Original Paper

A Knowledge Graph of Combined Drug Therapies Using Semantic Predications From Biomedical Literature: Algorithm Development

Jian Du[1*], PhD; Xiaoying Li[2*], PhD

[1]National Institute of Health Data Science, Peking University, Beijing, China

[2]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

[*]all authors contributed equally

Corresponding Author:
Xiaoying Li, PhD
Institute of Medical Information
Chinese Academy of Medical Sciences
No 69, Dongdan North Street
Dongcheng District
Beijing, 100005
China
Phone: 86 10 52328792
Email: lixiaoying@imicams.ac.cn

Abstract

Background: Combination therapy plays an important role in the effective treatment of malignant neoplasms and precision medicine. Numerous clinical studies have been carried out to investigate combination drug therapies. Automated knowledge discovery of these combinations and their graphic representation in knowledge graphs will enable pattern recognition and identification of drug combinations used to treat a specific type of cancer, improve drug efficacy and treatment of human disorders.

Objective: This paper aims to develop an automated, visual approach to discover knowledge about combination therapies from biomedical literature, especially from those studies with high-level evidence such as clinical trial reports and clinical practice guidelines.

Methods: Based on semantic predications, which consist of a triple structure of subject-predicate-object (SPO), we proposed an automated algorithm to discover knowledge of combination drug therapies using the following rules: 1) two or more semantic predications (S_1-P-O and S_i-P-O, i = 2, 3…) can be extracted from one conclusive claim (sentence) in the abstract of a given publication, and 2) these predications have an identical predicate (that closely relates to human disease treatment, eg, “treat”) and object (eg, disease name) but different subjects (eg, drug names). A customized knowledge graph organizes and visualizes these combinations, improving the traditional semantic triples. After automatic filtering of broad concepts such as “pharmacologic actions” and generic disease names, a set of combination drug therapies were identified and characterized through manual interpretation.

Results: We retrieved 22,263 clinical trial reports and 31 clinical practice guidelines from PubMed abstracts by searching “antineoplastic agents” for drug restriction (published between Jan 2009 and Oct 2019). There were 15,603 conclusive claims locally parsed using the search terms “conclusion*” and “conclude*” ready for semantic predications extraction by SemRep, and 325 candidate groups of semantic predications about combined medications were automatically discovered within 316 conclusive claims. Based on manual analysis, we determined that 255/316 claims (78.46%) were accurately identified as describing combination therapies and adopted these to construct the customized knowledge graph. We also identified two categories (and 4 subcategories) to characterize the inaccurate results: limitations of SemRep and limitations of proposal. We further learned the predominant patterns of drug combinations based on mechanism of action for new combined medication studies and discovered 4 obvious markers (“combin*,” “coadministration,” “co-administered,” and “regimen”) to identify potential combination therapies to enable development of a machine learning algorithm.

Conclusions: Semantic predications from conclusive claims in the biomedical literature can be used to support automated knowledge discovery and knowledge graph construction for combination therapies. A machine learning approach is warranted to take full advantage of the identified markers and other contextual features.

(JMIR Med Inform 2020;8(4):e18323) doi: 10.2196/18323

KEYWORDS

combined drug therapy; knowledge graph; knowledge discovery; semantic predications

Introduction

Background

Combination drug therapy is a therapeutic intervention in which multiple drugs are administered, particularly in patients with malignant neoplasms [1,2]. Compared with single-agent therapy, the synergistic interaction of combined medications significantly improves drug efficacy, shortens disease course, delays or avoids drug resistance, and reduces both toxicity and other side effects without loss of efficacy. The combination of several existing drugs with compatible mechanisms of action has been reported as an alternative approach to advance the success of drug repositioning [3]. The characteristics of combination therapies make them a practical alternative to standard approaches, with the potential to save billions of dollars on research and development of new drugs, particularly in the absence of effective monotherapies for many types of cancer and other diseases (such as autoimmune and psychiatric conditions), and more than 6700 rare diseases for which no therapies are available [3].

In recent decades, massive efforts have been made to employ combined therapeutic agents to improve treatment of human disorders such as specific cancers [2,4], malignancies such as lymphocytic leukemia [1], and hypertension [5]. PubMed houses over 175,000 publications found by searching the MeSH (Medical Subject Headings) heading "Drug Therapy, Combination" (Jan 2009 to Oct 2019). We used innovative information retrieval and semantic web technologies to discover knowledge about therapeutic drug combinations, then presented the findings in a visually intuitive knowledge graph. The resulting knowledge graph will not only support machine-understandable information for curing disease and drug efficacy screening, but also provide insights to quickly develop new therapies for untreated diseases.

In this paper, we propose a systematic, automated approach to discover knowledge about combination drug therapies in the biomedical literature (especially clinical trial reports and clinical practice guidelines with high evidence levels), and integrate the findings into knowledge graphs with customized organization and visualization. This entails the following:

1. Propose an automated algorithm to discover knowledge about combination drug therapies based on semantic predications extracted from conclusive claims in biomedical literature
2. Customize a knowledge graph to emphasize the specified drugs being combined rather than traditional triples (eg, one drug TREATS one disease)
3. Retrieve published clinical trial reports and clinical practice guidelines for algorithm verification and validation, followed by manual identification of accurate knowledge about combination drug therapies, as well as interpretation of inaccurate findings
4. Characterize the major patterns of combinations according to mechanism of action for new combined medication studies and identify potential markers as key features for machine learning-based drug combination discovery.

In the following sections, we review related work on knowledge graphs and drug-disease knowledge discovery. We then present our methodology to develop an automated algorithm to discover knowledge about combination drug therapies. A large number of clinical trial reports and clinical practice guidelines were retrieved from PubMed for algorithm verification and validation, followed by manual biocuration to verify accurate results for knowledge graph construction and to interpret inaccurate results. In the discussion we characterize the main patterns of drug combinations according to their mechanisms of action to inform new combination studies and identify markers of potential combined drug therapies to inform machine learning–based algorithm development.

Related Work

Knowledge Graph

A knowledge graph is a network-based representation of the semantic relationship between entities. Its principles have been developed by industry and academia, particularly by the semantic web community. In 1982, Hoede and Stokman used large graphs to represent knowledge extracted from medical and sociology texts [6], resulting in an expert system for quick searching and decision support for automated queries. In 2012, Google formally introduced their knowledge graph after compiling over 3.5 billion facts and relationships among 500 million objects, which is essentially a semantic enhancement of the search engine to help search real-world objects quickly and easily. At the end of 2016, Microsoft announced a large graph of concepts harnessed from billions of web pages and search logs for short text understanding, called the Concept Graph. Other frequently mentioned applications are Yahoo Spark, Facebook's entity graph, Wikidata, Freebase, Baidu's Knowledge Graph, and Sogou's Knowledge Cube. Although these products differ in their architecture, operational purpose, and supported technologies, they constitute a family of knowledge graphs and together represent the precursor to a new generation of semantic search and knowledge discovery.

Many other studies on biomedical knowledge graphs have been performed since 2012, playing an indispensable role in biomedical knowledge services. Remarkable achievements encompass the organization of health information from heterogeneous textual [7], disease-symptom association learning from electronic medical records [8], presenting relationships between cells and cytokines [9], extraction of human disorder biomarkers [10], and predicting drug efficacy [11]. However, knowledge graphs have not yet been applied to organize and manage biomedical information related to combination drug therapies, especially when such knowledge comes from the direct empirical evidence of clinical research.

Biomedical Drug-Disease Knowledge Discovery

Studies on biomedical knowledge discovery mainly focus on the semantic relationships, associations, and interactions between biomedical entities such as diseases, drugs, signs or symptoms, target organ, genes, biomarkers, and targets. One of the most important tasks is to identify the exact relationship between a drug and disease, especially for "treatment." Many information retrieval techniques and methods have been used to approach this problem based on predefined rules [12,13] or natural language processing [14-19] combined with machining learning [17-19]. Although predefined rules offer promising precision from biomedical texts, they are insufficient and perform poorly when parsing big data due to the noisy and variable syntactic structures within large-scale scientific texts. In comparison, natural language processing-based algorithms have generally been more successful and relatively flexible by virtue of features that parse context in literature.

Semantic Knowledge Representation, or SemRep, is a natural language processing tool based on the Unified Medical Language System (UMLS) [20]. This high-quality tool for extracted semantic predication has already been utilized for a broad range of applications such as the construction of a biomedical knowledge graph [21], identification of apparent contradictions [22], labeling for semantic relationships [23], and detection of drug-drug interactions [24] or drug-gene targets [25]. Here, we extend the application scope of SemRep by using semantic predications from conclusive sentences (eg, the conclusion section) of abstracts in biomedical literature, rather than the whole abstract, to automatically discover knowledge about combination drug therapies. The conclusion statement of a paper is the essential knowledge unit that synthesizes the knowledge content of an article and is validated by the experiment reported within the article.

Methods

Using Conclusive Sentences in the Abstract of a Publication as Knowledge Claims

There is a vast amount of published biomedical literature easily available in digital and printed format due to the rapid advance of information technology. For example, the cumulative citations of PubMed resources have exceeded 25 million, expanding with an annual growth of 0.9 million [26]. The huge amount of literature encourages the emergence of automated knowledge discovery, which could help scientists keep up with the latest scientific developments and academic achievements.

Scientific publications can be considered records of knowledge claims on a research question, supported by empirical evidence. These knowledge claims are often succinctly described in the abstract of a publication. The abstract is the most frequently accessed section of a publication and the only section used as source information in indexing databases such as PubMed. In this study, we parsed abstracts from PubMed for conclusive claims identified by the key words "conclusion*" and "conclude*" (Table 1) in order to discover knowledge about combination drug therapies.

Semantic Predication Interpretation Using SemRep

SemRep is a well-developed semantic knowledge interpreter that retrieves semantic predications (in terms of subject-predicate-object) to extract information from biomedical texts. For example, for the first claim in Table 1, SemRep would interpret the 7 semantic predications shown in Table 2, and the predications with "INFER" in the predicate was inferred based on two existing predications.

As a natural language processing driven tool, SemRep takes full advantage of UMLS knowledge sources including the Metathesaurus and Semantic Network. Briefly, the subject and object of semantic predication returned by SemRep are the preferred names of biomedical concepts in the UMLS Metathesaurus, while the predicates were derived from semantic relationships in the UMLS Semantic Network. An evaluation based on sample data with semantic type "Chemicals and Drugs" has allowed SemRep to achieve a promising degree of precision (83%) [20], which will contribute to the development of algorithms for automated knowledge discovery for combination drug therapy.

Table 1. Examples of conclusive claims from PubMed abstracts.

PMID_Ab[a]	Claim
19322566.ab.15	*CONCLUSION*: A combination of GTI-2040, capecitabine and oxaliplatin is feasible in patients with advanced solid tumors.
28101592.ab.10	In *conclusion*, FCM regimen allows excellent long-lasting response in previously untreated patients with FL.
21198717.ab.10	WHAT IS NEW AND *CONCLUSION*: The use of novel agents such as thalidomide, bortezomib and lenalidomide for RRMM is highly prevalent in France from the first relapse.
23197589.ab.8	We *conclude* that intraventricular rituximab in combination with MTX is feasible and highly active in the treatment of drug-resistant CNS NHL that is refractory or unresponsive to IV rituximab.

[a]PMID_Ab: PubMed reference number, abstract, sentence in which the information appears.

Table 2. Examples of SemRep semantic predications based on a biomedical claim.

Example claim	Predicate	Object
19322566.ab.15 CONCLUSION: A combination of *GTI-2040*, *capecitabine* and *oxaliplatin* is feasible in patients with *advanced solid tumors*.		
Advanced Malignant Solid Neoplasm	PROCESS_OF	Patients
GTI2040	TREATS	Patients
GTI2040	*TREATS(INFER)*	*Advanced Malignant Solid Neoplasm*
capecitabine	TREATS	Patients
capecitabine	*TREATS(INFER)*	*Advanced Malignant Solid Neoplasm*
oxaliplatin	TREATS	Patients
oxaliplatin	*TREATS(INFER)*	*Advanced Malignant Solid Neoplasm*

Development of an Algorithm for Discovering Knowledge About Combination Drug Therapy

The UMLS-based SemRep underpins biomedical knowledge discovery applications with its broad coverage and high-quality extracted semantic predications. SemRep enables interpretation of 30 semantic predicates [27], such as "PREVENTS," "TREATS," and "INHIBITS."

To develop our algorithm to automatically discover knowledge about combination drug therapies, we focused on 4 semantic predicates closely related to disease treatment: "TREATS," "INHIBITS," "PREVENTS," and "DISRUPTS" (also inferences with "INFER" such as "TREATS(INFER)"). We also adopted the UMLS Semantic Types "Chemicals and Drugs," "Disease or Syndrome," and their child types to restrict the subject and object of SemRep output to drug and disease.

Knowledge about combined drug therapy is detected under the hypothesis that (1) two or more semantic predications (S_1-P-O and S_i-P-O, i=2, 3...) are extracted from one conclusive claim in the abstract of a given biomedical publication, and (2) they have an identical object (eg, disease) and predicate (eg, treats) but different subjects (eg, drugs). Referring again to the example used in Table 2, the method provided straightforward discovery of the combined medication knowledge "GTI2040+capecitabine+oxaliplatin-TREATS-Advanced Malignant Solid Neoplasm."

Generally, the algorithm could be expressed by the following formula (Textbox 1):

Textbox 1. Algorithm text.

Algorithm: Drug combination knowledge discovery

Input: Semantic predications S_1-P-O and S_i-P-O (i=2, 3...) from one conclusive claim in a biomedical abstract

Output: Combined drug therapy knowledge S_1+S_i-P-O, where all of the following conditions are satisfied:

1. P∈ {TREATS","INHIBITS","PREVENTS","DISRUPTS"}
2. S_1∈ Chemicals and Drugs
3. S_i∈ Chemicals and Drugs, i≥2
4. O∈ Disease

Automated Filtering to Focus on Specific Drug and Disease Names

Knowledge about combined drug therapies primarily pertains to specified drugs and diseases; thus, the generic names of these biomedical entities should be filtered out automatically.

Filtering out Pharmacologic Actions

In the biomedical domain, the phrase "pharmacologic actions" stands for a broad category of chemical actions and uses that result in the prevention, treatment, cure, or diagnosis of disease. Typical subclasses include "Antineoplastic Agents," "Lipid Regulating Agents," and "Anti-Inflammatory Agents". In the UMLS Metathesaurus, these terms and phrases have been assigned the semantic type "Chemicals and Drugs" and several child types, which would not differ with the specific drug name for our study. To selectively filter out these pharmacologic actions, 497 headings from the MeSH thesaurus were systematically collected based on the tree structure shown in Figure 1 (left).

Figure 1. Automatic filtering of pharmacologic actions (left) and generic disease names (right).

Filtering out the Generic Names of Diseases

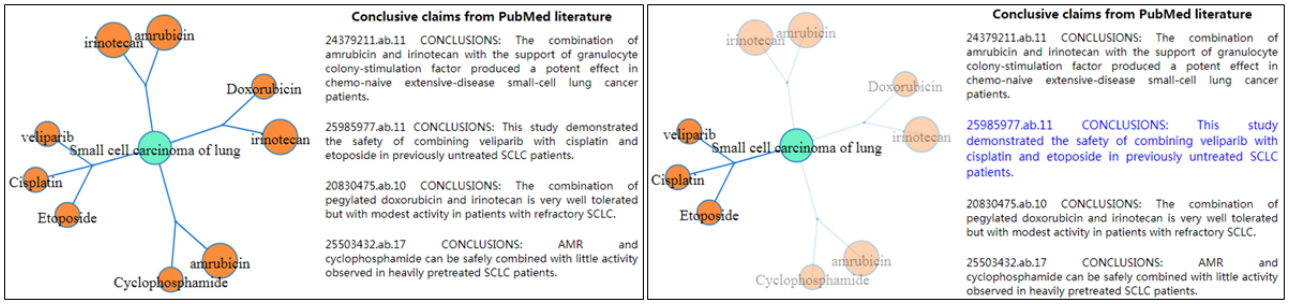
The top-level names of diseases were automatically filtered by disease (class C in the MeSH tree structure) and its direct hyponyms with tree number from C01 to C26, totaling 27 terms. This filtering was applied because the terms are better regarded as classes of disorders rather than specific diseases (Figure 1 [right]).

The Construction and Visualization of Knowledge Graph About Combined Drug Therapy

The knowledge graph is an evolving technology widely used for massive knowledge organization and presentation in the era of big data and artificial intelligence due to its ability to mine machine-understandable knowledge and information. In terms of data structure and storage, knowledge graphs store knowledge in the form of subject-predicate-object (usually called a semantic triple). Traditionally, to visualize a domain knowledge graph, the subjects and objects of triples are intuitively displayed as nodes in a graph, with the predicates presented as various edges linked to subjects and objects accordingly.

In this paper, to emphasize the combined drugs, knowledge about combined drug therapies (S_1+S_i)-P-O (i≥2) discovered by the proposed algorithm will be demonstrated such that the combined drugs will be first bound together and then directed to a specified disorder, while the supporting conclusive claims are shown on the right (Figure 2, left). Upon selecting the linked edge of interest, the specific claim regarding the combined medication will be amplified and highlighted (Figure 2, right). The JavaScript libraries Data-Driven Document (D^3) [28] was utilized to visualize the knowledge graph.

Figure 2. Customized knowledge graph visualization (left) and the conclusive claim being highlighted (right).

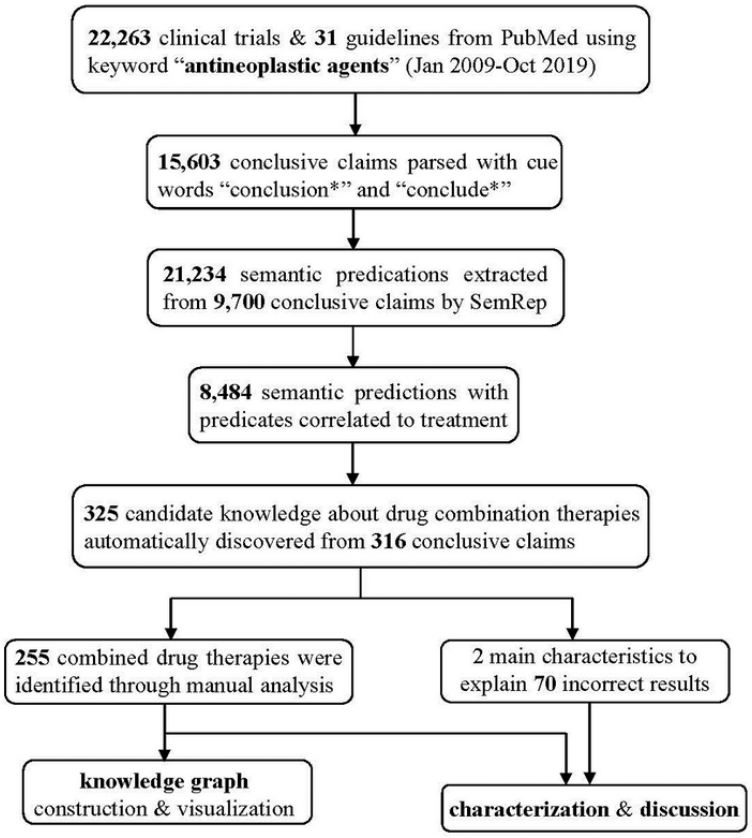


Results

Data Acquisition and Experimental Setup

A summary of the steps taken to discover and identify combined drug therapies is shown in Figure 3. We retrieved 22,263 clinical trial reports and 31 clinical practice guidelines of PubMed abstracts for algorithm verification and validation, with the subject majored on “antineoplastic agents” for drug restriction (Jan 2009 to Oct 2019). The following PubMed queries were used to identify clinical articles:

Figure 3. Study design.



1. Clinical trial reports: ((“clinical trial” [Publication Type] OR “clinical trial, phase I” [Publication Type] OR “clinical trial, phase ii” [Publication Type] OR “clinical trial, phase iii” [Publication Type] OR “clinical trial, phase iv” [Publication Type]) OR “clinical study” [Publication Type]).
2. Clinical practice guidelines: “guideline” [Publication Type]

Using the keywords “conclusion*” and “conclude*”, 15,603 conclusive claims were locally segmented and preserved, then pushed into the batch mode of SemRep for semantic predication extraction. Initially, there were 21,234 semantic predications extracted from 9700 conclusive claims, while 8484 predications had semantic predicates focusing on disease treatment (“TREATS,” “INHIBITS,” “PREVENTS,” and “DISRUPTS”). We then employed the automated algorithm to discover knowledge about combined drug therapies while automatically filtering out pharmacologic actions and generic disease names. As a result, 325 candidate groups of semantic predications about combined drug therapies were discovered from 316 conclusive claims for further analysis and characterization.

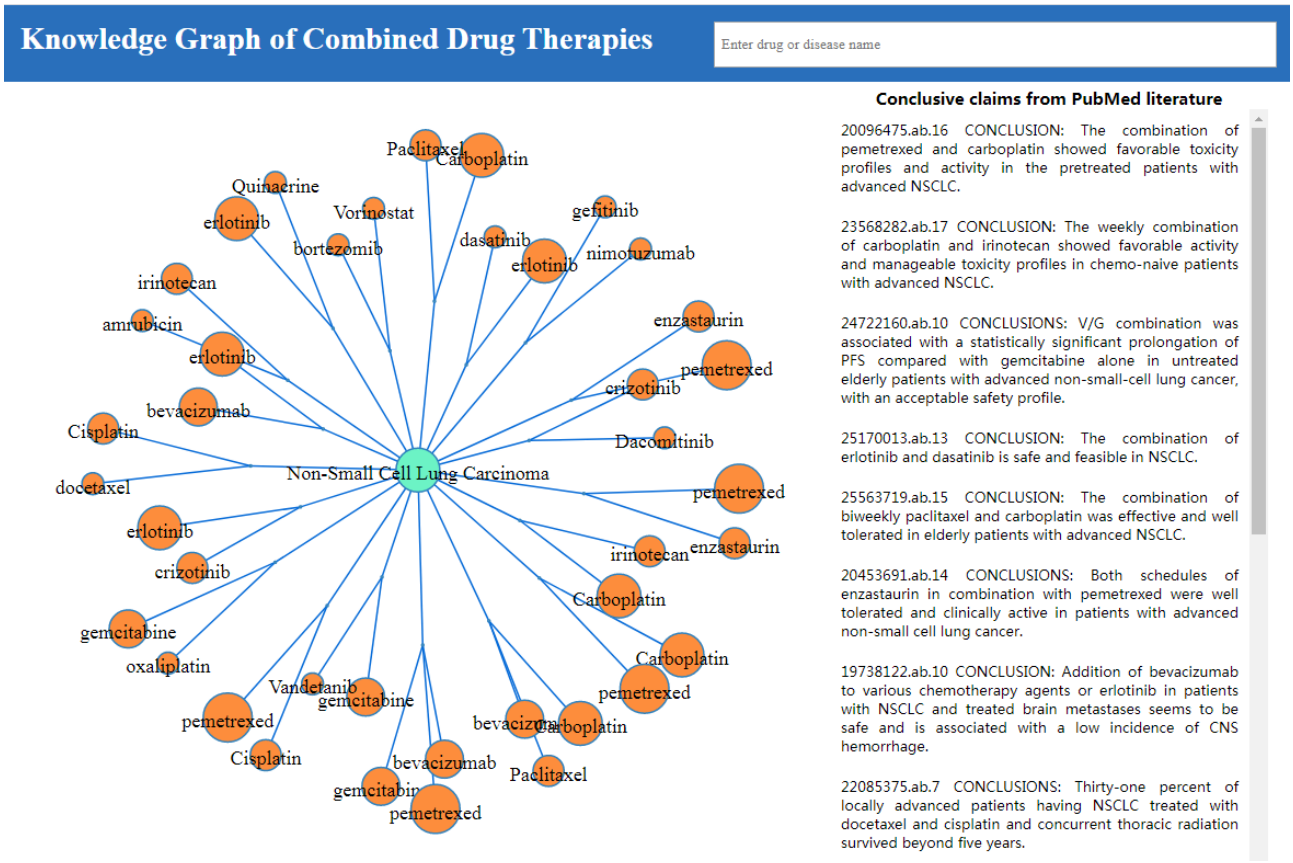
Evaluation

Two biocurators annotated 325 candidate groups of semantic predications about combined medications, which were automatically discovered by the algorithm based on SemRep’s semantic predications from 316 conclusive claims. The primary criteria of the biocuration process were that (1) the discovered drugs were combined to treat the specific disease in a given claim, and a single therapy should be identified; (2) the efficacy of combined therapeutic must be promising and negation was disallowed; and (3) the drug name and disease name should be properly recognized by SemRep. Both biocurators independently evaluated all the candidates groups and identified 255 and 239 combined drug therapies (agreement rate 93.73%). Their disagreements mainly lay in the SemRep object “advanced cancer,” which came from more specific terminal malignancies studied in the conclusive claims (such as “advanced carcinomas of the head and neck” in PMID [PubMed ID] 21947123). After consulting a biomedical scientist with specific clinical knowledge, we accepted this kind of text mapping, acknowledging that advanced cancers usually spread from where they started to other parts of the body. Eventually, 255 of 325 (78.46%) groups of semantic predications were identified to be accurate drug combinations (Multimedia Appendix 1), while 70 were determined to be inaccurate and further classified into 2 categories: limitations of SemRep and limitations of proposal.

Knowledge Graph Construction Based on Identified Knowledge About Combined Medications

Of the 255 identified combined drug therapies, 210 (82.35%) represented combinations of two drugs, 43 (16.86%) combined 3 agents, and 2 (0.78%) included 4 combined medications. These accurate drug combinations as well as their supporting claims were then used to build the knowledge graph based on customized data structure $((S_1+S_i)$-P-O, i≥2). Figure 4 shows a snapshot by searching for “Non-Small Cell Lung Carcinoma”.

Figure 4. Knowledge graph of combined drug therapies centered at "Non-Small Cell Lung Carcinoma".



Characteristics of Inaccurate Results

There were 70 groups of semantic predications from the automated discovery which, upon manual inspection, were deemed inaccurate due to limitations of SemRep (25/70, 35.7%), or limitations of the proposed algorithm (45/70, 64.3%). These were further categorized to include Named Entity Recognition (NER; 8/70, 11.4%) and Semantic Predicate Extraction (SPR) error (17/70, 24.3%), as well as single therapy (40/70, 57.1%) or multiple combined therapies (5/70, 7.1%). Table 3 summarizes the inaccurate results and their characteristics.

Limitations of SemRep

NER is one of the key tasks for knowledge discovery and information retrieval, usually implemented before SPR. In SemRep, NER will be executed by MetaMap, a highly configurable program mapping the biomedical entity to the UMLS Metathesaurus. However, due to the relatively limited coverage of the UMLS Metathesarus or the ambiguity of a given biomedical text, MetaMap may inadequately identify an entity, resulting in an improper semantic subject or object. For the first example in Table 3, "ED-SCLC" represents the abbreviation of "extensive-stage disease, small-cell lung cancer," which is expected to map to "Small cell lung cancer extensive stage" (Concept Unique Identifier: C0278726), but not "Widespread Disease" (CUI: C0849867).

SPR error is another example of SemRep imprecision. In particular, the keyword "failed" was sometimes ignored by SemRep when it appeared in a biomedical text (see the second example in Table 3), resulting in the semantic predicates "TREATS" instead of "NEG_TREATS." To reduce frequency at which negative predications are extracted, we plan to preprocess conclusive claims to filter out negations before SemRep interpretation.

Limitations of the Proposed Algorithm

A majority (40/70, 57.1%) of inaccurate results from the automated algorithm were references to single therapies primarily in comparative clinical studies of two or more individual agents. SemRep's predicate "COMPARED_WITH" may provide a means to filter out these predications. It is common for two or more combined drug therapies to be studied in one published clinical trial (the last claim in Table 3). Future work will focus on these issues to improve the performance of the proposed algorithm.

Table 3. Characteristics of inaccurate results from proposed automatic algorithm.

Explanation	No.	Example	PMID_tx[a]
Limitations of SemRep			
NER error	8	bevacizumab-TREATS-*Widespread Disease* Cisplatin-TREATS-*Widespread Disease* Etoposide-TREATS-*Widespread Disease*	19826110.ab.12 CONCLUSION: The addition of bevacizumab to cisplatin and etoposide in patients with *ED-SCLC* results in …
SPR error	17	ASA 404-*TREATS*-Non-Small Cell Lung Carcinoma Carboplatin-*TREATS*-Non-Small Cell Lung Carcinoma Paclitaxel-*TREATS*-Non-Small Cell Lung Carcinoma	21709202.ab.11 CONCLUSION: The addition of ASA404 to carboplatin and paclitaxel, although generally well tolerated, *failed* to improve frontline efficacy in advanced NSCLC.
Limitations of proposal			
Single Therapy	40	pemetrexed-TREATS-Non-small cell lung cancer metastatic erlotinib-TREATS-Non-small cell lung cancer metastatic	23661337.ab.9 CONCLUSION: Both pemetrexed and erlotinib had *comparable* efficacy in pre-treated patients with metastatic NSCLC.
Multiple combined therapies	5	Custirsen-TREATS-Hormone refractory prostate cancer docetaxel-TREATS-Hormone refractory prostate cancer Mitoxantrone-TREATS-Hormone refractory prostate cancer	21788353.ab.15 CONCLUSION: Custirsen plus *either* docetaxel *or* mitoxantrone was feasible in patients with progressive mCRPC following first-line docetaxel therapy.

[a]PMID_tx: PubMed identifier, abstract, sentence number, and associated text

Discussion

Major Patterns of Combinations According to the Mechanisms of Drugs Being Combined

Among 255 identified combined drug therapies, there were 142 specific drugs after duplicate removal. Classifying by mechanism, 125/142 (88.03%) are antineoplastic agents with 46/142 (32.39%) cytotoxic drugs, 59/142 (41.55%) targeted drugs, 11/142 (7.75%) immunotherapies, 3/142 (2.11%) hormonal drugs, and 6/142 (4.23%) other antineoplastic agents or adjuvant drugs.

We investigated the patterns of identified knowledge based on the mechanism of antineoplastic agents and counted the number of drug combinations under each pattern (Table 4). Although there were fewer cytotoxic drugs than targeted agents, the most common pattern (68/255, 26.67%) were combinations of two cytotoxic drugs, which may provide statistical and practical insights to study new combination of antineoplastic agents for precision medicine. If an antineoplastic agent A produces the same cytotoxic effect as another drug B, and a combination of A and a third cytotoxic agent C has been approved to treat a specific malignancy, our findings suggest the feasibility of a novel combination of B and C (Table 4). Other possible combinations such as A+B and A+B+C may also be valuable to explore. Since various combinations can be followed to develop combined therapies, it is important to be aware of and remain current on all available clinical studies that may be relevant. Our knowledge graph will not only provide a visual representation of existing drug combinations, but also assist practitioners and experts to take full advantage of publicly disseminated clinical trials.

Table 4. Major patterns of combined medication based on mechanisms of antineoplastic agents.

Combinations	Number of Instances
Cytotoxic + Cytotoxic	68
Targeted + Cytotoxic	45
Targeted + Targeted	22
Targeted + Cytotoxic + Cytotoxic	17
Cytotoxic + Other antineoplastic agent/adjuvant drugs	15
Immunotherapy + Targeted	13
Targeted + Other antineoplastic agent/adjuvant drugs	11
Immunotherapy + Cytotoxic	10
Cytotoxic + Cytotoxic + Cytotoxic	6
Others	48

Combined Drug Therapies Discovered in Published Clinical Trials and Clinical Practice Guidelines

All of the combined drug therapies identified in this study were from published clinical trial reports, none of which has been included in clinical practice guidelines. We identified 28 of 31 (90.32%) abstracts in guidelines listed in PubMed by searching "antineoplastic agents" (Jan 2009 to Oct 2019). However, only 4/31 (12.90%) contained conclusive claims with the key words "conclusion*" and "conclude*", with topics for single therapy (PMID: 20390116), intra-arterial chemotherapy (PMID: 23828325), curriculum in surgical oncology (PMID: 27145931), or drug management (PMID: 30381047). We then manually read the remaining guidelines and identified two combined drug therapies in one publication (PMID:21821491). We thus conclude that our method of parsing conclusive claims from PubMed abstracts may not be suitable for clinical practice guidelines, as a considerable number of these publications (87.10%) do not contain the necessary key words. Using structured abstracts after conversion or applying additional key words like "summar*" may improve the acquisition of conclusive claims. Although mentions of combined drug therapies are limited in clinical practice guidelines, our study focused on the discovery of combination therapies from published clinical trials, which inform the development of clinical practice guidelines.

The Markers to Identify Potential Combined Drug Therapies

The word "combin*" (namely "combine" or "combination") is generally used to indicate the combined medication, an assumption affirmed by the data sampled here. Among 316 conclusive claims to automatically identified in this study (Table 5), 171 (54.11%) contain the marker "combin*" and 170 discuss drug combinations, while one described a combination of a drug and radiotherapy. We also noted "coadministration" (2 occurrences) and "co-administered" (1 occurrence) are markers similar to "combin*", as is "regimen" (22 occurrence, 21 of which were for combined drug therapies) being an abbreviation of "antineoplastic combined chemotherapy regimens" [29]. These markers will become key features in the development of our next deep learning–based knowledge discovery algorithm. After SemRep extraction of semantic relations from conclusive claims in the biomedical literature, we plan to add the Bidirectional Encoder Representations from Transformers [30] model as a binary classifier using annotated data from two dimensions: the supporting conclusive claims and the factuality of semantic predications. The claims containing at least one of the identified markers will be used to classify the corresponding groups of semantic predications into positive knowledge about combined drug therapies.

Table 5. Major makers to identify combined drug therapies.

Markers	Occurrence	Combined drug therapy	Other therapy
combin*	171	170	drug & radiotherapy
coadministration	2	2	N/A[a]
co-administered	1	1	N/A
regimen (without markers above)	22	21	Single therapy

[a]N/A: not applicable.

The Utility and Major Applications of the Knowledge Graph for Combined Drug Therapies

The knowledge graph of combined drug therapies will be an appropriate supplement to most leading knowledge bases, similar to SemMedDB [31], which is a widely used publicly available repository extracted from biomedical literature by SemRep. However, the lack of knowledge concerning combinatorial effects is an important limitation of SemMedDB. Our study seeks to fill this gap by providing the combined

medications to enrich the coverage and information provided by SemMedDB and other biomedical knowledge systems.

The proposed knowledge graph has two major applications. An information retrieval system can utilize the knowledge from our graph to integrate various external sources of knowledge and information. Since the subjects and objects of the presented combined medications were drawn from the UMLS Metathesaurus by SemRep, it should be straightforward to integrate our graph with UMLS's source vocabularies for information retrieval, such as DrugBank, Disease Ontology, NCI thesaurus, SNOMEDCT, etc. Another major application is precision medicine and clinical decision-making support. Combined drug therapies provide an alternative to conventional single therapies especially for malignant disorders. In order to pursue clinical and therapeutic approaches to optimal disease management based on individual variations in a patient's genetic profile, it is useful for an expert working with the treatment of a specific cancer to know which other therapies could also fit in that clinical practice. Manually reading the tremendous literature to find available combinations is undoubtedly laborious and time-consuming. Our knowledge graph will help experts quickly and easily identify efficacious combined therapies that may not be immediately evident by a manual survey of published clinical studies.

Conclusions

We have shown that semantic predications extracted from large-scale conclusive claims in biomedical research literature can be used to automatically discover and build a customized knowledge graph to represent existing knowledge about combination therapies. We found that additional filtering and evaluation steps were needed to accurately identify drug combinations from candidate results automatically discovered by the proposed algorithm. From 22,263 published clinical trials retrieved from PubMed, we automatically discovered 325 candidate groups of semantic predications, 255 of which (78.46%) were manually verified as accurate. Two major categories and four subcategories were identified to characterize 70 inaccurate results. To address this precision error, we conclude that additional filtering, context analysis, and feature extraction are required to eliminate single therapies and incorrect semantic predications of SemRep output through active learning [32] or a factuality analyzer program [33].

The proposed algorithm can be generalized to automatically discover generic combined medications for all human disorders, not just malignant neoplasms. It is also likely that a larger number of combined drug therapies could be identified in other types of biomedical publications, such as meta-analysis and comparative studies, in which combined medications are frequently addressed.

By characterizing the major patterns of combinations according to the individual drug mechanisms, we found that combinations of two cytotoxic drugs are the most common for cancer treatment. Moreover, four apparent markers ("combin*", "coadministration", "co-administered" and "regimen") were extracted as key features to further develop the machine learning-based knowledge discovery algorithm.

Acknowledgments

This work was funded by the National Natural Science Foundation of China, grant number 71603280 and the Young Elite Scientists Sponsorship Program by China Association for Science and Technology, grant number 2017QNRC001.

Authors' Contributions

JD supervised the project and administered the work. XYL sampled data and implemented the experimental testing. XYL prepared the initial draft of the manuscript and JD revised it. Both authors provided contributions to the final version of the paper and approved it.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Discovered combined drug therapies.
[TXT File , 24 KB-Multimedia Appendix 1]

References

1. OTA D, AKATSUKA S, NISHI T, KATO T, TAKEUCHI M, TSUJI M, et al. Phase I Study of Combination Therapy With Weekly Nanoparticle Albumin-bound Paclitaxel and Cyclophosphamide in Metastatic Breast Cancer Patients. Anticancer Res 2019 Dec 06;39(12):6903-6907. [doi: 10.21873/anticanres.13910]
2. Morel D, Jeffery D, Aspeslagh S, Almouzni G, Postel-Vinay S. Combining epigenetic drugs with other therapies for solid tumours — past lessons and future promise. Nat Rev Clin Oncol 2019 Sep 30;17(2):91-107. [doi: 10.1038/s41571-019-0267-4]
3. Sun W, Sanderson PE, Zheng W. Drug combination therapy increases successful drug repositioning. Drug Discovery Today 2016 Jul;21(7):1189-1195. [doi: 10.1016/j.drudis.2016.05.015]
4. Kumar MS, Yadav TT, Khair RR, Peters GJ, Yergeri MC. Combination Therapies of Artemisinin and its Derivatives as a Viable Approach for Future Cancer Treatment. CPD 2019 Nov 14;25(31):3323-3338. [doi: 10.2174/1381612825666190902155957]

5. Printz C. Two‐drug combination benefits patients with chronic lymphocytic leukemia. Cancer 2019 Dec 11;126(1):13-13. [doi: 10.1002/cncr.32647]
6. Nurdiati S, Hoede C. 25 years development of knowledge graph theory: the results and the challenge. Memorandum 2008:1876.
7. Salahuddin A, Mushtaq M, Materson BJ. Combination therapy for hypertension 2013: An update. Journal of the American Society of Hypertension 2013 Sep;7(5):401-407. [doi: 10.1016/j.jash.2013.04.013]
8. Shi L, Li S, Yang X, Qi J, Pan G, Zhou B. Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services. BioMed Research International 2017;2017:1-12. [doi: 10.1155/2017/2858423]
9. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a Health Knowledge Graph from Electronic Medical Records. Sci Rep 2017 Jul 20;7(1). [doi: 10.1038/s41598-017-05778-z]
10. Lamurias A, Ferreira JD, Clarke LA, Couto FM. Generating a Tolerogenic Cell Therapy Knowledge Graph from Literature. Front. Immunol 2017 Nov 29;8. [doi: 10.3389/fimmu.2017.01656]
11. Vlietstra WJ, Zielman R, van Dongen RM, Schultes EA, Wiesman F, Vos R, et al. Automated extraction of potential migraine biomarkers using a semantic graph. Journal of Biomedical Informatics 2017 Jul;71:178-189. [doi: 10.1016/j.jbi.2017.05.018]
12. Vlietstra WJ, Vos R, Sijbers AM, van Mulligen EM, Kors JA. Using predicate and provenance information from a knowledge graph for drug efficacy screening. J Biomed Semant 2018 Sep 6;9(1). [doi: 10.1186/s13326-018-0189-6]
13. Lee CH, Khoo CSG, Na JC. Automatic identification of treatment relations for medical ontology learning: An exploratory study. 2004 Jul 13 Presented at: In: Proceedings of the Eighth International ISKO Conference . Wurzburg, Germanyrgon Verlag. FREE Full text; 2004; London p. 245-250.
14. Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinformatics 2013 Jun 6;14(1):181-191. [doi: 10.1186/1471-2105-14-181]
15. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated Acquisition of Disease-Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. Journal of the American Medical Informatics Association 2008 Jan 01;15(1):87-98. [doi: 10.1197/jamia.m2401]
16. Zhao M, Yang CC. Drug Repositioning to Accelerate Drug Development Using Social Media Data: Computational Study on Parkinson Disease. J Med Internet Res 2018 Oct 11;20(10):e271. [doi: 10.2196/jmir.9646]
17. Bchir A, Karaa WBA. Extraction of drug-disease relations from MEDLINE abstracts. 2013 Jun 22 Presented at: InWorld Congress on Computer and Information Technology (WCCIT). IEEE; 2013; Sousse, Tunisia p. 1-3. [doi: 10.1109/wccit.2013.6618759]
18. Wu G, Liu J, Wang C. Predicting drug-disease interactions by semi-supervised graph cut algorithm and three-layer data integration. BMC Med Genomics 2017 Dec 28;10(S5). [doi: 10.1186/s12920-017-0311-0]
19. Zhou H, Lang C, Liu Z, Ning S, Lin Y, Du L. Knowledge-guided convolutional networks for chemical-disease relation extraction. BMC Bioinformatics 2019 May 21;20(1). [doi: 10.1186/s12859-019-2873-7]
20. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics 2003 Dec;36(6):462-477. [doi: 10.1016/j.jbi.2003.11.003]
21. Cong Q, Feng Z, Li F, Zhang L, Rao G, Tao C. Constructing Biomedical Knowledge Graph Based on SemMedDB and Linked Open Data. 2018 Dec 3 Presented at: InIEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018; Madrid, Spain p. 1628-1631. [doi: 10.1109/bibm.2018.8621568]
22. Rosemblat G, Fiszman M, Shin D, Kilicoglu H. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. Journal of Biomedical Informatics 2019 Oct;98:103275. [doi: 10.1016/j.jbi.2019.103275]
23. Liu Y, Bill R, Fiszman M, Rindflesch T, Pedersen T, Melton GB, et al. Using SemRep to label semantic relations extracted from clinical text. 2012 Nov 03 Presented at: In: AMIA annual symposium proceedings. American Medical Informatics Association, , . FREE Full text Medline; 2012; Chicago, Illinois p. A.
24. Zhang R, Cairelli MJ, Fiszman M, Rosemblat G, Kilicoglu H, Rindflesch TC, et al. Using semantic predications to uncover drug–drug interactions in clinical data. Journal of Biomedical Informatics 2014 Jun;49:134-147. [doi: 10.1016/j.jbi.2014.01.004]
25. Fathiamini S, Johnson AM, Zeng J, Araya A, Holla V, Bailey AM, et al. Automated identification of molecular effects of drugs (AIMED). J Am Med Inform Assoc 2016 Apr 23;23(4):758-765. [doi: 10.1093/jamia/ocw030]
26. Accessed November 19. 2019 Nov 01. Medline pubmed production statistics URL: https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html [accessed 2019-11-01]
27. Kilicoglu H, Rosemblat G, Fiszman M, Rindflesch TC. Constructing a semantic predication gold standard from the biomedical literature. BMC Bioinformatics 2011 Dec 20;12(1). [doi: 10.1186/1471-2105-12-486]
28. Bostock M, Ogievetsky V, Heer J. D³ Data-Driven Documents. IEEE Trans. Visual. Comput. Graphics 2011 Dec;17(12):2301-2309. [doi: 10.1109/tvcg.2011.185]
29. MeSH Thesaurus. URL: https://meshb.nlm.nih.gov/record/ui?ui=D000971 [accessed 2020-04-22]
30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Computation and Language, 1-16. FREE Full text 2019 May 04.

31. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 2012 Oct 08;28(23):3158-3160. [doi: 10.1093/bioinformatics/bts591]
32. Vasilakes J, Rizvi R, Melton GB. Evaluating active learning methods for annotating semantic predications. JAMIA open, . FREE Full text 3074 2018;1(2):0594-0282. [doi: 10.1093/jamiaopen/ooy021]
33. Kilicoglu H, Rosemblat G, Rindflesch TC. Assigning factuality values to semantic relations extracted from biomedical research literature. PLoS ONE 2017 Jul 5;12(7):e0179926. [doi: 10.1371/journal.pone.0179926]

Abbreviations

CUI: Concept Unique Identifier
MeSH: Medical Subject Headings
NER: Named Entity Recognition
PMID: PubMed ID/reference number
SemRep: Semantic Knowledge Representation
SPR: Semantic Predicate Extraction
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 21.02.20; peer-reviewed by Z Xing, WC Su; comments to author 20.03.20; revised version received 26.03.20; accepted 29.03.20; published 28.04.20

Please cite as:
Du J, Li X
A Knowledge Graph of Combined Drug Therapies Using Semantic Predications From Biomedical Literature: Algorithm Development
JMIR Med Inform 2020;8(4):e18323
URL: http://medinform.jmir.org/2020/4/e18323/
doi: 10.2196/18323
PMID: 32343247