Exploring Novel Computable Knowledge in Structured Drug Product Labels

Scott A. Malec, PhD¹, Richard D. Boyce, PhD¹

¹University of Pittsburgh Department of Biomedical Informatics, Pittsburgh, PA

Abstract

This paper introduces a database derived from Structured Product Labels (SPLs). SPLs are legally mandated snapshots containing information on all drugs released to market in the United States. Since publication is not required for pre-trial findings, we hypothesize that SPLs may contain knowledge absent in the literature, and hence “novel.” SemMedDB is an existing database of computable knowledge derived from the literature. If SPL content could be similarly transformed, novel clinically relevant assertions in the SPLs could be identified through comparison with SemMedDB. After we derive a database (containing 4,297,481 assertions), we compare the extracted content with SemMedDB for recent FDA drug approvals. We find that novelty between the SPLs and the literature is nuanced, due to the redundancy of SPLs. Highlighting areas for improvement and future work, we conclude that SPLs contain a wealth of novel knowledge relevant to research and complementary to the literature.

Introduction

Translational researchers have long recognized the value of integrative approaches to complex problems. The large-scale interpretation of multi-modal, heterogeneous, distributed sources of observational data often requires robust background information to help interpret and ultimately transform these data into new knowledge.¹⁻⁷ Researchers have exploited computable knowledge extracted from the literature in drug safety, drug repurposing, and oncology applications using Semantic MEDLINE, or SemMedDB.⁸⁻¹⁶ SemMedDB is a repository of structured knowledge extracted using a semantic interpreter of biomedical text. To build off the success of previous text mining projects, there is a pressing need to seek out new sources of knowledge. Another source of relevant, but yet to be mined pharmaceuticals-related knowledge, is content embedded in the narrative text of drug product labeling.

Drug product labeling standards are written into federal law and administered by the FDA. Since 2006, the Code of Federal Regulations has required submissions be sent to the FDA in an electronic format known as Structured Product Labeling (SPL).¹⁷ The SPL format is intended to make labels readable to both computers and humans. To that end, SPLs use a general technology standard called eXtensible Markup Language (XML). SPL has also been certified as a Health Level Seven International (HL7) standard for interoperability of electronic health information.

SPLs exist for all prescription and over-the-counter drugs approved for marketing in the United States. Each SPL summarizes knowledge about a drug based on pre-market studies and post-marketing information. These include information on: safety (e.g., black box warnings and reported adverse reactions), approved indications, clinical pharmacology, use in special populations, and drug-drug interactions. Because there is no legal requirement for pre-clinical or in vitro studies to be published, many of the knowledge claims summarized in SPLs might not be present in the published peer-reviewed biomedical literature. The purpose of this study is to describe knowledge claims present in SPLs, compare these knowledge claims with knowledge claims extracted from the literature, and determine the extent of novel knowledge in the SPLs. Another goal is to introduce and report on a newly created resource, called SemMedDB_SPL, that represents structured knowledge claims extracted from the SPLs of all prescription drugs currently marketed in the United States.

Background

Structured knowledge is knowledge that is in a computable format, meaning that the knowledge content in such a format that it is readable by computer programs. One convenient, computable representation of knowledge is that of a semantic predication. Semantic predications, also known as “triples,” consist of two concepts that relate to each other through some predicate (i.e., verb) such as “CAUSES” and “TREATS.”¹⁸⁻²⁰ For instance, “ibuprofen CAUSES gastrointestinal_hemorrhage” is one such semantic predication. Semantic predications have been referred to as the “atoms of thought.”²¹ In philosophy and cognitive science, these are referred to variously as propositions or assertions,
but in practice, the “proposition” refers to the normalized form (triple) and the assertion is the source sentence in the literature from which the semantic predication was derived.

SemRep is a symbolic natural language processing tool that was developed for the purposes of extracting, translating, and loading knowledge in the form of semantic predications by researchers the National Library of Medicine. SemRep was used to build Semantic MEDLINE (SemMedDB). SemMedDB stores structured knowledge extracted from titles and abstracts of peer-reviewed biomedical literature stored in MEDLINE. A number of research studies have used the structured knowledge including drug safety and drug repositioning.

Another potential source of relevant knowledge is embedded in the narrative text of SPLs. Although there have been several studies that have used natural language to extract specific kinds of knowledge, such as indications and adverse drug reactions, from SPLs, no studies to date have applied SemRep to extract a broad range of semantic predications. Because many SPLs report knowledge from unpublished pre-market studies, we hypothesize that semantic predications present in SPLs will complement those extracted from the scientific literature. Our basic assumption is that the extent of “novel” structured knowledge may be measured using the number of semantic predications captured in the SPLs but absent among the predications extracted from the literature. In this study, we report on a pipeline for extracting semantic predications from SPLs. Semantic predications extracted by the pipeline are stored in a new resource that we call SemMedDB_SPL. We describe the new resource and report on its coverage for several newly approved drugs relative to the extracted knowledge in the latest version of SemMedDB.

Methods

Extracting knowledge from SPLs. SPLs for prescription drugs were manually downloaded from the DailyMed website hosted by the National Library of Medicine. A custom parser that we developed for SPLs translated the XML format into a relational database. SPL sections are coded using LOINC and contain a mix of narrative and tables tagged with HTML. In order to run SemRep on SPL content, we exported the text of selected sections by querying the relational database using LOINC codes, saving the query results into individual text files, and then parsing the text files to separate the narrative text from table content. Only the narrative content was processed using SemRep. We processed text from the following SPL sections: adverse reactions, boxed warning, clinical pharmacology, clinical studies, contraindications, description, dosage and administration, drug interactions, how supplied, inactive ingredients, indications and usage, overdosage, precautions, and use in specific populations. The process of extracting structured knowledge in the form of semantic predications is illustrated below in (Figure 1).

![Figure 1: Workflow diagram of structured knowledge extraction process running the SemRep NLP system. The steps were as follows: 1.) download the SPL XML files from DailyMed, 2.) parse the XML and populate a relational (SQL) database, 3.) separate SPL narrative text from table content, 4.) run SemRep NLP to extract semantic predication knowledge content, and 5.) load resulting predications into a SQL database for further analysis.](image)

To extract meaningful assertions, we ran SemRep version 1.8 on the narrative text extracted from the SPLs. SemRep was run in batch mode with the anaphora flag activated (to help assign meaning to assertions with pronouns) using the UMLS 2018AB dictionary on a server running Ubuntu Linux v.16.04 (64 GB RAM, four-core Xeon processors). Finally, we extracted predications from the raw SemRep output, along with the source sentences and accompanying metadata. The final data was stored in a relational database (Postgres v11.5-1). We used a combination of python, SQL, shell scripts (bash), and R scripts to implement the experiments in the present study.

404
Figure 2: Tables in the SemMedDB_SPL schema that contains computable knowledge in semantic predication format. The predications table from SemMedDB was also included to facilitate rapid comparison.

Figure 2 shows the two core content tables in the new database. For the sake of providing convenient access for querying the newly extracted SPL structured knowledge and SemMedDB, we also added the predications table (smdbpredications in Figure 2) from SemMedDB version 40R (with SemRep version 1.8 run without anaphora resolution, released July 10, 2019). The database also includes a SPL metadata table (structuredProductLabelMetadata) to make its simple to navigate between predications and the source files.

We then wrote queries to explore the features of the new resource and test if it held knowledge indicative of gaps in the peer-reviewed literature that might be of clinical interest. Our analysis proceeded as follows:

1. **Exploratory analysis of structured knowledge in SPLs.** We collected statistics (counts, frequencies) and explored distributional information by generating visual summaries. We generated heatmap visualizations normalized by total predication count of the distribution of predicate types by SPL section type.

2. **Examination of novel content.** To find novel structured knowledge, we ran SQL queries that filter out novel SPL-derived predications from the pre-existing, literature-derived predications in SemMedDB.

3. **Investigate comparative coverage of newly released drugs.** If there are new drugs in the structured knowledge extracted from the SPLs, it may be helpful to use the coverage of these new drugs as a metric of comparison of literature-derived versus SPL-derived structured knowledge. Since there is coverage in both the literature and the SPLs, it is convenient to use SQL to discover what is potentially novel in the SPLs.

The first two analyses are general analysis and without specific inclusion criteria. For the third analysis, we only examined the drugs with indications approved by the FDA between January 2015-August 2019. Drugs had to have at least one mention in the subject position of a predication in SemMedDB and also have at least one mention in the subject position of predicates in SemMedDB_SPL. This was to ensure that at least some structured knowledge extracted from the scientific literature existed in SemMedDB. Finally, to help us focus on predications where drugs are causal agents, drug mentions had to be in the subject position of predications.

**Results**

We extracted 4,297,481 semantic predications from the SPLs (373,061 SPL sections for 37,749 pharmaceutical therapies). The clinical pharmacology and precautions SPL sections yielded the most predications (Table 1). SemRep extracted predications with 62 predicate types (some of these are negated, so “TREATS” becomes “NEG_TREATS”). The five most common predicate types in general were “PROCESS_OF” (n = 986,243), “TREATS” (744,794), “ISA” (366,616), “ADMINISTERED_TO” (287,794) and “LOCATION_OF” (273,879). Counts for clinically relevant predicates include “CAUSES” (172,654), “AFFECTS” (115,798), “INHIBITS” (98,486), “PREDISPOSES” (54,270), “STIMULATES” (49,193) and “COMPLICATES” (7,668). 86 SPL sections died or errored during processing. An examination of the logs revealed the culprits for these processing failures. Specifically, SemRep timed out while processing unruly strings such as expansive lists of adverse events and unruly chemical names burdened with extensive diacritical marks.
Table 1: Statistics for the semantic predications extracted by SemRep from SPLs by SPL section type.

<table>
<thead>
<tr>
<th>Section Name</th>
<th>Section count</th>
<th># Predications</th>
<th># of Predications/Section</th>
<th># Unique Predications</th>
</tr>
</thead>
<tbody>
<tr>
<td>adverse_reactions</td>
<td>29,259</td>
<td>519,884</td>
<td>17.77</td>
<td>21,595</td>
</tr>
<tr>
<td>boxed_warning</td>
<td>10,490</td>
<td>104,463</td>
<td>9.96</td>
<td>4,137</td>
</tr>
<tr>
<td>clinical_pharmacology</td>
<td>31,228</td>
<td>781,157</td>
<td>25.01</td>
<td>38,263</td>
</tr>
<tr>
<td>clinical_studies</td>
<td>15,818</td>
<td>345,019</td>
<td>21.81</td>
<td>17,547</td>
</tr>
<tr>
<td>Contraindications</td>
<td>26,770</td>
<td>145,742</td>
<td>5.44</td>
<td>7,624</td>
</tr>
<tr>
<td>Description</td>
<td>27,638</td>
<td>91,846</td>
<td>3.22</td>
<td>7,811</td>
</tr>
<tr>
<td>dosage_and_administration</td>
<td>26,598</td>
<td>275,063</td>
<td>10.34</td>
<td>13,886</td>
</tr>
<tr>
<td>drug_interactions</td>
<td>21,951</td>
<td>350,929</td>
<td>15.99</td>
<td>16,125</td>
</tr>
<tr>
<td>how_supplied</td>
<td>3,506</td>
<td>8,492</td>
<td>2.42</td>
<td>2,317</td>
</tr>
<tr>
<td>inactive_ingredient</td>
<td>90</td>
<td>189</td>
<td>2.1</td>
<td>73</td>
</tr>
<tr>
<td>indications_and_usage</td>
<td>29,612</td>
<td>298,604</td>
<td>10.08</td>
<td>15,111</td>
</tr>
<tr>
<td>Overdosage</td>
<td>24,766</td>
<td>145,883</td>
<td>5.89</td>
<td>4,716</td>
</tr>
<tr>
<td>Precautions</td>
<td>18,623</td>
<td>857,711</td>
<td>46.06</td>
<td>29,515</td>
</tr>
<tr>
<td>use_in_specific_populations</td>
<td>12,960</td>
<td>372,489</td>
<td>28.74</td>
<td>20,050</td>
</tr>
</tbody>
</table>

To narrow down the range of predicate types to analyze, we selected a subset of predicate types that most frequently occur with a pharmaceutical substance subject semantic type (“phsu”) or disease/syndrome object semantic type (“dsyn”), finding these semantic types to be useful starting points for drug safety and drug repurposing use cases. The predicates we included in our analysis were “TREATS,” “PREVENTS,” “PREDISPOSES,” “CAUSES,” and “INTERACTS_WITH,” among others. The full list of predicates chosen are reflected in the labels along the y-axes of the Figures 3 and 4.

Figure 3: This heatmap illustrates the relative distribution of predicates (along the y-axis) in each SPL section-type (x-axis). Red indicates a higher predicate frequency with lighter colors to white, indicating lower frequency by predicate type. The numbers in the legend indicate predicate counts on a natural log scale.
Examination of novel content. We analyzed SemMedDB_SPL to determine the extent of novel content. The heatmap in Figure 4 illustrates the breakdown of novel content by SPL section. The precautions and clinical pharmacology SPL sections (followed by the drug_interactions SPL section) were notable for having substantially more novel predications than other SPL sections. Of the 4,297,481 total predications, only 142,311 of the semantic predications are unique. Each predication is repeated an average of 30.20 times. By comparison, the current release of SemMedDB (version 40R) has 97,972,561 semantic predications (extracted from 29,115,337 abstracts), with 19,836,608 of them being unique (with each predication repeated an average of 4.86 times).36

Investigate comparative coverage of newly released drugs. There was a total of 183 drugs in the list of drugs released in the last five years by the FDA. Of the 103 drugs from that list that were in SemMedDB_SPL, only 19 were also found in SemMedDB. Samples of novel predications for newly released drugs are enumerated below as a listing in Figure 5 and a summary of the drug coverage is provided in Table 2.

Cannabidiol: coexists_with cannabis_sativa_plant
Ivabradine: causes toxic_effect, interacts/neg_interacts with metformin, causes teratogenesis, affects abnormal cardiac conduction
Daclizumab: isa antibodies,_monoclonal,_humanized, interacts_with protein_subunits, affects lactation_established_(finding), affects immune
Deflazacort: coexists_with acetic_acid, interacts_with ppg_gene, coexists_with enflaza, coexists_with methylene chloride, neg_inhibits abc2_gene, neg_inhibits ppg_gene
Mepolizumab: interacts_with il5_protein,_human, affects fetus, neg_augments adverse_effects, affects lactation_established_(finding), treats asthma, compares_with placebos, isa antibodies,_monoclonal,_humanized, treats infant
Prucalopride: treats health_volunteers, treats social_group, neg_higher_than antineoplaston_a10, interacts_with htr2a gene, inhibits cytochrome_p450, treats old_age, affects decreased_renal_function, affects renal_function, compared_with antineoplaston_a10
Striptentol: affects lactation_established_(finding)
Fibanserin: treats renal_insufficiency, treats cyp2d6_poor_metabolizer, coexists_with paroxetine, treats health_volunteers, stimulates infed, inhibits infed, causes syncope, affects drug_kinetics, treats patients, causes hypotension_severe, stimulates digoxin, treats hepatic_impairment, treats cyp2c9_poor_metabolizer, interacts_with nr4a2_protein,_human
Sulfinamide: isa monoamine_oxidase_b_inhibitor_(product), coexists_with water, neg_isa proteins, coexists_with methanol, isa acids, causes discoloration_of_skin, coexists_with dimethyl sulfoxide, inhibits abc2_g2_protein,_human, affects drug_kinetics, affects lactation_established_(finding), treats hepatic_impairment, affects drug_kinetics, affects lactation_established_(finding)
Secnidazole: coexists_with smad4-as1_gene, affects qtc_interval, neg_treats bacteria, affects lactation_established_(finding)
Tafenoquine: inhibits membrane_transport_proteins, affects/neg_affects qtc_interval, inhibits pou2f2_protein,_human, inhibits slc47a1 gene, neg_interacts_with sloc1b1 gene, causes unspecified_abortion, inhibits metformin, interacts_with proteins, affects drug_kinetics, neg_interacts_with sloc1b1 gene, interacts_with p-glycoproteins, causes toxic_effect

Figure 5: This listing outlines a subset of novel predications extracted from the SPLs for drugs released by the FDA to market (January 2015-August 2019) that also had coverage in SemMedDB. The subject (pharmaceutical drug) of each predicate is underlined with a colon at the beginning, followed by each novel predicate and object, respectively. Potential side-effects and drug-gene mechanisms of strong pharmacogenomic importance have been bolded.
Table 2: Coverage of drugs recently released by the FDA (January 2015 - August 2019) by # of predications per UMLS CUI. Search performed by querying the predications table using the subject_name string would yield slightly different results.

<table>
<thead>
<tr>
<th>Drug name</th>
<th># in SemMedDB (unique indexed articles)</th>
<th># in SemMedDB_SPL (unique SPLs)</th>
<th># Novel Predications (unique SPLs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cannabidiol (CUI: C0006863)</td>
<td>2797 (1142)</td>
<td>2 (1)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Ivabradine (CUI: C0257190)</td>
<td>1800 (651)</td>
<td>7 (2)</td>
<td>6 (2)</td>
</tr>
<tr>
<td>Daclizumab (CUI: C0663182)</td>
<td>1477 (657)</td>
<td>7 (2)</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Deflazacort (CUI: C0057258)</td>
<td>668 (256)</td>
<td>9 (2)</td>
<td>7 (2)</td>
</tr>
<tr>
<td>Mepolizumab (CUI: C0969324)</td>
<td>548 (185)</td>
<td>14 (1)</td>
<td>9 (1)</td>
</tr>
<tr>
<td>Prucalopride (CUI: C0913506)</td>
<td>460 (160)</td>
<td>13 (1)</td>
<td>10 (1)</td>
</tr>
<tr>
<td>Stiripentol (CUI: C0075262)</td>
<td>308 (112)</td>
<td>1 (1)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Flibanserin (CUI: C0754280)</td>
<td>245 (80)</td>
<td>18 (3)</td>
<td>15 (2)</td>
</tr>
<tr>
<td>Safinamide (CUI: C1098261)</td>
<td>166 (64)</td>
<td>12 (1)</td>
<td>8 (1)</td>
</tr>
<tr>
<td>Secnidazole (CUI: C0074246)</td>
<td>148 (65)</td>
<td>7 (2)</td>
<td>6 (2)</td>
</tr>
<tr>
<td>Tafenoquine (CUI: C0903411)</td>
<td>144 (61)</td>
<td>22 (2)</td>
<td>19 (2)</td>
</tr>
</tbody>
</table>

Our query of SemMedDB_SPL for as yet undiscovered knowledge for the newly released drugs yielded many predications of potential interest, many of which are listed in Figure 5. We compared the coverage for older versus newer statins by counting unique predications with the statin name in the subject position. We found that approximately half of the predications were novel for the older statins (Simvastatin [55% or 60 of 108 predications] and atorvastatin [54.8% or 46 of 84]), but that all of the predications were novel for the newer statins (alirocumab [“Praluauten”] and evolocumab [“Repartha”]), with 10 and 9 distinct mentions in the subject position, respectively.

Discussion

SPLs are essentially snapshots of all knowledge for all drugs that have approval for marketing in the United States. SPLs are continuously updated, as product labels are legally required to be up to date. The model provided for the current release of SemMedDB_SPL shows the feasibility of automatically extracting predications from SPLs. We hypothesized that there was evidence in the SPLs that were not present in the literature. We found that across the sections that there are novel predications that are present in the SPLs but not in the predications extracted from the indexed biomedical literature. The distribution of predication types by SPL section was uneven. The yield of predications extracted varied considerably by section type, with precautions, use_in_specific_populations, and clinical pharmacology being the most fecund, as per Table 1 above, while other sections yield substantially fewer predications.

We also found that many predications are of potential clinical relevance as per Figure 5. The relational table structure enables linking from predications to the source texts in the sentences table. For example, the semantic predication “flibanserin CAUSES syncope” links to the following source sentence retrieved from the box warning section: “The concomitant use of ADDYI and moderate or strong CYP3A4 inhibitors increases flibanserin concentrations, which can cause severe hypotension and syncope [(see CONTRAINDICATIONS and WARNINGS)].”

Novelty and redundancy. We defined novelty as the presence structured knowledge present in the SPLs (SemMedDB_SPL) but absent in the structured knowledge extracted from the literature (SemMedDB). We tested our assumption that the SemRep NLP system could accurately map from detected concept unique identifier or CUI in the
UMLS to concept mentions in the SPL narrative text and in a different process on the accuracy of concept identification in the indexed peer-reviewed biomedical literature.

As long as SemRep maps to the correct concept unique identifier or CUI in the UMLS that encodes those strings in the subjects of semantic predications, we were be able to compare the two knowledge sources. However, there may be variants that are not in the UMLS that we might have missed.

Novelty between the literature and SPLs is nuanced: some sections might hold statements that are more precautionary than scientific.32 The frequent repetition of predications may indicate that the predications derived from SPLs may be more redundant than those derived from the literature. There is known to be a many to many relationship between SPLs and drugs. Accordingly, SPLs for each product can be repeated many times. Since we did not select canonical labels for each drug, much of the duplication that has been noted is likely due to repeat narrative in the SPLs. Repetition and redundancy need to be addressed. This could be resolved by picking a canonical name for each drug to address this issue. The “structured” part of product labels may be more aspirational than a verdict. For example, predications having a “CAUSES” predicate extracted from precaution sections or black box warning sections may be uninformative, as causal predications in such sections may be issued by pharmaceutical companies as blanket coverage to indemnify themselves against lawsuits springing from potential side-effects.

We found that the distribution of predication types by the SPL section type is uneven. The yield of predications extracted using SemRep across section types varied substantially, with precautions, use_in_specific_populations, and clinical pharmacology being the most fecund, as per Table 1 and Figures 3 and 4 above.

**Limitations and directions for future work.** As a pilot project toward a more comprehensive biomedical database of structured causal knowledge and with such a crude measure of novelty, this work is subject to several limitations. First, we extracted the data using plain text output instead of the more informative XML output. The XML output contains metadata concerning the confidence of the predication. Such information could help inform machine learning models that incorporate information describing the degree of belief. Second, we were puzzled initially why so many recently released drugs were missing from SemMedDB. We performed a PubMed search for several of the missing drugs and associated side-effects, and found publications that mention drugs absent in SemMedDB. One explanation is that the SemMedDB_SPL and SemMedDB are not strictly comparable: in the process of constructing SemMedDB, the 2006AA version of the UMLS lexicon was applied without anaphora resolution, whereas the 2018AB version of the UMLS lexicon was applied to the SPLs with anaphora resolution. Forthcoming analyses should address this issue by using the same lexicons. We also suspect that much knowledge is missing because of the NLP: SemRep is noted for high precision, but low recall.41 NLP both facilitates and limits text the helpfulness of research. As methods for extracting knowledge from the free text with high confidence improves, we can expect that efforts in this domain to improve accordingly, particularly with efforts afoot to improve extraction recall for causal language.42 Finally, we analyzed only the latest SPLs available. One would expect that the longer a drug has been on the literature, the less novelty. In future work, we hope to track the accretion of new knowledge longitudinally by analyzing archived versions of the SPLs. In this way, we could take snapshots of what was known about the drug historically at various time points.

**Potential applications.** The schema is under revision to address longitudinal (archived) labeling information. Other potential applications of SemMedDB_SPL include literature-based discovery (LBD), causal feature selection for statistical and causal graphical modeling, and knowledge engineering. Recent developments in LBD support ingesting structured knowledge to help researchers generate hypotheses about potential therapies.12,13,24,43

**Causal learning.** Mathematical formalisms called graphical causal models have emerged that can learn causal structures from observational data.44–46 Observing that conditional dependence (and independence) results from causal relationships, causal graphical modeling methods work in the opposite direction, by learning dependencies to infer causal relationships.44,47 However, these methods often perform better when domain knowledge is available.1,14,21 However, the domain expertise of live human experts cannot scale. SemMedDB_SPL could be exploited as a
component in a translational pipeline to stand in for theoretical domain expertise to help interpret retrospective observational data.

Discovering contradictions. Though not the focus on the present paper, research into contradictions and scientific retraction are active areas of research.\textsuperscript{49–51} In particular, contradictory knowledge claims could be a productive source of leads for hypothesis generation. Researchers may ask, for example: do findings from pre-clinical trials contradict those in the peer-reviewed literature? More specifically, are there cases where the pre-trial findings indicate a “CAUSES” predicate and the literature report a “PREVENT” or “NEG_CAUSES”?

More comprehensive qualitative, and quantitative, analysis of the “novel” content remains. For example, are the extracted predications correct? Are some static predicates such as “ISA” or “PROCESS_OF” really relevant for drug safety or drug repurposing purposes? Are any of the novel relationships are mentioned in the literature but missed by SemRep? Is any of the missing knowledge available in other structured knowledge resources?

Conclusion

The present work presents an initial pilot study intended to create a new resource of structured knowledge for use and reuse by other researchers in the area of drug repurposing, drug safety, and drug discovery. We have provided the code we used publicly so that other researchers can reproduce and build upon our efforts to elucidate novel or clinically relevant content contained in SPL text narratives.\textsuperscript{32} We intend for SemMedDB\textsubscript{SPL} to contribute critical structured knowledge to pharmaceuticals-based research pipelines, and to complement existing structured knowledge resources such as SemMedDB. To that end, we have made the materials created for this paper publicly available at http://github.com/dbmi-pitt/SemMedDB\textsubscript{SPL}.

Acknowledgments

This work is support by a training grant from the University of Pittsburgh Department of Biomedical Informatics grant T15LM007059. I want to thank Harry Hochheiser of the University of Pittsburgh DBMI and Halil Kilicoglu of the University of Illinois Urbana-Champaign for their thoughtful feedback on earlier drafts of this manuscript.

References


14. Malec S. Using the Literature to Identify Confounders. UT SBMI Dissertations (Open Access) [Internet]. 2018 Jan 1; Available from: https://digitalcommons.library.tmc.edu/uthshis_dissertations/41


52. Alamri A. The Detection of Contradictory Claims in Biomedical Abstracts [Internet] [phd]. University of Sheffield; 2016 [cited 2019 Aug 15]. Available from: http://etheses.whiterose.ac.uk/15893/


56. Alamri A. The Detection of Contradictory Claims in Biomedical Abstracts [Internet] [phd]. University of Sheffield; 2016 [cited 2019 Aug 15]. Available from: http://etheses.whiterose.ac.uk/15893/