Research Paper ■

Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study

Xiaoyan Wang, MPhi, George Hripcsak, MD, MS, Marianthi Markatou, PhD, Carol Friedman, PhD

Abstract **Objective:** It is vital to detect the full safety profile of a drug throughout its market life. Current pharmacovigilance systems still have substantial limitations, however. The objective of our work is to demonstrate the feasibility of using natural language processing (NLP), the comprehensive Electronic Health Record (EHR), and association statistics for pharmacovigilance purposes.

Design: Narrative discharge summaries were collected from the Clinical Information System at New York Presbyterian Hospital (NYPH). MedLEE, an NLP system, was applied to the collection to identify medication events and entities which could be potential adverse drug events (ADEs). Co-occurrence statistics with adjusted volume tests were used to detect associations between the two types of entities, to calculate the strengths of the associations, and to determine their cutoff thresholds. Seven drugs/drug classes (ibuprofen, morphine, warfarin, bupropion, paroxetine, rosiglitazone, ACE inhibitors) with known ADEs were selected to evaluate the system.

Results: One hundred thirty-two potential ADEs were found to be associated with the 7 drugs. Overall recall and precision were 0.75 and 0.31 for known ADEs respectively. Importantly, qualitative evaluation using historic roll back design suggested that novel ADEs could be detected using our system.

Conclusions: This study provides a framework for the development of active, high-throughput and prospective systems which could potentially unveil drug safety profiles throughout their entire market life. Our results demonstrate that the framework is feasible although there are some challenging issues. To the best of our knowledge, this is the first study using comprehensive unstructured data from the EHR for pharmacovigilance.

■ **J Am Med Inform Assoc.** 2009;16:328–337. DOI 10.1197/jamia.M3028.

Introduction

In the 1960s, the tragedy of thalidomide affected nearly 10,000 children worldwide.[1] Thalidomide had been marketed as an effective sedative-hypnotic and antiemetic medication during the late 1950s and early 1960s. Before its release, inadequate tests were performed to assess the drug's safety. Tragically the drug caused major birth defects in children in countries where the drug was prescribed to pregnant women. The field of postmarketing drug safety has received a great deal of attention ever since. Pharmacovigilance systems have been introduced in the biomedical domain and have played a key role in drug safety monitoring during the last fifty years. Although a randomized clinical trial (RCT) is considered a gold standard for determining the risks and benefits of a drug, it is generally recognized that premarketing RCTs may not detect all safety issues related to a particular drug before its use in clinical practice. First, premarketing RCTs have inherent limitations due to small numbers, short duration and restrictive inclusion criteria. These RCTs are not powered to detect uncommon (incidence of 1 in 1,000), rare (incidence of 1 in 10,000) or long-term (latency of > 6 mo) adverse drug events (ADEs).[2] Second, clinical trial experiences (i.e., dosing regimen, duration of administration, or concomitant therapies) from restricted subjects may not mirror the actual use in a diverse population in terms of age, race, gender, comorbidities, etc.[2,3] Third, premarketing RCTs are designed to prove efficacy. Safety is a big issue in the RCTs for drug development; however, they are not powered to identify rare events. The efficacy data of a drug is generally more robust and well-established based on premarketing RCTs, while less is known concerning safety profiles.[2,3]

It is therefore vital to establish safety profiles over the market life of a drug that incorporates comprehensive clinical data and more diverse populations. Various databases and data mining algorithms have been developed to support pharmacovigilance tasks. There has been consider-

Affiliations of the authors: Department of Biomedical Informatics (XW, GH, CF), Department of Biostatistics, Columbia University (MM), New York, NY.

The authors thank Lyudmila Shagina for assistance with MedLEE, and Dr. Herbert Chase for his assistance on reviewing the results obtained in this study and construction of the reference standard. Dr. Friedman was supported in part by grants R01 LM007659 and R01 LM008635 and Dr. Hripcsak was supported by R01 LM06910 from the National Library of Medicine, and Dr. Markatou was supported by DMS-0504957 from the National Science Foundation.

Correspondence: Carol Friedman, PhD, Department of Biomedical Informatics, Columbia University, 622 West 168th Street, VC5, New York, NY 10032; e-mail: <friedman@dbmi.columbia.edu>.

Received for review: 10/09/08; accepted for publication: 01/31/09.

able work in developing surveillance systems to monitor large stable populations.[4,5] Data mining algorithms using spontaneous reporting systems, pharmacoepidemiology databases and Electronic Health Records (EHR) systems have produced some interesting results.[6–8] The success of current pharmacovigilance systems, however, is hampered by limitations inherent in the pharmacovigilance databases and in the pitfalls of data mining algorithms.[9] Additionally, most of the work in the field has been based on retrospective cohorts that lack the potential to be extended to prospective studies.

For a long time, pharmacovigilance researchers have been seeking a real time, continuous and prospective approach. Towards this goal, we propose a high throughput system that demonstrates the relevance and significance of using the EHR for pharmacovigilance. Data mining algorithms in pharmacovigilance have focused on coded and structured data, and therefore miss important clinical data that is relevant for pharmacovigilance. Some important ADEs, such as "fever" and "feeling suicidal", are generally only available in the narrative EHR reports. This paper discusses a framework that enables automated active pharmacovigilance by applying natural language processing (NLP) and association statistics on comprehensive unstructured clinical data from EHR systems. We present a proof of principle that it is feasible to develop a methodology that could unveil drug safety profiles and novel adverse events in a timely fashion. Thus, our work differs from related work in pharmacovigilance in that we use NLP to transform large amounts of comprehensive clinical data to a form useable by association statistics. Additionally, the system we propose has the potential to uncover new ADEs prospectively.

Background

Traditional work in pharmacovigilance originally focused on the medical evaluation of an individual case report or the literature.[10] Subsequent work involved establishment and adoption of spontaneous reporting systems (SRS) for pharmacovigilance by regulatory authorities.[10,11] Increasing access to multiple streams of data, such as data in EHRs, pose many new challenges and possibilities for use in the detection of novel adverse signals.[7,12] Meanwhile, data mining algorithms such as disproportionality analysis (DPA), correlation analysis and multivariate regression have been developed and integrated into pharmacovigilance databases to detect adverse signals of drugs.[11,13]

Pharmacovigilance Databases

Spontaneous reporting systems (SRSs) have been the primary means for providing postmarketing safety information on drugs since 1960. Prominent SRS databases include the Adverse Event Reporting System (AERS) in the United States, the United Kingdom's Yellow Card Scheme of the Medicines and Healthcare Products Regulatory Agency (MHRA), the European Agency for the Evaluation of Medical Products (EMEA) and the World Health Organization (the WHO Uppsala Monitoring Center).[4,5,14,15] In addition, there are other databases associated with adverse reporting, such as the Vaccine Adverse Event Reporting Systems (VAERS), and the Manufacturer and User Facility Device Experience Database (MAUDE).[16–18] Surveillance based on SRS databases has been a cornerstone for the early detection of safety issues related to drugs. Analysis of spontaneous reports has been a critical component for the removal of more than 20 drugs/drug products from the market due to safety problems.[8] The SRS databases, however, have several limitations. First, the potential ADE reports are often incomplete and inaccurate due to voluntary reporting. Second, SRSs are often criticized for biased reporting and substantial underreporting.[19] Third, sample distributions related to submission regulations, geographic marketing and population may diverge for different drugs.

Pharmacoepidemiology databases have been created to provide relevant information for detecting new ADEs.[12,20] These databases are advantageous because they include clinical information over long periods for large numbers of patients. Some pharmacoepidemiology databases include New Zealand Intensive Medicines Monitoring Programme (IMMP) databases, the medicine monitoring unit (MEMO) databases and general practice research databases (GPRD) in the UK.[6,21,22] The GPRD provides complete profiles of over 3 million patients, including demographics, medical diagnoses, treatments, hospitalizations, etc, along with the dates and locations of events.[6,23] Pharmacoepidemiology databases are typically used to refute, confirm or strengthen signals from other approaches but not to discover new hypotheses.[24] Although these databases contain a substantial amount of comprehensive information in both structured and in textual form, only a very small amount of data are recorded as structured information and therefore the majority of information cannot be accessed by the pharmacovigilance applications. In addition, the Medicare and Medicaid databases have also been used for pharmacovigilance.[25,26]

Researchers have looked beyond SRS databases and pharmocoepidemiology databases to search for safety signals and have started to use EHRs for ADE detection.[27] For example, Berlowitz, and colleagues used prescription and laboratory test data, and found that the interaction of β blockers and warfarin could affect the risk of hemorrhaging in patients with congestive heart failure.[28] A big advantage for using the EHR for pharmacovigilance is the potential to perform active and real time surveillance, and the probable reduction of errors caused by biased reporting. However, most of the clinical information in patient records also consists of unstructured narratives, such as discharge summaries, progress reports, or nursing notes, and therefore much of the data is also inaccessible for pharmacovigilance purposes. Only structured and coded data in EHRs have been used to detect novel ADEs.

Data Mining Algorithms Used in Pharmacovigilance

Traditional pharmacovigilance involves "case by case" manual evaluation of reports or the literature. Manual assessment of large quantities of data, however, is challenging and costly. Data mining algorithms, as techniques of extracting valuable and interesting information from large complex databases, have since been developed and applied in quantitative pharmocovigilance to discover potential ADEs.[29]

Most of the data mining algorithms have explored some form of disproportionality analysis (DPA). The DPA algorithms involve calculating surrogate observed-to-expected ratios in which each potential drug-ADE pair is compared to background across all other drugs and events in the database. The simplest approach using DPA algorithms involves tabulating each drug-candidate ADE as a contingency table, and then calculating frequentist metrics, such as relative reporting ratio (RRR) and/or reporting odds ratio (ROR).[11,30] More complex algorithms, such as gamma-Poisson shrinker (GPS) and multi-item gamma-Poisson shrinker (MGPS) were developed using Bayesian statistics.[13,31,32] In addition, approaches involving multiple regression modeling have been used to deal with higher-order associations, such as interactions of β blockers and warfarin on haemorrhagic events.[28] Traditional statistics methods such as sequential probability ratio testing have been used in pharmacoepidemiology and EHR databases.[11,27] Notably, all these methods have been applied only on the structured data.

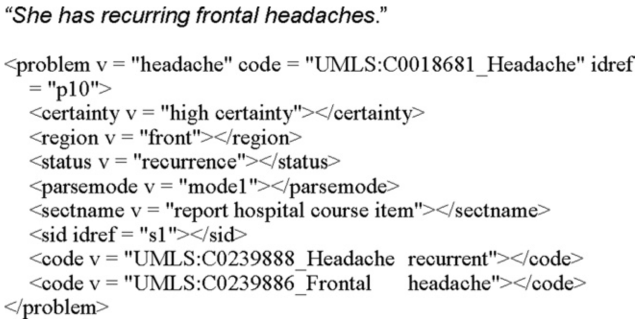
Natural Language Processing and Knowledge Discovery

As discussed above, clinical information in narrative reports is not accessible for pharmacovigilance applications, and is buried in either biomedical literature or narrative clinical reports. Natural language processing (NLP), a high throughput technology, has been applied in biomedicine for decades.[33] The NLP systems have been developed to identify, extract, and encode information within biomedical literature and clinical narratives. Some systems include MEDSYNDIKATE, MetaMap, SemRep, MedLEE, and BioMedLEE.[34–38] There also have been some NLP techniques applied to detect ADEs from narrative reports of EHR systems.[39,40] However, these focus on ADE detection and patient safety, but not on knowledge discovery and pharmacovigilance.

An increasing number of text mining researchers focus on extracting and establishing associations between entities from textual data,[41] and NLP has come to play an essential role in automatically extracting entities and relations within documents. Co-occurrence statistics is widely used to determine relations between the entities, and has proven to be effective in acquiring associations between biological and clinical entities.[42,43] Rindflesch and colleagues extracted drug and disease entities from the Mayo Clinic notes using SemRep and constructed a repository of drug-disease co-occurences to validate inferences produced by SemRep about drug treatments for diseases.[38,44]

Related Work in Our Laboratory

The MedLEE (Medical Language Extraction and Encoding) system, which is used in this study, has been deployed to extract and encode information in clinical narratives for a large number of different applications and studies.[36,37] For a given report, MedLEE generates a set of structured findings, such as problem (*headache*), or medication (*ibuprofen*), along with associated modifiers, such as certainty (*no, high certainty*), status (*previous, recent*), body location (*chest*), and section (Hospital Course). The output of MedLEE is consistent with frames, and has the format Type-Value-Modifiers, where Type is the type of information in the frame, Value is the value and Modifiers are a sequence of frames containing the same format where each modifier frame denotes a certain type of qualifying information.



"She has recurring frontal headaches."

```
<problem v = "headache" code = "UMLS:C0018681_Headache" idref
   = "p10">
   <certainty v = "high certainty"></certainty>
   <region v = "front"></region>
   <status v = "recurrence"></status>
   <parsemode v = "mode1"></parsemode>
   <sectname v = "report hospital course item"></sectname>
   <sid idref = "s1"></sid>
   <code v = "UMLS:C0239888_Headache  recurrent"></code>
   <code v = "UMLS:C0239886_Frontal      headache"></code>
</problem>
```

Figure 1. Example of simplified MedLEE output in XML format for the sentence *She has recurring frontal headaches*.

A simplified example of MedLEE output in XML format is shown in Fig 1 for the sentence "She has recurring frontal headaches". In Fig 1, the primary finding is **problem** with value **headache**, which has a **certainty** modifier **high certainty** corresponding to *has*, a **region** modifier **front** corresponding to *frontal*, a **status** modifier **recurrence** corresponding to *recurring*, and additional modifiers, which provide contextual information. In addition, codes are computed for primary findings and certain modifiers. In this work, we used UMLS codes. A code **C0018681** that is an attribute of the **problem** tag was assigned to the primary finding **headache** in the sentence without regard to modifiers. Additional codes, which are XML tags called **code**, correspond to the primary finding along with modifiers, and are intended to be as specific as possible. In this example, two UMLS codes, **C0239888** and **C0239886**, were assigned corresponding to *recurring headache* and to *frontal headache* respectively. The UMLS does not include a single code corresponding to *recurring frontal headache*; if it did, that single more specific code would have been assigned instead of the two former less specific codes.

Cao et al used MedLEE and co-occurence statistics to discover disease-finding associations in discharge summaries,[45,46] and later Chen et al used similar methods to detect disease-drug associations and their trends in both discharge summaries and the literature.[42,43] A χ^2 statistic was used as a measure of significance for associations, but because the large volume of data was sufficient to make any hypothesis test significant in a simple analysis, the χ^2 statistic was calibrated by a volume test adjustment denoted by $\varepsilon(\chi^2)$ and automatic determination of cutoff point for the number of true associations. The present study continues to build upon the previous work of our group by adapting the combination of NLP and statistical methods to acquire potential drug-ADE associations.

Methods

Materials and System Framework

The framework we propose for detecting drug-ADE associations from narrative reports involves five major phases, as shown in Fig 2: (1) collecting the set of reports to be mined; (2) processing the reports using NLP to encode clinical entities; (3) selecting drug and possible ADE entities; (4) reducing inappropriate information using a filter that excludes possible confounding factors, such as diseases/symptoms occurring before the use of therapeutic drugs, and another filter that excludes entities which were negated

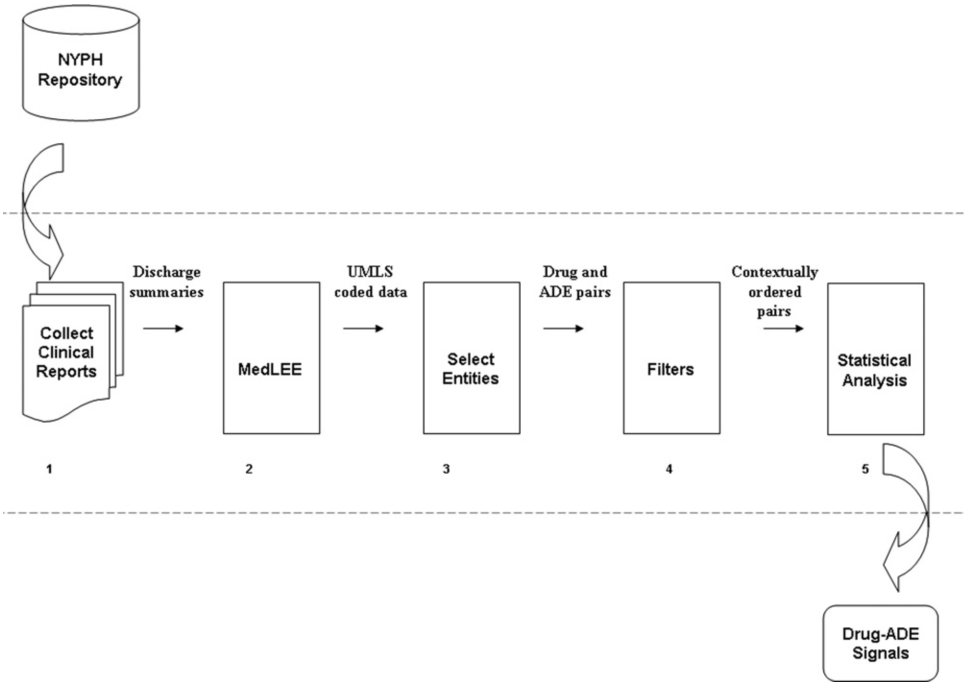


Figure 2. Overview of System Framework. The framework for detecting drug-ADE associations from narrative reports has five major phases: (1) data collection: collecting the set of reports to be mined; (2) data extraction: processing the reports using NLP to encode clinical entities; (3) data selection: selecting drug and possible ADE entities; (4) data filtering: excluding possible confounding information using two filters; and (5) statistical analysis: determining co-occurring drug-ADE candidates, and applying statistical methods to reveal associations between drugs and ADE candidates. The strength of associations were calculated and cutoffs were determined by co-occurence statistics adjusted by volume tests.

or which were noted as unlikely to have occurred; and (5) determining drug-ADE co-occurring pairs, and then applying statistical methods to reveal associations between drugs and ADE candidates. For this initial study we used discharge summaries of inpatients, and focus on drug-ADE detection occurring during hospital stays.

Phase 1 Data Collection

The data warehouse in NYPH maintains a variety of structured and unstructured patient information in the form of narrative reports, coded laboratory data, and pharmaceutical orders. Textual discharge summaries dictated in 2004 were collected for this study.

Phase 2 Data Extraction

The MedLEE system was used to parse and transform discharge summaries into a structured representation consisting of entities specified by UMLS CUIs and modifiers, as described in the Background Section. In this study, we modified MedLEE encoding to exclude some temporal modifiers (e.g., *exacerbated*) and degree modifiers (e.g., *slightly*) within the codes to avoid obtaining highly specific codes, such as **exacerbated dyspnea** (**C0853326**).

Phase 3 Data Selection

The UMLS codes that were extracted in the previous phase and that corresponded to the following semantic classes were used to select entities which were possible ADEs: *Finding* (T033), *Disease* or *symptom* (T047), *Mental or behavioral dysfunction* (T048), *Sign or symptom* (T184), and *Neoplastic process* (T191). Similarly, the UMLS codes that were extracted and that corresponded to the semantic classes *Pharmacologic Substance* (T121), *antibiotic* (T195), and *Clinical Drug* (T200) were used to select the medication entities. The UMLS table MRREL, which includes RxNorm (RxNorm vocabulary at the National Library of Medicine), NCI (national cancer institute), and PDQ (physician data query), defines several types of relationships between concepts that are related to generic classes and trade names of drugs, such as *trade name-of* and *has-trade name*. This was used to map all trade names to their generic names.[47–49]

Phase 4 Data Filtering

In phase 4, two filters are used to eliminate some clinical entities. One filter eliminates findings associated with modifiers corresponding to certain *certainty* values (negation, low certainty, workups), *past* events, or *family history* events. The second filter attempts to eliminate drug-indication sequences which are in the wrong chronological order. For pharmacovigilance, it is critical to obtain potential ADEs and drugs occurring within the appropriate time sequence (i.e., adverse drug events cannot occur before a drug is given). To avoid those events associated with indications occurring before the drug event, an additional contextual filter consisting of the section where the clinical information occurred was applied as a coarse estimation of the correct temporal order of events. Drugs mentioned in sections other than **Hospital Course** and **Medications**, were filtered out to eliminate medications not given during the Hospital Course. Typically, the medications section at NYPH contains medications the patient is on during a hospital stay, which may contain outpatient medications, which are usually continued during a hospital stay. Disease or pathological events mentioned in certain sections, such as *Chief Complaints*, *Diseases at Admission* and *History of Present Illness* (HPI) were filtered out since most of these are conditions related to the diseases and symptoms the patients have.

This filter was determined based on our initial work. When we did not include this filter, many of the drug associations we obtained corresponded to therapeutic associations which consisted mainly of the diseases and symptoms the drugs treated and were therefore not potential ADEs.

Phase 5 Drug-ADE Association

In this step, the χ^2 statistic adjusted with volume tests is used on the remaining co-occurring drug-ADE pairs to determine possible signals. First, drug-ADE pairs are collected for each report. This is accomplished using co-occurence within each discharge summary of the drug-ADE entities so that each drug entity that has not been filtered out is paired with each potential ADE entity that has not been filtered out. The pairs are then combined for all the reports and overall frequencies of the drug and ADE entities as well as the drug-ADE pairs are obtained as needed for the χ^2 statistic. Contingency tables for each drug-ADE pair across all the possible drugs and potential ADEs values were generated. All tables that had a frequency of less than 2 were excluded because they were unlikely to yield meaningful statistical results.

To test the hypothesis of no association between a drug and an ADE, the χ^2 statistic was used. For a detailed description of the method and the definition of cutoff point see Cao and colleagues.[45,46] In the present study, because the data are 2×2 tables with the same row margins, we computed the adjustment to the χ^2 p value that corresponds to tables with fixed row margins. Fixed row margin tests are partially conditional tests, where the conditioning argument is the variable that describes the row marginals. This conditioning guarantees that the margins of the table do not provide any evidence either in favor or against the null hypothesis of independence (i.e., no association). Fixed row margin volume tests have similar interpretation with the unconditional volume tests, that is, they can be interpreted as a distance from the surface of independence. The larger the distance, the stronger the association. For details on the method of computation of the fixed margin test and cutoffs see Cao et al.[45,46] For potential ADEs associated with a particular drug, a ranked list for potential drug-ADE pairs was generated based on the strength of the statistics. A no-intercept linear regression model was constructed for the drug to identify the cut-off point. Examples of contingency tables and calculation of χ^2 are presented in Table 1.

Evaluation

Six drugs and one drug class were chosen for evaluation, each of which had known short-term side effects. Among them were (1) three drugs (ibuprofen, morphine, warfarin), that have been on the market for a long-time and have known short-term side effects, (2) three drugs (bupropion, paroxetine, rosiglitazone) for which new ADEs were detected after 2004, and (3) one drug class (ACE inhibitors). The three drugs in group 1 were used to evaluate if the system could "detect" known drug-ADEs, and the three drugs in group 2 were used to simulate a historic roll back design where we used discharge summaries from 2004 to test whether the system has the potential to discover drug-ADEs before they became known (i.e., the ADEs associated with these drugs first became known after 2004). The drug class of ACE inhibitors was used to see if the system could detect the ADEs common to all drugs of one drug class.

To evaluate our drug-ADE detection system, a reference standard was constructed by a practicing physician who was presented with the drugs in the study. The physician formed the reference standard by summarizing the ADEs for each drug/drug class based on his medical knowledge and Micromedex, a well-respected, evidence-based and reliable reference material.[50] Many of the other drug reference databases contain long lists of adverse events for particular drugs, which constitute common allergic reactions and uncertain adverse drug effects. Common allergic reactions were not included in the reference standard if the expert determined that they were more related to the patients' physical conditions and not to the specific drug. By contrast, severe and rare adverse events that were supported by evidence demonstrating association with a particular drug were included.

Quantitative Evaluation

Recall and precision were used to assess the performance of our method. Recall was calculated as the ratio of the number of distinct potential drug-ADE pairs that were identified by our method over the total number of the corresponding drug-ADE pairs in the reference standard (i.e., TP/(TP + FN)). Precision was measured as the ratio of the number of distinct potential drug-ADE pairs returned by our method that were correct according to the reference standard divided by the total number of drug-ADE pairs found by our method (i.e., TP/(TP + FP)).

Qualitative Evaluation

The physician also analyzed and classified the drug-ADE associations detected by our system into four classes, again

Table 1 ■ Examples of Potential Drug-ADE Contingency Tables and Associations*

		2 × 2				
Drug	Potential ADE	Drug/+ADE+	Drug/+ADE−	Drug−/ADE+	Drug−/ADE−	$\varepsilon(\chi^2)$
Bupropion	feeling suicidal	11	177	116	24770	0.592585
Bupropion	motor retardation	5	183	25	24861	0.517752
Bupropion	tinnitus	4	184	20	24866	0.449479
Bupropion	extrapyramidal sign	3	185	10	24876	0.441856
Bupropion	stiffness	4	184	21	24865	0.439430
Bupropion	sleeplessness	9	179	162	24724	0.392614

ADE = adverse drug event.

*This table shows the first six drug-ADE pairs and is a subset of the potential drug-ADE pairs for bupropion.

using his knowledge and knowledge in reliable reference materials such as Micromedex and the Physicians' Desk Reference (PDR)[50]: (1) **known ADEs**: events that are known adverse drug events from premarketing RCTs and postmarketing surveillance. For example, the pair "paroxetine-dizziness" was classified as a known drug-ADE pair because "dizziness" is one of the known ADEs for paroxetine according to our reference standard; (2) **indication associations**: events that are indications (symptoms/disease) which the drug treats or symptoms directly associated with the disease indications. For example, the pair "Paroxetine-hallucinations" was classified as an indication association because paroxetine is used to treat an obsessive-compulsive disorder, and hallucinations are a manifestation of the disorder; (3) **remote indication associations**: events that are known to be consequences of the indication that the drug treats through a clinically plausible pathway. For example, the pair "ibuprofen-joint swelling" was considered a remote indication association because ibuprofen is used to treat pain, which could be caused by arthritis, and joint swelling is one of the symptoms of arthritis; and (4) **unknown associations**: events that are either conceptually poorly defined or currently unknown to be associated with the drug or with indications that the drug treats; For example, the pair "Paroxetine-thicken" was classified as unknown because "thicken" is a conceptually poorly defined concept.

Results

Data Statistics

The data in this study included 25,074 discharge summaries from NYPH. Co-occurrence data of drugs and potential ADEs in the corpus are summarized in Table 2a. One thousand nine hundred ninety-seven (1,997) unique drug concepts and 732 adverse event concepts were extracted. For the seven drugs or drug classes in our evaluation set, co-occurence data of drugs and potential ADEs are described in Table 2b.

Results of Quantitative Evaluation

One hundred thirty-two ADEs were found to be associated with the selected seven drugs. Overall, recall and precision were 0.75 and 0.31 for known ADEs respectively. Our analysis showed that the recall of the ACE inhibitor drug class was higher than that of individual drugs (1.00 vs. 0.63–0.87), whereas precision is similar for the class and the individual drugs (0.35 vs. 0.20–0.43).

Results of Qualitative Evaluation

We determined that 31% of the potential drug-ADE associations were known ADEs, 30% were indication associations, 33% were remote indication associations and 6% were unknown associations. Results of the qualitative analysis are shown in Table 3. The ACE inhibitors are known to cause cough among 30% of the patients, and our analysis showed that "cough" was ranked as the second among all the ACE inhibitors-ADEs Association.

Table 2a ▪ Summary of the Data that was Selected

Data in the Corpus	Count
2004 discharge summaries	25074
Total drug occurrences	143828
Total potential ADE occurrences	103362
Unique drug concepts	1997
Unique potential ADE concepts	732

ADE = adverse drug event.

Table 2b ▪ Summary of Co-Occurence Data for the Selected Seven Drugs

Drug	Total Documents	2 × 2 Tables*	Cutoff†
Ibuprofen	583	125	21
Morphine	490	128	22
Warfarin,	2040	189	10
Bupropion	188	124	32
Paroxetine	468	137	16
Rosiglitazone	287	119	10
ACE inhibitors	2482	257	14

ACE = angiotensin-converting enzyme.

*2 × 2 tables reflect number of potential drug-ADE associations for each drug.

†The cut-off represents the total number of potential drug-ADE associations selected as possible signals when ordered by $\varepsilon(\chi^2)$.

Results of the qualitative evaluation also suggested that our system has the potential to discover novel ADEs. Some ADEs associated with paroxetine such as "feeling suicidal", were detected as candidates for both indications/treatment and for new ADEs. Potential ADEs associated with rosiglitazone, such as "chest pain" and "shortness of breath", are symptoms of cardiovascular diseases, which were first discovered as a new type of ADE in 2007, leading to a new box warning of rosiglitazone.

Discussion

Our findings demonstrate that the system we proposed is feasible for pharmacovigilance. We were able to identify known drug-ADE with a performance of 75% for recall and 31% for precision. More importantly, our historic roll back experiments indicated that the system can potentially detect new ADEs prospectively.

The current study incorporates several important features. First, while other studies have focused on structured and coded data, this work took a completely different route using narrative data as the starting point for pharmacovigilance. The application of NLP unlocks rich information occurring in narrative reports. Data mining algorithms in pharmacovigilance miss important clinical data that is relevant for pharmacovigilance. Some important ADEs such as "fever" and "feeling suicidal" are generally only available in the narrative EHR reports. Some events, such as *nose bleeding* may occasionally be available as structured data (ICD9 code: 784.7), but documentation for this symptom may be found more frequently in clinical notes. The ability of NLP systems to extract a broad variety of events from clinical reports provides a valuable access to clinical information that would not be available otherwise. The work could be extended to combine structured with unstructured data. For example, structured data consisting of abnormal laboratory results and pharmacy orders would provide complementary information for detection of ADEs. Second, other studies analyzed associations in their databases using DPA, regression or other methods, where this study focused on the association statistics adjusted with volume tests, which have been

Table 3 ■ Qualitative Evaluation

		Associations Detected‡			
Drug* (Treatment Indications)	Reference Standard†	Known ADEs	Indication Associations	Remote Indication Associations	Unknown Associations
Ibuprofen (pain of rheumatoid arthritis, osteoarthritis, menstrual cramps, or mild to moderate pain)	constipation, diarrhea, dizziness, gas, headache, heartburn, nausea, stomach pain or upset	headache, achalasia, nausea, constipation	pain, pleuritic pain, chest pain, ache, referred pain, apyrexial, fever, chill, night sweat, hot flush	joint swelling, lesion, bacterial abscess, erythema, oral lesion, hemoptysis	
ACE inhibitors (hypertension and congestive heart failure)	cough, diarrhea, dizziness, headache, tiredness	cough, lethargy, dizziness, diarrhea, headache	chest pain, shortness of breath, syncope, orthopnea, pain	hyponatremia, decreased body weight,	vomiting, asymptomatic
Rosiglitazone (diabetes)	headache, weight gain, **symptoms of heart failure**	headache, **chest pain**, **left atrial hypertrophy**, **shortness of breath**	syncope, vertigo	tremor, pins and needle, cyanosis, colic abdominal	non-productive cough, erythema
Bupropion (depression and smoking cessation aid)	constipation, dizziness, drowsiness, dry mouth, headache, pruritus increased sweating, loss of appetite, nausea, vomiting, nervousness, restlessness, taste changes, trouble sleeping, weight changes, seizure, tinnitus **suicidal thoughts**	dizziness, abnormal sensation, difficulty, drugged state, fatigue, constipation sleeplessness, seizure, tinnitus, pruritus **feeling suicidal**	suicidal, visual hallucinations, moody, emotional, tremor, nightmare	motor retardation, fall, jumpy, stiffness, early satiety, extrapyramidal sign, energy increased, malingerer, rale, urge incontinence, bulimia, yellow sputum, emaciation	
Paroxetine (mental depression, obsessive-compulsive disorder, panic disorder, generalized anxiety disorder, social anxiety disorder)	agitation, chest congestion, chest pain, chills, cold sweats, confusion, difficulty breathing, dizziness, muscle pain or weakness, skin rash, **suicidal thoughts**	pain chest, drowsiness, orthostasis, dyspnea, agitation, dizziness, **feeling suicidal**	verbal auditory hallucinations, sleeplessness	Syncope, hormonal changes, intoxication, sleepy, thinness, numbness	thicken

ACE = angiotensin-converting enzyme.

*Only five out of seven drugs in our evaluation set are listed in this table due to space limitations; the conditions that drugs treat are shown in parentheses following the drug name.

†The reference standard constructed by the physician is shown in the second column, and new ADEs discovered after 2004 are shown in bold.

‡The associations obtained from our methods were categorized by the expert into four classes shown in the last four columns. New ADEs detected after 2004 based on data of 2004 are shown in bold.

shown to provide more clinically meaningful cutoffs for clinical associations.[48,51] Third, while work in current pharmocovigilance practice focused on retrospective investigations, our system highlights the potential for prospective surveillance which detects novel ADEs automatically and actively.

It is, however, a constant challenge to accurately identify and evaluate safety signals in a timely manner for pharmacovigilance. One factor affecting precision could be that some adverse events were detected correctly but they have not yet become known. Another factor is due to temporality and dependencies of clinical events. To infer causal associations between drugs and potential ADEs, it is important to recognize temporal sequences between drugs and these potential ADEs. In this investigation, we tackled the problem of temporality in text using an extremely simple contextual filter consisting of the sections in the discharge summary where the information was found. The strategy was somewhat successful since we were able to detect 75% of the known-ADEs whereas without that strategy we almost always detected indication associations. By contrast, the precision was only 31% because 63% of the associations (30% of the indication associations and 33% of the remote indication associations) in our qualitative analysis should have been eliminated but were not. Most of the false positives classified as indications were caused by two types of confounding information. One type was related to the diseases and indications the patient had because their associations with the medications were therapeutic associations and not side effects (e.g., "Paroxetine-dizziness"). This showed that our strategy for handling temporal information was not effective enough. Another type of confounder was due to indirect associations which occurred when a medication used to treat a particular disease and manifestations of the disease formed statistically significant pairs (e.g., "Paroxetine-hallucinations").

We have experimented with drugs which are used to treat the same diseases but have different safety profiles to differentiate ADEs from possible indication confounders. Our preliminary data have shown that rosiglitazone was associated with heart failure symptoms while other diabetes drugs tested (metformin, glipizide) were not. This does not, however, confirm that these heart failure symptoms are actually ADEs rather than treatment indications. Rosiglitazone may be more likely to be prescribed to severe and late stage diabetic patients, and these patients might be likely to develop comorbidities, such as heart disease, more often than patients on the other drugs. Stratification may solve the problem but it may be challenging. In addition, variable selection may have to be decided externally by an expert.

Inspired by work in Bioinformatics of characterizing interactions between genes, we applied mutual information (MI) and its property of data processing inequality (DPI) to help differentiate the direct and indirect types associations between clinical entities. This information theoretical approach using MI and DPI showed some promise for reducing false positives due to indirect associations, and is the focus of another paper.[52] As a further line of research, more sophisticated statistical methods that are able to account for the structure of a large database, and which are extensions of the proposed methodology, can be devised to differentiate between the different types of associations observed. This will also involve use of more sophisticated temporal models, use of information from other sources of clinical data, such as medication mentions in prior notes and prescription information. Additionally, use of other sources of knowledge, such as DXPlain or QMR will be explored.[53,54]

Another challenge related to our methods concerns the granularity of the codes corresponding to diseases and symptoms. The disadvantage of using highly granular terms is the dilution of the signals among medically similar ADEs. For example, there are more than 150 codes for cough in the UMLS corresponding to highly specific terms such as "cough on exercise", "postural cough", "brassy cough", and "increased frequency of cough". In this study, the modified version of MedLEE, which excluded some of modifiers, worked well for symptoms but this issue needs to be explored further. A commonly used coding system in the pharmacovigilance field is the Medical Dictionary for Regulatory Activities (MedDRA) which contains five hierarchical levels.[55,56] The level of "preferred term" is often used in pharmacovigilance systems because it is considered the appropriate level of granularity for that purpose. Similarly, granularity information in other knowledge sources in the UMLS (e.g., parent/child relations in MRREL, the UMLS knowledge source specifying relationships among entities) should also be helpful in meeting the challenge. Development of statistical approaches should also be considered. For example, Berry and Berry cleverly applied Bayesian methods to "borrow strength" and enhance diluted "signals" across multiple similar ADEs.[57] In subsequent studies, we will explore using MedDRA, the UMLS, and statistical approaches to help solve the granularity problem.

Our study had several limitations. Some of the limitations were caused by UMLS codes that were not well-defined. For example, a code "thicken" was an entity that was frequently extracted by MedLEE and then subsequently statistically associated with one of the drugs. A manual review indicated that some concepts, such as "thickened sigmoid" and "thickened valve leaflet" were encoded correctly as combinations of "thicken" with body location qualifiers but there were no individual UMLS codes that corresponded to either of the two complete concepts. Therefore, the single and poorly defined concept "thicken" was used. Another limitation is that, for our initial study, we included narrative reports for inpatients. As a result, our findings should be understood in the context of a sick patient population, which affected our results in several ways. First, the details in the documentation may be different. For example, "temperature increased slightly" might be less documented for outpatients than for inpatients because an increased temperature could be perceived as being more burdensome for inpatients because they are sicker. Second, inpatients may be more prone to have ADEs due to their weakened conditions and because they are more likely to be taking multiple medications. However, this limitation is due to the type of reports we focused on and not the methodology. In subsequent studies we will adapt our methods to a corpus consisting of multiple outpatient visits as well as hospital admissions, so the information relating to a patient will span a longer period and so that we will obtain a more varied patient population including healthier patients. However, that will likely introduce new challenges. A further limitation of this investigation is the fact that the reference standard was obtained using only one expert, and only seven drugs were tested. A more comprehensive evaluation will be undertaken in our future work.

Although the methods discussed in this work focuses on pharmacovigilance, the same methodology could readily be broadened to include adverse events associated with vaccines, devices, and procedures. Extension to these patient safety domains would involve using MedLEE to extract these types of events rather than just drug events, and using similar statistical processes to determine signals, although determination of the thresholds would have to be optimized. In future research, we will experiment with such an extension.

While our study proves the feasibility of our method more work is needed to establish our method as a surveillance system. The success of this system relies on the following two fundamental components of the proposed method: (1) use of NLP to generate useful data, and (2) creation of a statistical methodology that successfully deals with the data obtained. Both of these components are important and supplement each other. In particular, the identification of the appropriate threshold(s) is at the heart of successfully extending this feasibility paper to a real time pharmacovigilance system. What we have used in this study, which is based on previously published work, is an operational threshold that provided reasonable results. In our future work, we will be experimenting with a variety of ideas for identifying and evaluating this threshold. If we are successful, precision should be significantly improved.

Conclusions

Establishing safety profiles over the market life of a drug accurately and timely is a constant challenge in the field of pharmacovigilance, and is critical for patient safety. In this study, we provided a high throughput model and method to identify drug safety signals by mining narrative reports in the EHR. We demonstrated the potential of the method. To the best of our knowledge, this is the first study demonstrating the use of unstructured patient data, NLP, and statistics for pharmacovigilance. This paper provides a framework for the development of automated, active and prospective pharmacovigilance which could potentially unveil drug safety profiles and novel adverse events in a timely fashion.

References ■

1. McBride W. Thalidomide and congential malformation. Lancet 1961;2:1238.
2. Amery WK. Why there is a need for pharmacovigilance. Pharmacoepidemiol Drug Saf 1999 Jan;8(1):61–4.
3. Griffin JP, Pharmacovigilance JR. Coll, Londres: Physician's, 1992 Apr;26(2):197–8.
4. Available at: http://www.emea.europa.eu/. Accessed 3/21/09.
5. Available at: http://www.fda.gov/cder/aers/default.htm. Accessed 3/21/09.
6. Wood L, Martinez C. The general practice research database: Role in pharmacovigilance. Drug Saf 2004;27(12):871–81.
7. Sturkenboom M. Other Database in Europe for the Analytic Evaluation of Drug Effects. Pharmacovigilance, 2nd edn, Wiley, 2007.
8. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969–2002: The importance of reporting suspected reactions. Arch Intern Med 2005 Jun 27;165(12):1363–9.
9. Stephenson WP, Hauben M. Data mining for signals in spontaneous reporting databases: Proceed with caution. Pharmacoepidemiol Drug Saf 2007 Apr;16(4):359–65.
10. Venulet J. Possible strategies for early recognition of potential drug safety problems. Adverse Drug React Acute Poisoning Rev 1988 Spring;7(1):39–47.
11. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol drug Saf Oct-Nov 2001;10(6):483–6.
12. Rothman KJ, GS. Modern Epidemiology 1998, 2nd edn, Lippincott Raven, PA.
13. Hauben M, Zhou X. Quantitative methods in pharmacovigilance: Focus on signal detection. Drug Saf 2003;26(3):159–86.
14. Available at: http://www.who-umc.org/DynPage.aspx. Accessed 3/21/09.
15. Available at: http://yellowcard.mhra.gov.uk/. Accessed 3/21/09.
16. Burwen DR, LVL, Braun MM, Houck P, Ball R. Evaluating adverse events after vaccination in the Medicare population. Pharmacoepidemiol Drug Saf 2007 Jul;16(7):753–61.
17. Banks D, WE, Burwen DR, et al. Comparing data mining methods on the VAERS database. Pharmacoepidemiol Drug Saf 2005 September;14(9):601–9.
18. Available at: http://www.fda.gov/cdrh/MAUDE.html. Accessed 3/21/09.
19. Eland IA, Belton KJ, van Grootheest AC, et al. Attitudinal survey of voluntary reporting of adverse drug reactions. Br J Clin Pharmacol 1999 Oct;48(4):623–7.
20. Halpern SD, Barton TD, Gross R, et al. Epidemiologic studies of adverse effects of anti-retroviral drugs: How well is statistical power reported. Pharmacoepidemiol Drug Saf 2005 Mar;14(3): 155–61.
21. Coulter DM, the New Zealand Intensive Medicines Monitoring Programme, Pharmacoepidemiol d, Saf. 1998. Marine;7(2): 79–90.
22. Evans JM, McNaughton D, Donnan PT, MacDonald TM. Pharmacoepidemiological research at the Medicines Monitoring Unit, Scotland: Data protection and confidentiality. Pharmacoepidemiol Drug Saf 2001 Dec;10(7):669–73.
23. Gelfand JM, Crawford GH, Brod BA, Szazpary PO. Adverse cutaneous reactions to guggulipid. J Am Acad Dermatol 2005 Mar;52(3 Pt 1):533–4.
24. Waller P. Pharmacoepidemiology—A tool for public health. Pharmacoepidemiol Drug Saf 2001 Mar–Apr;10(2):165–72.
25. Hershman D, NA, Jacobson JS, et al. Acute myeloid leukemia or myelodysplastic syndrome following use of granulocyte colony stimulating factors during breast cancer adjuvant chemotherapy. J Natl Cancer Inst 2007 Febr;99(3):96–205.
26. Ray WA, Murray KT, Hall K, Stein CM. Atypical antipsychotic drugs and the risk of sudden cardiac death. N Engl J Med 2009 Jan 15;360(3):225–35.
27. Brown JS, Kulldorff M, Chan KA, et al. Early detection of adverse drug events within population-based health networks: Application of sequential testing methods. Pharmacoepidemiol Drug Saf 2007 Dec;16(12):1275–84.
28. Berlowitz DR, Miller DR, Oliveria SA, et al. Differential associations of beta blockers with haemorrhagic events for chronic heart failure patients on warfarin. Pharmacoepidemiol Drug Saf 2006 Nov;15(11):799–807.
29. Hauben M, Madigan D, Gerrits CM, Walsh L, Van Puijenbroek EP. The role of data mining in pharmacovigilance. Expert Opin Drug Saf 2005 September;4(5):929–48.
30. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf 2002;25(6):381–92.

31. Hauben M. Application of an empiric Bayesian data mining algorithm to reports of pancreatitis associated with atypical antipsychotics. Pharmacotherapy 2004 September;24(9):1,122–9.
32. DuMouchel W, Smith ET, Beasley R, et al. Association of asthma therapy and Churg–Strauss syndrome: An analysis of postmarketing surveillance data. Clin Ther 2004 Jul;26(7):1092–104.
33. Baruch JJ. Progress in programming for processing English language medical records. Ann N Y Acad Sci 1965 Aug 6;126(2):795–804.
34. Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. Proc AMIA Symp 2001: 17–21.
35. Hahn U, Romacker M, Schulz S. MEDSYNDIKATE—A natural language system for the extraction of medical information from findings reports. Int J Med Inform 2002 Dec 4;67(1–3):63–74.
36. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004 Sept–Oct;11(5):392–402.
37. Chen L, Friedman C. Extracting phenotypic information from the literature via natural language processing. Stud Health Technol Inform 2004;107(2):758–62.
38. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003 Dec;36(6):462–77.
39. Bates DW, Evans RS, Murff H, et al. Detecting adverse events using information technology. J Am Med Inform Assoc 2003 Mar–Apr;10(2):115–28.
40. Honigman B, Lee J, Rothschild J, et al. Using computerized data to identify adverse drug events in outpatients. J Am Med Inform Assoc 2001 May–Jun;8(3):254–66.
41. Rebholz-Schuhmann D, Kirsch H, Arregui M, et al. EBIMed—Text Crunching to Gather Facts for Proteins from MEDLINE, BioInformatics 2007 Jan 15;23(2):e237–44.
42. Narayanasamy V, Mukhopadhyay S, Palakal M, Potter DA, Trans M. Mining transitive associations among biological objects from text. J Biomed Sci 2004 Nov–Dec;11(6):864–73.
43. Cohen KB, Hunter L. Getting started in text mining. PLoS Comput Biol 2008 Jan;4(1):e20.
44. Rindflesch TC, Tanabe L, Weinstein JN, Hunter LE. Extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 2000:517–28.
45. Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. AMIA Annu Symp Proc 2005:106–10.
46. Cao H, Hripcsak G, Markatou M. A statistical methodology for analyzing co-occurrence data from a large sample. J Biomed Inform 2007 Jun;40(3):343–52.
47. Liu S, MW, Moore R, et al. Prescription for electronic drug information exchange. IT Prof 2005;7(5):17–23.
48. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: An initial study. J Am Med Inform Assoc 2008 Jan–Feb;15(1):87–98.
49. Available at: http://www.nlm.nih.gov/research/umls/. Accessed 3/21/09.
50. Available at: http://www.micromedex.com/. Accessed 3/21/09.
51. Chen E, Stetson PD, Lussier YA, et al. Detection of practice pattern trends through natural language processing of clinical narratives and biomedical literature. AMIA Annu Symp Proc 2007;11:120–4.
52. Wang X, Hripcsak G, Friedman C. Characterizing environmental and phenotypic associations using information theory and electronic health records. AMIA Summit on Translational BioInformatics, 2009.
53. Barnett O, Hoffer E, Feldman M, et al. 20 years later—What have we learned. AMIA Annu Symp Proc 2008;6:1201–2.
54. Parker RC, Miller RA. Creation of realistic appearing simulated patient cases using the internist-1/QMR knowledge base and interrelationship properties of manifestations. Methods Inf Med 1989 Nov;28(4):346–51.
55. Brown EG. Using MedDRA: Implications for risk management. Drug Saf 2004;27(8):591–602.
56. Bousquet C, Henegar C, Louet AL DP, Jaulent MC. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. Int J Med Inform 2005 Aug;74(7–8):563–71.
57. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. Biometrics 2004 Jun;60(2):418–26.