Gunther Schadow. 2010. Querying the national drug file reference terminology (ndfrt) to assign drugs to decision support categories. *Studies in health technology and informatics*, 160 Pt 2:1095–9.

Justin Spence, Monica Chintapenta, Hyanggi Irene Kwon, and Amie Taggart Blaszczyk. 2017. A Brief Review of Three Common Supplements Used in Alzheimer's Disease. *The Consultant Pharmacist: The Journal of the American Society of Consultant Pharmacists*, 32(7):412–414.

Alyssa A. Sprouse and Richard B. van Breemen. 2016. Pharmacokinetic Interactions between Drugs and Botanical Dietary Supplements. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, 44(2):162–171.

Johann Stan, Dina Demner-Fushman, Kin Wah Fung, Sonya E. Shooshan, Laritza Rodriguez, and Olivier Bodenreider. 2014. Title : A supervised machine learning framework for the extraction of drug-drug interactions from structured product labels.

Luis Tari, Saadat Anwar, Shanshan Liang, James Cai, and Chitta Baral. 2010. Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. volume 26, page i547–i553.

Luc Paul Frank Vaes and Leslie Hendeles. 2000. Interactions of warfarin with garlic, ginger, ginkgo, or ginseng: nature of the evidence. *The Annals of pharmacotherapy*, 34:1478–82.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Yefeng Wang, Terrence J. Adam, and Rui Zhang. 2016. Term Coverage of Dietary Supplements Ingredients in Product Labels. *AMIA Annual Symposium proceedings*, 2016:2053–2061.

Yefeng Wang, Divya R. Gunashekar, Terrence Adam, and Rui Zhang. 2017. Mining adverse events of dietary supplements from product labels by topic modeling. In *MedInfo*.

David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082.

Rui Zhang, Terrance J. Adam, Gyorgy Simon, Michael J. Cairelli, Thomas C. Rindflesch, Serguei V. S. Pakhomov, and Genevieve B. Melton. 2015. Mining biomedical literature to explore interactions between cancer drugs and dietary supplements. In *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*.

Yaoyun Zhang, Heng-Yi Wu, Jun Xu, Jingqi Wang, Ergin Soysal, Lang Li, and Hongwei Xu. 2016. Leveraging syntactic and semantic graph kernels to extract pharmacokinetic drug drug interactions from biomedical literature. volume 10, page 67.

Yijia Zhang, Wei Zheng, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier. 2018. Drug–drug interaction extraction via hierarchical rnns on sequence and shortest dependency paths. In *Bioinformatics*.

Wei Zheng, Hongfei Lin, Ling Luo, Zhehuan Zhao, Zhengguang Li, Yijia Zhang, Zhihao Yang, and Jian Wang. 2017. An attention-based effective neural model for drug-drug interactions extraction. In *BMC Bioinformatics*.

A Supplement and drug identifiers

We generate lists of supplement and drug entities based on UMLS Concept Unique Identifiers (CUIs) using a semi-automated method. For supplements, we identify NCI thesaurus (NCIT) concepts such as "Dietary Supplement" (NCIT: C1505, CUI: C0242295), "Vascular Plant" (NCIT: C14336, CUI: C0682475), and "Antioxidant" (NCIT: C275, CUI: C0003402) as likely parents of supplement terms. We recursively extract child entities of these parent classes from UMLS, deriving an initial list of supplements. To improve recall, we extract supplement names from the TRC Natural Medicines database,[9] perform fuzzy string matching to entities in UMLS, and add any identified CUIs to our list of supplements. The list is manually reviewed to remove non-supplement entities, those for which we could not identify any marketed supplement or medicinal uses. Following curation, we retain 2139 unique supplement entities.

Test dataset	Num. pairwise instances	RoBERTa-DDI (Trained on DDI-2013 and NLM-DailyMed)	RoBERTa-DDI (Trained on DDI-2013 only)
DDI-2013 (All)	5688	0.88	0.89
DDI-2013 (DrugBank)	5251	0.89	0.90
DDI-2013 (Medline)	437	0.73	0.77
NLM-DailyMed	927	0.84	0.70
All	6615	0.87	0.85

Table 3: F1-scores of RoBERTa-DDI trained using different training data. Test data contains all pairwise combinations of entities in test sentences.

Similarly, we generate a corresponding list of drug CUIs from parent entity "Pharmacologic Substance" (NCIT: C1909, CUI: C1254351) and any UMLS entity with a DrugBank identifier. Fuzzy name matching between drugs on drugs.com[10] and UMLS entities is used to identify drugs and experimental chemicals missed through UMLS search alone. Due to the significantly larger number of drugs compared to supplements, manual curation of this list is impractical at this time. This process generates a list of 15252 unique drug CUIs. Any entity that is identified as both a supplement and a drug is categorized exclusively as a supplement for the purposes of this work.

Similar supplement and drug entities are merged, such as those with overlapping names, e.g., entities corresponding to UMLS C0006675, C0006726, C0596235, and C3540037 all describe variants of Calcium and are merged under the supplement entity C3540037 ("Calcium Supplement"). The

[9] https://naturalmedicines.therapeuticresearch.com/

[10] https://drugs.com/

Model	Reference	F1 (classification)	F1 (detection)
Bi-LSTM (w/ max and attentive pooling)	Sahu and Anand (2017)	0.69 (macro-F1)	-
Hierarchical Bi-LSTM + Attention + dependency path	Zhang et al. (2018)	0.73 (unspecified)	-
Bi-LSTM (w/ attention and negative instance filtering)	Zheng et al. (2017)	0.77 (unspecified)	0.84
BioBERT embeddings	Chauhan et al. (2019)	0.72 (macro-F1)	0.87
BERT-large embeddings fine-tuned on DDI-2013	Peng et al. (2019)	0.79 (micro-F1)	-
RoBERTa-DDI fine-tuned on DDI-2013	-	0.82 (micro-F1)	0.89

Table 4: Baseline models for DDI detection and reported performance on the DDI-2013 test set. Results are shown for classification (5-way classification) and detection (binary classification).

canonical CUI representing a cluster is selected manually. Drug, supplement, and canonical mappings are provided in our data repository.

B DDI model performance

We train RoBERTa-DDI on a combination of DDI-2013 and NLM-DailyMed training data. In Table 3, we report the F1-scores of model variants on the test data. We show the performance of the final variant of RoBERTa-DDI (trained on both DDI-2013 and NLM-DailyMed) as well as a variant trained only on DDI-2013 training data (last column), which performs best on the DDI-2013 test set, but suffers when tested on NLM-DailyMed. We also further break down performance on the DrugBank and Medline sub-corpora within DDI-2013.

The DDI-2013 dataset is used as a benchmark dataset for DDI detection and classification, and is part of the BLUE benchmark suite (Peng et al., 2019). RoBERTa-DDI outperforms recently-reported SOTA performance on DDI detection in the DDI-2013 dataset using BioBERT (Lee et al., 2019) (F1 = 0.87) (Chauhan et al., 2019). Peng et al. (2019) also report SOTA performance on the DDI-2013 classification task, achieving 0.79 micro-F1 using a tuned BERT-large model. For comparison, we show the results of RoBERTa-DDI trained on DDI-2013 multi-class classification, which achieves 0.82 micro-F1 on DDI-2013 classification. We provide previously reported SOTA performance metrics on DDI-2013 in Table 4. We note that because the interaction classes are unbalanced in the DDI-2013 dataset, reported classification micro- and macro-F1-scores in previous work are not directly comparable.

The inclusion of the NLM-DailyMed corpus increases training data diversity and should improve generalization for the task of detecting SDI/SSI evidence. Thus, although RoBERTa-DDI trained on DDI-2013 has the highest performance on the DDI-2013 test set, RoBERTa-DDI trained over all training data performs the best overall, and we use this model variant to classify evidence for `SUPP.AI`.