Summarizing Drug Information in Medline Citations

Marcelo Fiszman MD PhD,[1] Thomas C. Rindflesch PhD,[2] Halil Kilicoglu MS[2]

[1]Graduate School of Medicine, University of Tennessee, Knoxville, TN
[2]National Library of Medicine, Bethesda, MD

Adverse drug events and drug-drug interactions are a major concern in patient care. Although databases exist to provide information about drugs, they are not always up-to-date and complete (particularly regarding pharmacogenetics). We propose a methodology based on automatic summarization to identify drug information in Medline citations and present results to the user in a convenient form. We evaluate the method on a selection of citations discussing ten drugs ranging from the proton pump inhibitor lansoprazole to the vasoconstrictor sumatriptan. We suggest that automatic summarization can provide a valuable adjunct to curated drug databases in supporting quality patient care.

INTRODUCTION

Adverse drug events and drug-drug interactions are a significant source of error in providing patient care [1]. Many institutions seek to minimize these events by referring to locally [2,3] or commercially curated drug databases, such as MicroMedex and DrugDigest [4,5], which contain potential drug-drug interactions. However, these resources are not always complete, and there is a delay between publication of relevant information and its appearance in the databases. Moreover, there are an increasing number of known interactions, not only between drugs, but also with respect to genes and genetic function (pharmacogenetics) [6,7], which are not always covered by these databases. Automatic approaches that allow convenient access to the research literature on drug information directly can assist clinicians in providing quality care to their patients.

We propose to identify adverse drug events and drug interactions in Medline citations using automatic summarization. We expand a methodology we previously introduced for summarizing medical text with respect to treatment of disease [8, 9] to apply to drug information. The results are presented to the user as an informative graph with links to source text and not only provide an overview of the research literature on a particular drug but can also address specific questions, such as "is this drug teratogenic?"

BACKGROUND

Automatic summarization

Although automatic summarization has been used in the biomedical domain [10], it has not been applied comprehensively to drug information. Afantenos [11] provides a survey for biomedical summarization, which documents the high prevalence of the extraction paradigm, in which summaries are constructed from sentences extracted from the source text. We follow the semantic abstraction paradigm [12] for automatic summarization, in which a summary is a semantic representation of the most important aspects of the content of the source documents.

SemRep

Abstraction summarization requires semantic predications as input, and we rely on SemRep [13], a natural language processing system for extracting semantic predications from medical text. SemRep depends on a partial syntactic analysis based on the SPECIALIST Lexicon [14] and MedPost tagger [15]. Semantic analysis is dependent on medical domain knowledge in the UMLS Metathesaurus and Semantic Network. Access to the Metathesaurus is provided by MetaMap [16].

As an example of SemRep output, the predications in (2) comprise a partial representation of the meaning of (1). Each predicate (TREATS and AFFECTS) is a relation from the UMLS Semantic Network, while the arguments ("Adrenal Cortex Hormones" and the symptoms) are concepts from the UMLS Metathesaurus.

(1) In the knee, injections of corticosteroids into the joint may relieve inflammation, and reduce pain and disability.

(2) Adrenal Cortex Hormones TREATS Inflammation
Adrenal Cortex Hormones AFFECTS Disability NOS (finding)
Adrenal Cortex Hormones AFFECTS Pain

Summarization for treatment of disease

In previous work [8, 9] we summarized Medline citations on treatment of disease. A query specifying a disorder of interest is submitted to PubMed. The citations returned are processed by SemRep, which produces a list of semantic predications representing

the content of the source citations. Summarization then reduces and generalizes this list through a transformation process that produces a "condensate" of the source predications. Finally, the condensate is presented to the user in graphical form. A schematic view of the system can be seen in Figure 1.

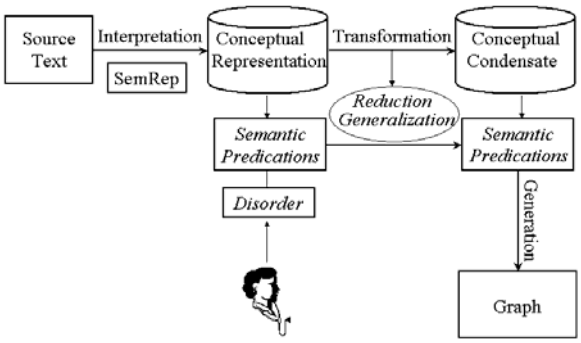

Figure 1. Summarization of Medline citations.

The core of this system is the transformation process, which consists of four phases: relevance, connectivity, novelty, and saliency. During relevance, the system ensures that all predications in the summary pertain to treatment of the topic disease. This is implemented through a predication schema [17], which contains predication templates. A template contains a predicate from the UMLS Semantic Network and argument "domains" expressing UMLS semantic types that can be associated with UMLS concepts serving as arguments of the predicate.

The schema for treatment of disease includes Semantic Network predicates ISA, CAUSES, TREATS, PREVENTS, LOCATION_OF, OCCURS_IN, and CO-OCCURS_WITH. The argument domains for the treatment schema are Disorders, Etiologic process, Treatment, and Body location. The Disorders domain includes such semantic types as 'Disease or Syndrome' and 'Neoplastic Process'. Etiologic process has 'Virus', 'Antibiotic', and Hazardous or Poisonous Substance', for example. Treatment includes 'Pharmacologic Substance', 'Therapeutic or Preventive Procedure', among others. Body location contains the UMLS anatomy semantic types.

In summarizing for treatment of degenerative polyarthritis, for example, the schema allows predications "Hyaluronan TREATS Degenerative polyarthritis" and "Entire knee region LOCATION_OF Degenerative polyarthritis" but eliminates "Ophthalmologic Surgical Procedures USES Hyaluronan."

The remaining phases of the transformation process further generalize and condense the list of predications. Connectivity adds additional predications that have a bearing on the topic disorder. For example, given the topic degenerative polyarthritis, "Hyaluronan TREATS Arthropathies NOS" is also included. Novelty eliminates predications having arguments which are too general to be useful, such as "Pharmacologic Substance TREATS Patients." This phase is implemented by checking for depth in UMLS Metathesaurus hierarchies. Finally, the Saliency phase [18] calculates frequency of occurrence of predications and keeps only those that appear more frequently than average.

The application of the transformation phase to the SemRep predications for 87 Medline citations on hyaluronan as a treatment for degenerative polyarthritis produces a condensate that is displayed as the graph in Figure 2, which provides users with an informative overview of the content of the citations. Each edge in the graph is linked to the citation text that generated the predication, allowing easy access to the relevant research literature. For example, the predication "Adrenal Cortex Hormones TREATS Degenerative polyarthritis" is linked to text *Corticosteroid compared with hyaluronic acid injections for the treatment of osteoarthritis of the knee*, the title of citation 15069162.

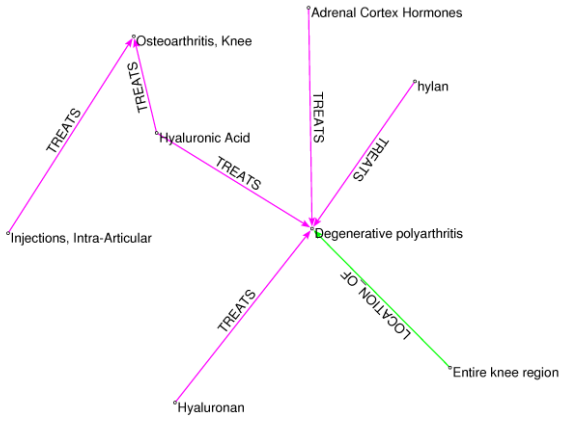

Figure 2. Summarization results for treatment of degenerative polyarthritis.

METHODS

Modifying the schema for drug information

For this study we adapted the summarization system for treatment of disease to drug therapy. The only change needed was to modify the schema to accommodate text discussing clinical and research aspects of drugs, such as disease indications, adverse

effects, and interactions. The drug schema contains the following Semantic Network predicates: AFFECTS, CAUSES, COMPLICATES, DISRUPTS, INTERACTS_WITH, ISA, PREVENTS, and TREATS. The argument domains for these predicates are Drugs, Chemicals, Physiology, Disorders, and Anatomy. Each domain is defined in terms of the UMLS semantic groups [19]. All semantic types for each argument domain are given in (3) through (7).

(3) Drugs: 'Amino Acid, Peptide, or Protein', 'Antibiotic', 'Hazardous or Poisonous Substance', 'Hormone', 'Nucleic Acid, Nucleoside, or Nucleotide', 'Organic Chemical', 'Pharmacologic Substance', 'Steroid', 'Vitamin'

(4) Chemicals: Drugs domain plus 'Biologically Active Substance', 'Carbohydrate', 'Eicosanoid', 'Element, Ion, or Isotope', 'Enzyme', 'Immunologic Factor', 'Inorganic Chemical', 'Lipid', 'Neuroreactive Substance or Biogenic Amine', 'Organophosphorus Compound'

(5) Physiology: 'Biologic Function', 'Cell Component', 'Cell or Molecular Dysfunction', 'Genetic Function', 'Mental Process', 'Molecular Function', 'Organism Function', 'Organ or Tissue Function', 'Physiologic Function'

(6) Disorders : 'Acquired Abnormality', 'Anatomical Abnormality', 'Congenital Abnormality', 'Disease or Syndrome', 'Injury or Poisoning', 'Mental or Behavioral Dysfunction', 'Neoplastic Process', 'Pathologic Function'

(7) Anatomy: 'Body Part, Organ, or Organ Component', 'Cell Component', 'Cell', 'Embryonic Structure', 'Fully Formed Anatomical Structure', 'Gene or Genome', 'Tissue'

The argument domains are combined with allowable predicates to form predication templates, which constitute the complete schema for drugs, as shown in (8).

(8) Drug schema
{Drugs} AFFECTS {Disorders}
{Drugs} AFFECTS {Physiology}
{Drugs} CAUSES {Disorders}
{Drugs} COMPLICATES {Disorders}
{Drugs} COMPLICATES {Physiology}
{Drugs} DISRUPTS {Anatomy}
{Drugs} DISRUPTS {Physiology}
{Drugs} INTERACTS_WITH {Chemicals}
{Drugs} ISA {Chemicals}
{Drugs} PREVENTS {Disorders}
{Drugs} TREATS {Disorders}

In applying the summarization system to text discussing drugs using the drug schema, Figure 3 illustrates the results of summarizing 508 Medline citations retrieved with the query "phenytoin."

Figure 3. Summarization results for Phenytoin.

In Figure 3, it can be seen, for example, that phenytoin (the central concept) is an anticonvulsant, which treats epilepsy and disrupts sleep.

Evaluation

We performed a linguistic evaluation on the quality of INTERACTS_WITH and the AFFECTS predications for a sample of ten drugs categorized as follows: Central nervous system: citalopram, paroxetine, phenytoin, and selegiline; Antiviral: efavirenz; Heart: enalapril; Gastrointestinal: lansoprazole and ranitidine; Vascular: sumatriptan; Skin: voriconazole.

We chose INTERACTS_WITH because of the importance of this relation in this context (drug-drug interactions only). We chose AFFECTS because we wanted to look more generally at the way drugs interact with disease and physiology.

A query was issued to PubMed consisting of the drug name; output was limited to English citations with abstracts. Output was further limited by date so that approximately the most recent 400 to 600 citations were retrieved.

The saliency phase eliminates predications occurring less frequently than average. We were interested in seeing its effect on accuracy, and thus evaluated both before and after this phase. After summarizing the set of citations for each of the ten drugs, the INTERACTS_WITH (before saliency) and AFFECTS (after saliency) predications were isolated in the summarized output and assessed for linguistic accuracy.

RESULTS

Table 1 indicates the number of citations, sentences, and predications retrieved for all drugs

and the number that were judged to be correct. Finally, precision is given.

	Before saliency	After saliency	Total
# Citations	122	130	252
# Sentences	149	157	306
# Predications	203	189	392
# Correct	117	148	265
Precision	58%	78%	68 %

Table 1. Results of the linguistic evaluation.

DISCUSSION

It is to be noted that the final phase of our transformation process, saliency, has a significant effect on accuracy. Overall, the majority of the errors were due to two phenomena: missed negation and complicated sentence structure. An example of missed negation is seen in (10) as the output for (9).

(9) Selegiline was found unable to inhibit deamination of beta-PEA.

(10)Selegiline INTERACTS_WITH Phenethylamine

Complex sentence structure, such as that seen in (11) led to the incorrect predication (12).

(11)After 10 days of incubation, the antifungal activities of ketoconazole (MIC at which 90% of isolates were inhibited [MIC90], 0.125 microg/ml), itraconazole (MIC90, 0.064 microg/ml), and voriconazole (MIC90, 0.125 microg/ml) appeared superior to those of fluconazole (MIC90, 128 microg/ml) and amphotericin B (MIC90, 1 microg/ml), with MICs in the clinically relevant range.

(12)Ketoconazole INTERACTS_WITH voriconazole

Clinical observations

We discuss three examples that illustrate the type of drug information provided by our summarization method and comment on possible implications linking the research literature to clinical practice.

Phenytoin

While examining the summary for phenytoin**,** we found an interaction with losartan**,** which seems to be mediated by genetics and is not listed in curated databases such as MicroMedex and DrugDigest. The sentence that produced this information (13) appears in the abstract of PMID 12235444, "Evaluation of potential losartan-phenytoin drug interactions in healthy volunteers," published in September, 2002.

(13)**Phenytoin** inhibited the CYP2C9-mediated conversion of **losartan** to E3174.

The summary also contains predications that both phenytoin and docetaxel can cause exanthema. This information appears in PMID 15500423, "Impact of phenytoin therapy on the skin and skin disease," November, 2004 (14) and 15088316, "Docetaxel (taxotere) induced subacute cutaneous lupus erythematosus: report of 4 cases," April, 2004 (15).

(14)**Phenytoin** can induce generalised eruptions that include: a maculopapular **exanthem**, …

(15)Pathogenetically, **docetaxel** may evoke a **lupus-like eruption** through its proapoptotic effects on replicating cells …

In a patient receiving both medications it could be difficult to determine the exact etiology of the dermatologic effects.

Paroxetine

The summary for paroxetine contains the predication that this drug interacts with noradrenaline transporter (12359676, "Inhibition of norepinephrine uptake in patients with major depression treated with paroxetine," October, 2002) (16).

(16)The study examined whether **paroxetine** inhibits the human **norepinephrine transporter** in addition to the human serotonin (5-HT) transporter in patients with major depressive disorder.

Although paroxetine is currently classified as a serotonine reuptake inhibitor, this study demonstrates that it also acts as a norepinephrine uptake inhibitor. The authors state that the clinical significance of the latter action is unknown. It is interesting to note that the antidepressant duloxetine acts on both receptors as well and has generated considerable interest in the psychiatric literature for the treatment of depression.

Lansoprazole

The results of this methodology for extracting drug information from the research literature facilitate interesting observations on some pharmacogenetic interactions of lansoprazole. One predication in the summary for this drug asserts an interaction with clarithromycin (17) (12161414, "Localization of [14C]clarithromycin in rat gastric tissue when administered with lansoprazole and amoxicillin," August, 2002).

(17)The amount of unchanged **clarithromycin** in … increased with co-administration of **lansoprazole** …

From a later study (18) (14664653, "Clinical pharmacology of proton pump inhibitors: what the practising physician needs to know," 2003), the system extracted the predication "CYP2C19 protein, human INTERACTS_WITH lansoprazole."

(18)… significant genetic polymorphisms for one of the cytochrome P450 (CYP) isoenzymes involved in PPI metabolism (**CYP2C19)** … has been shown to substantially increase plasma levels of omeprazole, **lansoprazole** and pantoprazole, but not those of rabeprazole.

Moreover, another predication, "clarithromycin AFFECTS Drug Kinetics," was extracted from a very recent publication (19) (15752376, "Effects of Clarithromycin on lansoprazole pharmacokinetics between CYP2C19 genotypes," March, 2005).

(19)Effects of **clarithromycin** on lansoprazole **pharmacokinetics** between CYP2C19 genotypes.

The conclusion of this study is that "… there are significant drug interactions between lansoprazole and clarithromycin in all CYP2C19 genotype groups …" This provides a potential explanation for co-administration of these medications in the gastrointestinal clinical setting.

CONCLUSION

We propose the use of semantic processing and automatic summarization for identifying drug information in Medline citations. We adapted an existing summarization method to apply to drug indications, adverse events, and interactions. Results are presented as an informative graph with links to the citation text underlying the summary, and we suggest that this method can serve as a useful supplement to curated drug databases in support of quality patient care.

Acknowledgments

This study was supported in part by the Intramural Research Programs of the National Institutes of Health, National Library of Medicine.

References

1. LT Kohn, JM Corrigan, MS Donaldson. To err is human: building a safer health system. Institute of Medicine. Nov, 1999.

2. Del Fiol G, Rocha BH, Kuperman GJ, Bates DW, Nohama P. Comparison of two knowledge bases on the detection of drug-drug interactions. Proc AMIA Symp. 2000;171-5.

3. Morimoto T, Gandhi TK, Seger AC, Hsieh TC, Bates DW. Adverse drug events and medication errors: detection and classification methods. Qual Saf Health Care. 2004 Aug;13(4):306-14.

4. http://www.micromedex.com/

5. http://www.drugdigest.org/

6. Wilke RA, Reif DM, Moore JH. Combinatorial pharmacogenetics. Nat Rev Drug Discov. 2005 Nov;4(11):911-8.

7. Ritchie MD, Carillo MW, Wilke RA. Computational approaches for pharmacogenomics. Pac Symp Biocomput. 2005;245-7.

8. Fiszman M, Rindflesch TC, Kilicoglu H. Summarization of an online medical encyclopedia. Medinfo. 2004;11(Pt 1):506-10.

9. Fiszman M, Rindflesch TC, Kilicoglu, H. Abstraction summarization for managing the biomedical research literature. Proc of the HLT-NAACL Workshop on Computational Lexical Semantics. 2004;76-83.

10. McKeown M, Chang SF, Cimino J, et al. PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information. Proc of JCDL. 2001.

11. Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. Artif Intell Med. 2005 Feb;33(2):157-77.

12. Hahn U, Mani, I. The challenges of automatic summarization. Computer. 2000; 33(11):29-36.

13. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J of Biomed Inf. 2003 Dec;36(6):462-77.

14. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994;235-9.

15. Smith L, Rindflesch T, Wilbur WJ. MedPost: a part-of-speech tagger for biomedical text. Bioinformatics. 2004;20(14):2320-1.

16. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp. 2001;17-21.

17. Jacquelinet C, Burgun A, Delamarre D, et al. Developing the ontological foundations of a terminological system for end-stage diseases, organ failure, dialysis and transplantation. Int J Med Inf. 2003; 70(2-3):317-28.

18. Hahn U, Reimer U. Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. In: Mani d Maybury, Eds. Advances in Automatic Text Summarization. Cambridge, London. MIT Press, 1999, pp. 215-232.

19. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo. 2001;10(Pt 1):216-20.