EXPLORING ADVERSE DRUG EFFECT DISCOVERY FROM DATA MINING OF CLINICAL NOTES

By

Joshua Carl Smith

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August, 2012

Nashville, Tennessee

Approved:

Professor Randolph A. Miller

Professor Joshua C. Denny

Professor W. Anderson Spickard, III

Professor S. Trent Rosenbloom

ACKNOWLEDGEMENTS

I would like to thank my thesis committee for their guidance throughout this work. I would especially like to thank my advisor, Randy Miller, who has guided my work over the past three years, helped me to develop my research interests, and taught me a great deal. I would also like to thank Cindy Gadd and Kevin Johnson for all of their work advancing my education in biomedical informatics.

I am grateful for many others who gave invaluable assistance. I would like to thank Josh Denny and his team, including Lisa Bastarache and Raquel Zink, for their assistance using KnowledgeMap and SecTag and for other technical support. I would like to thank Tom Lasko for his input on data mining, Steve Brown for providing information on the NDF-RT, Jonathan Schildcrout for his help with project design, and Mike Assink for his assistance with the SD. I would like to thank Cindy Chen and Hui Nian for their help with the statistical analysis of the DEB.

I would also like to express my gratitude to the Department of Biomedical Informatics. I appreciate the support of the students, faculty, and staff, without which I could not have completed this project. I am grateful to the National Library of Medicine and the National Institutes of Health for their support (T15 007450, R01 LM007995, R01 LM010828).

Finally, I am very grateful for the support of my friends and family. I know I would not be where I am today without my parents, Richard and Nancy Smith, and their unconditional love, support, and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP ..... natural language processing

H&P ..... history and physical exam

AE ..... adverse effect

ADE ..... adverse drug event

ADR ..... adverse drug reaction

EMR ..... electronic medical record

FDA ..... Food and Drug Administration

NSAID ..... nonsteroidal anti-inflammatory drug

MI ..... myocardial infarction

EMA ..... European Medicines Agency

WHO ..... World Health Organization

SRS ..... spontaneous reporting system

AERS ..... Adverse Event Reporting System

EudraVigilance ..... European Union Drug Regulating Authorities Pharmacovigilance

PRR ..... proportional reporting ratio

ROR .....reporting odds ratio

BCPNN .....Bayesian confidence propagation neural networks

IC.....information component

EBS ..... empirical Bayes screening

PPV .....positive predictor value

VAE ..... vaccine adverse event

NDF-RT .....National Drug File-Reference Terminology

UMLS .....Unified Medical Language System

CUI.....Concept Unique Identifier

KMCI .....KnowledgeMap Concept Identifier

SD ..... Synthetic Derivitive

MEDLINE.....Medical Literature Analysis and Retrieval System Online

MeSH ..... Medical Subject Headings

SPL.....structured product label

DEB..... Drug indications and adverse effects Evidence Base

CHAPTER I

INTRODUCTION

When considering both recommended and excessive medication dosages, almost all drugs have potentially serious adverse effects. Adverse effects (AEs) are any harmful or unintended reactions to medication. AEs can occur at doses normally used for treatment or because of overdose. Due to known limitations in premarketing medication trials (1–3), identification of adverse effects requires vigilant post-marketing surveillance. Agencies and investigators have recognized the potential for Electronic Medical Record Systems (EMRs) to characterize clinical correlations in large numbers of patients using prescription medications.

This thesis project explores one aspect of using EMRs to detect single drug ingredient AEs – specifically, the feasibility of mining History and Physical (H&P) exam notes to detect concurrent mentions of single drug ingredients and clinical findings (including both symptoms and diseases). In the remainder of this document, the term "drug" will refer to a single (active) drug ingredient (as opposed to a multi-component "combination" medication or an inert ingredient), unless otherwise stated. Our approach identifies drug and finding concepts using Natural Language Processing (NLP) of clinical text. We hypothesize (based on previous work (2; 4–6)) that drug-finding pairs occurring in a higher-than expected number of records signify an underlying relationship between drug and finding. If one can distinguish which pairs occur for known reasons, then one

can postulate that the remaining pairs occur for unknown reasons -- especially previously unrecognized potential adverse effects.

Our project involved three phases. In Chapter III of this thesis, I describe the creation and evaluation of a knowledge base (KB) of known drug-finding pairs. I developed this KB from multiple existing reference sources containing structured data. I used automated methods to extract, reformat, and combine the information. In Chapter IV, I describe the extraction of drug-finding pairs from a corpus of de-identified H&P notes using NLP. Further processing required generalization of specific drug terms (e.g., mapping a brand name or a specific dose form to a common generic drug descriptor). I calculated statistical measures of the strength of correlation between drug and finding concepts that appeared across large numbers of notes. Finally, I discuss application of the (Chapter III) drug-finding KB to categorize the (Chapter IV) correlations as either known and/or unknown drug-finding correlations. I discuss the project results and the limitations of this study.

CHAPTER II

BACKGROUND

Introduction

Most drugs have potentially serious, medication-specific adverse effects (AEs) at therapeutic or excessive doses (we detail our definition of AE in Figure 1). The U.S. Food and Drug Administration (FDA) requires multi-phase clinical testing (described below) in an attempt to ensure efficacy and safety of prescription medications sold in the marketplace. Due to known limitations in premarketing clinical trials (1–3), identification of adverse effects also requires vigilant post-marketing surveillance. Some AEs remain unknown until after a drug has been approved for clinical use and used by large numbers of people.

After extensive in vitro and animal research, developers of new medications must undertake FDA-mandated pre-marketing clinical trials using human subjects in three phases (7). In Phase I testing, researchers administer the new drug to a small group of volunteer subjects, generally between 20 and 80 people. This initial testing determines the metabolism and pharmacologic action of the drug in humans, determines a safe dosage range, and identifies side effects associated with different dosages. Phase I trials may include healthy individuals or patients with specific diseases. In Phase II, researchers test the drug in a controlled trial on a somewhat larger group, usually between 100 and 300 volunteers. This phase studies the effectiveness of the drug in treating patients with one or more targeted conditions. Phase II also further evaluates safety and

determines common side effects. A drug still promising after the first two phases undergoes Phase III trials. These include both controlled and uncontrolled studies typically enrolling 1000-3000 individuals. Phase III trials determine the overall risk-benefit relationship for the new drug in treating a particular condition.

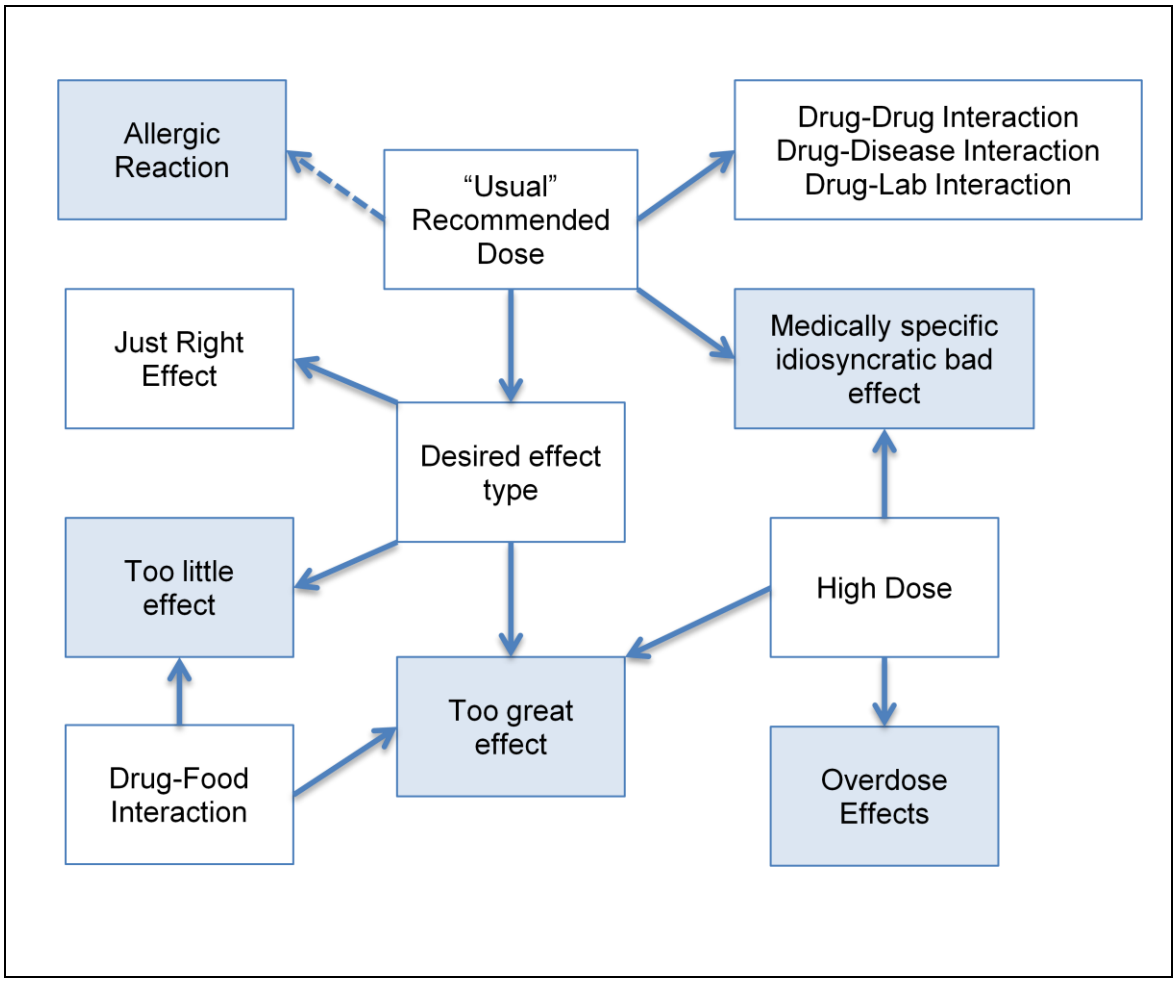

Figure 1. Illustration of drug effects. We define a "single drug ingredient" AE as those effects shaded in blue. Arrows represent a "may lead to" relationship.

Some AEs do not arise in pre-market clinical trials for a variety of reasons (1). First, such clinical trials are relatively small (typically fewer than 3000 people), so they cannot detect rare side effects that occur in fewer than 1 in 10,000 patients. Upon

reaching the market, hundreds of thousands, or even millions, of people might use a new drug (3). Second, inclusion criteria for FDA pre-marketing trials are often restrictive in terms of age, race, gender, and health status. While necessary for accurate testing, a small, uniform population sample is unlikely to mimic the diverse spectrum of individuals who will eventually use the drug. Additionally, controlled trials rarely emulate the exact conditions of medication use by the public at large. Pre-marketing trials cannot and do not fully explore effects of comorbid conditions, ranges of dosing, duration of administration, and interactions with other medications taken simultaneously. Unfortunately, this means that many medication AEs only appear after a large number of people have taken a drug over much longer periods of time than typical durations of pre-marketing trials (2; 3).

Examples of AEs discovered in post-marketing surveillance

Rofecoxib (Vioxx), a COX-2 selective nonsteroidal anti-inflammatory drug (NSAID), was marketed by Merck to treat arthritis, acute pain in adults, and dysmenorrhea (8). The FDA approved the sale of rofecoxib on May 21, 1999, and it became one of the best-selling prescription drugs worldwide. On September 30, 2004, after over 80 million people had taken the medication, Merck voluntarily withdrew the drug because of an increased risk of myocardial infarction (MI) and stroke that only appeared after 18 months of use of the medication (9). Rofecoxib is no longer sold in the United States, but was available for over 5 years before these serious effects were discovered. Valdecoxib (Bextra), another COX-2 NSAID, was also withdrawn from the market due to an increased risk of heart attack and stroke. Due to these problems, the

FDA warns against the use of Celecoxib (Celebrex), another popular COX-2 NSAID, by patients at risk of heart disease, but there is at best contradictory evidence that it is associated with similar side effects.

Rosiglitazone (Avandia) is a thiazolidinedione class drug that was widely used in the treatment of Type II diabetes. The drug was approved by the FDA in 1999 and by 2006 had sales of approximately $2.2 billion (10). In 2007, an analysis published in the *New England Journal of Medicine (NEJM)* showed that rosiglitazone use increased the risk of MI and other adverse cardiovascular effects (11). Despite some criticism of the NEJM study, and subsequent studies that provided mixed results, public confidence in, and sales of rosiglitazone dropped rapidly(12). Three years after the NEJM study, on September 23, 2010, the European Medicines Agency (EMA) removed the drug from the European market and the FDA imposed significant restrictions on its use in the United States (13; 14). While the NEJM study eventually discovered these AEs, the question remains whether information was available earlier that might have led to a more timely detection. Another anti-diabetic drug in the thiazolidinedione class, pioglitazone (Actos), has a lower risk for MI than rosiglitazone, but was linked to increased risk of bladder cancer in 2011 after four years on the market. As a result, France suspended the sale of pioglitazone, and the FDA has issued a warning that it should not be used in patients with a history of bladder cancer (15).

The HMG-CoA reductase inhibitors (commonly known as statins) have been available in the U.S. since the 1980s. This class includes the frequently prescribed drugs simvastatin (Zocor), atorvastatin (Lipitor), and rosuvastatin (Crestor). More than 20 million Americans currently take statins to lower their cholesterol levels (16). On

February 28, 2012, the FDA announced that statin users have a dose-related increased risk for memory loss and for developing type II diabetes mellitus (17). Using a conservative estimate that 1/200 patients treated with higher doses of the statins develop diabetes, suggests that up to 100,00 new statin-induced cases of diabetes would occur in the US alone (16).

Finally, in recent news, on April 11, 2012, Johnson and Johnson was ordered to pay $1.2 billion in fines on charges that they minimized or concealed the dangers associated with the drug risperidone (18). Risperidone is an anti-psychotic drug used to treat schizophrenia, bipolar disorder and behavior problems in teenagers and children with autism. The adverse effects that were allegedly minimized or withheld include weight gain, increased risk of diabetes, and stroke.

Prevalence of ADRs/AEs

Studies have shown a surprisingly high prevalence of adverse drug reactions AEs and ADRs, in terms of both generic effects, such as allergic reactions, and medication-specific pharmacological effects. In 2004, a British study of 20,000 inpatients showed that 6.5% of admissions were associated with ADRs. The ADR directly led to the admission in 80% of those cases (19). A 2008 systematic review of 25 studies involving over 100,000 patients found that approximately 5.3% of hospital admissions were associated with ADRs. The ADR rate was higher in elderly patients, who more commonly receive multiple medications (20). In 2010, a study of patients admitted to a 1250-bed hospital in Dordrecht, The Netherlands, found that 19% to 29% of admissions

to the Departments of Internal Medicine, Cardiology, and Pulmonology were due to ADRs (21). Adverse effects are a serious threat to patient safety.

The withdrawal of drugs such as rofecoxib and rosiglitazone, as well as improved warnings for pioglitazone and the statins, illustrates that FDA monitoring can eventually detect dangerous AEs in the market. Nevertheless, the key question remains whether one could detect such important and unanticipated AEs sooner after a drug reaches the marketplace (1).

Pharmacovigilance & Post-marketing surveillance

The science of pharmacovigilance comprises detection, assessment, understanding, and prevention of AEs (22), including AEs at both normal and excessive doses. Concerted and coordinated pharmacovigilance efforts began in the 1960s, including those of the World Health Organization (WHO) and the FDA. These followed the widespread, highly publicized, and tragic side effects of thalidomide administration during pregnancy. Pharmacovigilance includes pre-marketing risk assessment of newly developed drugs, ongoing risk minimization, and post-marketing surveillance (23). Post-marketing surveillance is essential for identifying a medication's AEs since it is unlikely that pre-marketing trials can be comprehensive enough to do so.

In the 1960s, Spontaneous Reporting, the process of healthcare professionals reporting suspected ADRs to a national agency or to the drug manufacturer, became the standard method of gathering data for post-marketing surveillance. Most countries collect such information in databases known as Spontaneous Reporting Systems (SRS). United States laws require drug manufacturers to submit reports of any suspected AEs to

the FDA. In addition, hospitals, healthcare professionals, and patients can submit spontaneous reports to the FDA through a program called MEDWATCH (24). The current FDA AE database, the Adverse Event Reporting System (AERS) contains over 2 million reports (24). In addition to the national regulatory authorities (NRAs) of each European country, the EMA maintains an EU-wide spontaneous reporting database known as EudraVigilance (European Union Drug Regulating Authorities Pharmacovigilance) (25)(26). In 1968, the WHO set up the International Drug Monitoring Programme. Today, it pools data from the spontaneous reporting systems of over 80 countries to assess post-marketing risk (27).

SRS are well known for under-reporting, over-reporting, and duplicate reporting. Clinicians often under-report due to lack of time, education, and financial compensation, as well as fear of revealing medication errors, and a generally negative attitude towards reporting activity (25). Additionally, the decision to report requires individual to make the connection between the event and the administration of the suspected drug; incidents that occur rarely, or seem disconnected from the drug administration for some reason, may go unreported. Examples of under-reporting include physicians not reporting effects that they do not consider significant or missing data in significant cases. Examples of over-reporting might include physicians reporting known, well-understood, and common AEs. Duplicate reports can be generated by physician or patient reports to the FDA followed by reports of the same incident by the drug manufacturer. SRS also lack information on the number of individuals actually consuming a drug and contain limited temporal information (28). Additionally, there are significant differences in the reporting of serious versus non-serious AEs, as well as trends relating to time on the market and the

number of prescriptions written (29). Oftentimes, these reports are focused on known AEs.

The Medical Dictionary for Regulatory Activities (MedDRA) includes concepts such as diseases, diagnoses, signs, symptoms, therapeutic indications, and medication errors(30) that categorize AEs. An international group of NRAs, including the FDA, EMA, WHO, and others, manages MedDRA. Trained staff use MedDRA to process and encode spontaneous reports. MedDRA has become the required or preferred terminology for reporting AEs throughout the world, facilitating collaborative data exchange and analysis (31).

Traditionally, efforts to monitor spontaneous reporting databases have focused on review of individual case reports. Newer methods facilitate analysis of aggregate data for purposes of detection and evaluation of what are known as "signals" (32). Signals are defined as threshold-based indicators that suggest a particular medication produces an AE, as detected by statistical, computational, or data-mining techniques in pharmacovigilance databases. Most methods use what is known as disproportionality analysis – that is, methods that measure the extent to which a given AE is disproportionally reported with a given drug (32). These methods have been validated in drug safety research, but they are only exploratory; one cannot draw conclusions solely from disproportionality analysis or any other single method (33). Definitive proof can only come from a convergence of experimental, clinical, and statistical research.

Signal Detection Methods and Examples

One of the most popular disproportionality analysis methods is the Proportional Reporting Ratio (PRR). Developed in 2001 by researchers using the UK SRS, the PRR is analogous to the concept of epidemiology concept of proportional mortality ratios (PMR) (34) (35). The PRR for a particular AE is defined as the ratio of the particular AE among all AEs for the given drug, divided by the ratio of the particular AE among all AEs for all drugs. One major drawback for the PRR, however, is that it is "numerator-based."

When estimating the prevalence of an AE with a particular drug, the "numerator" is the number of users experiencing an AE, while the "denominator" is the total number of individuals using the drug (34). Since the focus of a spontaneous report is on a user experiencing an AE, SRS do not contain information on the total number of users taking the drug in the population. Therefore, most methods are numerator-based.

One of the strengths of the PRR is that underreporting of adverse events should not influence the PRR, given that the AE in question is equally underreported as the aggregate of other AEs in the SRS. Unfortunately, the PRR also suffers from the same weaknesses of the PMR. The size of the numerator influences the size of the denominator and thus distorts the PRR; therefore a drug can appear to increase the risk of an AE solely because it reduces the risk of another AE (36). The PRR is frequently used in the EMA's EudraVigilance (37).

Another common numerator-based method is the Reporting Odds Ratio (ROR). It is similar to the PRR, but it is computed by viewing the SRS database as a case-control study; unlike the PRR, the ROR excludes any cases from the denominator that are suspected of being related to exposure of the drug in question. So unlike the PRR, the

ROR is an unbiased estimator of the risk ratio (36). Researchers frequently use the ROR to analyze data from the FDA's AERS (30).

Other statistical measures include Pearson's Chi-Square and the Poisson probability test. The chi-square test is used as a test of independence between two variables – whether a patient was exposed to the drug and whether the patient suffered the adverse effect. The Poisson distribution, often used to model rare events in large samples, is popular as well. Research has shown that methods such as the PRR, ROR, Poisson probability test, and the Chi-square test are broadly comparable when four or more cases per combination have been collected (38). Correlation analysis, multivariate regression, and Bayesian logistic regression are used, as well.

In addition to statistics, data mining methods have been used in post-marketing signal detection as well. Bayesian Confidence Propagation Neural Networks (BCPNN), a form of Bayesian data mining, used ideas from Information Theory to calculate a quantity known as the information component (IC) for each possible drug-event combination in a database. Signal detection is based upon this value and the time trend of the data (34). It is one of the primary methods used by the WHO. Empirical Bayes Screening (EBS), another Bayesian data mining technique, has been applied to the FDA's AERS. EBS ranks drug-event combinations according to how large the number of reports for a given combination is compared with what would be expected if the two were statistically independent (31). Unlike BCPNN, which provides a stand-alone measure for each drug-event combination, EBS provides only an overall ranking of drug-even combinations. A more thorough discussion on the strengths and weaknesses of many of these methods can be found in (33; 35; 36; 39).

Harpaz and Friedman at Columbia have done groundbreaking work in applying data-mining techniques to investigate drug interaction adverse effects – AEs that are caused by a specific combination of two or more drugs. Using popular association rule mining techniques, they investigated associations in the FDA's AERS between sets of drugs and sets of findings (40; 41). Similarly, they later applied the biclustering paradigm to the AERS to identify drug groups that share a common set of AEs. Application of that information could allow researchers to gain insight into the etiology of AEs (42).

As late as 2005, it was suggested that no major stakeholders had the goal of hypothesis-free examination of large databases in efforts to find new AEs (1). While methods exist which attempt to compensate for the shortcoming of SRS (43), methods that focus on new data sources with fewer limitations must also be explored.

Identifying AEs From EMR Data

Spontaneous reporting databases once stood as the only systems that had the necessary data in machine-readable formats for large-scale AE signal detection. However, with the proliferation of EMRs, substantive drug and finding data now exists at most healthcare institutions, providing an important opportunity for post-marketing surveillance (44).

In recent years, there have been several attempts to scan EMRs for AEs. In 2001, Honigman, et al., (45) were able to identify known AEs in the ambulatory setting using diagnosis codes, allergy rules, computer event monitoring, and text searching. These methods had a sensitivity of 58%, specificity of 88%, and a positive predictor value

(PPV) of 7.5%. In 2003, Murff, et al., (46) searched free text discharge summaries for trigger words that indicated a possible adverse event. The study achieved sensitivity, specificity, and PPV of 69%, 48%, and 52%, respectively. In the 2004-2007 NIH-sponsored TIME (Tools for Inpatient Monitoring Using Evidence) project, Miller and colleagues at Vanderbilt correlated inpatient laboratory test abnormalities with CPOE-based medication orders to discover time dependencies of known AEs and attempt to find new AEs (47–49). In 2004, Field, et al., (50) examined multiple strategies for identifying AEs in older patients in ambulatory clinics, including manual review of clinician and administrative incident reports, electronic codes, and automated text searching of patient notes for known drug-AE combinations. They found far more instances of AEs than were actually labeled in the records, suggesting that one should use multiple strategies to detect AEs in clinical notes. In 2009, Hazlehurst, et al., (51) used NLP to detect AEs in clinical notes. As part of the Vaccine Safety Datalink collaboration, researchers modified an existing NLP tool so that it could recognize possible general vaccine adverse events (VAEs), and specifically gastrointestinal-related VAEs. The authors believed their reported sensitivity, specificity, and PPV (75%, 97%, and 89%, respectively) improved on previous work due to more sophisticated NLP methods.

At Columbia, Friedman, Wang, and colleagues showed that NLP could effectively identify disease, symptom, and AE concepts in EHRs. In 2008, they reported extracting diseases and related symptoms with a recall of 90% and precision of 92% from discharge summaries (52). They also used NLP to extract disease-drug co-occurrence statistics discharge summaries, as well as from Medline articles (6). Building on their previous work, Wang and Friedman were later able to use NLP to extract both disease-

symptom and drug-AE pairs from clinical notes, filtering out sections of the note that are not directly related to patient experiences (such as "family history") to improve precision and recall to 0.92 and 0.90 for disease-symptom pairs and 0.31 and 0.75 for drug-AE pairs (5). In 2009, Wang & Friedman, et al., used NLP to identify drugs and findings in a collection of 25,074 discharge summaries. Using co-occurrence statistics, they found correlations between seven drugs/drug classes and their known AEs. They had a recall of 0.75 and a precision of 0.31 for the known AEs, and based upon dates of the discharge summaries, showed that novel AEs would likely have been detected using their methodology (4).

The FDA's recent Sentinel Initiative aims to "create a linked, sustainable system that will draw on existing automated healthcare data from multiple sources to actively monitor the safety of medical products continuously and in real time" (53). Sentinel will monitor drug safety and, eventually, all FDA-regulated products. This will include data mining of healthcare information stakeholders (i.e., insurance companies and hospitals). After a successful pilot program, the FDA is developing and implementing Sentinel in stages. Key project areas include: evaluation of potential data sources; evaluation of existing methods of signal detection; engagement of patients, consumers, and healthcare professionals; evaluation of potential database models; and the evaluation of privacy regulations.

Drug Indication and Adverse Effect Information Sources

When analyzing the co-occurrences among drugs and findings in EHRs, correlations (signals) will be identified not only between drugs and novel AEs, but also

between drugs and indications and drugs and known AEs. In order to discover new AEs, one must distinguish these unknown signals from those that are already understood. Several resources exist that catalog drug indications and adverse effects. DailyMed is a NLM-operated website that makes FDA-approved prescription drug labels available to the public (54). Commercial resources such as Micromedex, First Databank (FDB), UpToDate, and many others provide drug information including treatments and dose amounts for a given conditions or AEs for specific medications (55–57).

Despite this wide variety of drug knowledge resources, there is no definitive source with all information in a structured format. In 2010, Wang, et al., attempted to compile drug indication information from a combination of sources: the FDA AERS, SemMed – a database generated from NLP on MEDLINE abstracts, and the National Drug File-Reference Terminology (NDF-RT) (58). In 2011, Li, et al., applied information from Micromedex, NDF-RT, and the AERS to identify the reasons for prescriptions for 20 drugs mentioned in EHR discharge summaries. They achieved 62% sensitivity, 93.9% specificity, 90% precision, and an F-measure of 73.9% (59).

Summary

As proposed by the MOMENT project (47–49) and similar to Wang, et al. (4), we used NLP on narrative clinical reports to identify co-occurrence of drugs and findings, and to study the feasibility of using this data for identifying novel adverse effects. Unlike Wang, et al., the current project uses History and Physical exam (H&P) notes instead of discharge summaries, and captures data from a much larger corpus of notes. While discharge summaries may contain major findings, therapies, and diseases, we believe

H&Ps provide a deeper and richer picture of patient findings and symptoms prior to what happened during a hospital admission. Since we are interested in discovering AEs related to prescription medications that the patient is taking outside of the hospital, we believe it is better to capture medications and symptoms at time of admission (when H&Ps are recorded). Similar to (5), we used clinical note section header data to restrict the sections from which we mine concepts, eliminating potential "false positive" terms from sections such as "Plan" and "Family Medical History." Additionally, we developed an automated methodology to generate a knowledge base of known drug-finding pairs, and to use this knowledge to identify known drug-finding pairs from our results. We did not intend to discover novel adverse effects at this stage of research, but we hoped to re-identify AEs that were discovered using post-marketing surveillance in the past and explore the potential for using this approach to discover novel associations in the future.

In summary, previous studies have examined the co-occurrence of drug and findings concepts using NLP, but others often performed their research on discharge summaries, which could confound drug effects prior to hospital admission with drug effects that occurred post admission. H&Ps contain a richer set of findings and symptoms because, while discharge summaries catalog the course of treatment in-hospital, H&Ps attempt to describe the patient's health more fully. Additionally, the Vanderbilt Synthetic Derivative provided us with the opportunity to use far more records than most previous studies, potentially allowing us to identify rare effects. Finally, through the creation and application of a drug-finding evidence base, we automatically identified known indications and adverse effects from the drug-finding correlations

discovered in the corpus of H&P notes, improving our ability to identify previously unknown correlations.

Project Specific Tools

Unified Medical Language System

The Unified Medical Language System (UMLS) is a collection of controlled vocabularies and ontologies in the biomedical sciences and healthcare. The UMLS is primarily composed of three knowledge sources (databases) – the Metathesaurus – database of vocabularies that contains information about biomedical and health-related concepts; the Semantic Network – information on the semantic types of and relationships between concepts in the Metathesaurus; and the SPECIALIST Lexicon – a source of lexical information for use with NLP tools (60–63).

We used UMLS version 2009AA throughout this project because it was the version of UMLS being currently used by our tools the KnowledgeMap Concept Identifier and MedEx (see below). We made significant use of the semantic types, UMLS Concept Unique Identifiers (CUIs), the UMLS co-occurring concepts table (MRCOC), and two drug vocabularies included in the UMLS – RxNorm and NDF-RT.

KnowledgeMap Concept Identifier

The KnowledgeMap Concept Identifier (KMCI) is a natural language processing (NLP) tool developed at Vanderbilt by Denny, Miller, Spickard, et al. Originally developed for use in medical education (64), it has been extended for use in clinical

research. KMCI is an NLP tool that indexes Unified Medical Language System (UMLS) concepts that occur in an input document. KMCI uses UMLS-derived resources, along with locally developed and publically available components, for word and term normalization, language processing, and concept identification. When candidate concepts have multiple UMLS matches, KMCI resolves ambiguous concepts using previously matched concepts and document context.

KnowledgeMap is used extensively at Vanderbilt for both medical education and research. KMCI is used to index concepts in the medical school curriculum to allow students easy access to information from course documents on a particular topic (64). KMCI has been used to extract EKG findings from clinical reports and correlate them with patient medication administration records in order to identify patients with prolonged QT intervals (65). KMCI was also used to recognize clinical text descriptions of colonoscopy screening events, status of the procedures' completion, and the dates the procedure was performed (66). It was later extended to better identify colorectal cancer in EMRs by recognizing three additional tests – flexible sigmoidoscopy, fecal occult blood testing, and double contrast barium enema – and determining whether testing was planned or completed and to estimate the date of completed tests (67). Another research project used KMCI to extract noun phrases from article titles in the American Journal of Epidemiology and, using heuristic rules, identify terms that contained epidemiologic exposure (68). KMCI has also been used for phenotype identification algorithms (69; 70).

SecTag

SecTag is another Vanderbilt-developed NLP tool used in clinical research. Developed by Denny, Miller, Spickard, et al., SecTag is used to identify section headers in clinical notes. The SecTag algorithm uses a locally developed lexicon of "clinical note section header" terms and heuristics to identify sections in H&P notes, such as "History of Present Illness," "Medications," or "Family Medical History." SecTag identifies not only major headings, such as "Cardiovascular Exam," but also subheadings within sections, such as "Cardiac Auscultation." In some instances, SecTag can also detect implied section headers, such as those for "Chief Complaint," using a modified Naïve Bayes algorithm combined with terminology-based rules. An evaluation of SecTag on 319 randomly select EMR H&P notes found 16,036 sections. Physician reviewers agreed with SecTag for 15,329 tags and identified 160 sections that were not recognized by the algorithm. The recall and precision of the SecTag algorithm were 99.0% and 95.6% for all sections, 98.6% and 96.2% for major sections, and 96.6% and 86.8% for unlabeled sections (71).

MedEx

MedEx is a Vanderbilt-developed NLP tool for extracting medications and medication-related information from natural language clinical notes. Developed by Xu, Denny, et al., MedEx extracts medication name, both generic and brand names, as well as other strength, dose, route, frequency, form, dose amount, intake time, duration, dispense amount, refill, and necessity, if they are present. Tested on discharge summaries, MedEx was shown to be very reliable in extracting not only drug names (F-measure 93.2%), but

also related information such as strength, route, and frequency, with F-measures of 94.5%, 93.9%, and 96.0% respectively (72). MedEx was later combined with other NLP tools, including SecTag, to participate in the 2009 i2b2 NLP challenge, placing second overall (73). It achieved an overall F-measure of 0.821 for exact matching with a precision of 0.839 and recall of 0.803 and an F-measure 0.822 for inexact matching (precision 0.866 and recall 0.782) (74).

MedEx has been used in studies to calculate the daily dose of drugs mentioned in clinical text. Specifically, MedEx was extended to normalize dose-related findings and calculate daily doses of the medication tacrolimus. Precision was greater than 0.90 and recall greater than 0.81 (75). MedEx was further extended to calculate weekly doses of the drug warfarin in another study. It determined patients' weekly doses with 99.7% recall, 90.8% precision, and 93.8% accuracy (76). MedEx is also extensively used at Vanderbilt to identify medications present in the Synthetic Derivative (SD) – the de-identified version of the Vanderbilt EMR used for research.

Vanderbilt Synthetic Derivative

The Synthetic Derivative (SD) is a comprehensive database containing nearly all clinical information present in the Vanderbilt "Star" EMR system (77). This includes laboratory values, billing codes, imaging and pathology reports, and clinical narratives from both the inpatient and outpatient setting. The SD is a de-identified resource; it is stripped of personal identifiers such as names, places, and addresses, dates are shifted by up to one year backward (consistent within each record but different across records), and medical record numbers are hashed to a new value consistent for each record. It is

updated nightly with new information from Star. The SD can be used as a stand-alone resource for clinical research, or as part of the BioVU program, in combination with genetic samples for phenome-genome analysis.

UMLS Co-Occurring Concepts (MRCOC)

The MRCOC contains aggregations of co-occurrences of concepts from multiple data sources, including MEDLINE (Medical Literature Analysis and Retrieval System Online), the AI/RHEUM Knowledge Base, and Canonical Clinical Problem Statement System (CCPSS) (78). AI/RHEUM contains co-occurrence of diseases and findings; CCPSS contains problem-problem co-occurrences extracted from patient records. Since we are attempting to classify drug-finding co-occurrences, we focused solely on the MEDLINE data.

The MEDLINE co-occurrence data in MRCOC was compiled from the MeSH (Medical Subject Headings) concepts designated as main topics of a given indexed journal article. Counts exist for the frequencies with which the first concept is qualified with MeSH qualifiers when it appears with the second concept. Separate counts and subheadings are provided for each direction of the relationship; that is, MRCOC gives the qualifying MeSH terms for concept one when it appears with concept two, as well as the qualifiers for concept two when it appears with concept one. MRCOC contains separate counts for recent MEDLINE entries, designated as MED and including entries from the 5 years prior to release (2003 – 2008), as well as entries from a preceding 5-year bloc, designated MBD (1998 - 2002) (79).

RxNorm

RxNorm is a standardized nomenclature of drugs and drug delivery devices developed by the National Library of Medicine (80). With RxNorm, medications are represented as "clinical drugs." Each clinical drug is defined by one or more ingredients, possible strengths, and dose forms. Since drug concepts are formed from these constituent pieces, clinical drugs can be mapped back to other doses or forms, and more importantly, generic ingredients. These mappings allow a user to link a brand name or dose-form drug with its generic ingredient concept(s) (81). RxNorm is included as part of the UMLS. We used the RxNorm version included in the UMLS2009AA release throughout this project.

National Drug File – Reference Terminology

The National Drug File – Reference Terminology (NDF-RT) is produced by the U.S. Department of Veterans Affairs Veterans Health Administration (VHA) since 2002 (82). It is an extension of the VHA National Drug File. The NDF-RT organizes the list of drugs into a formal representation used for modeling drug characteristics. Among other things, this includes ingredients, dose form, physiologic effect, mechanism of action, related diseases, and 25 distinct relationship types between concepts. Information exists for approximately 80,000 orderable compositions associated with 4000 active ingredients (83). Among other features, the NDF-RT contains 25 distinct relationship types between concepts (81; 84). Both RxNorm and NDF-RT contain mappings among drug concepts. These mappings allow a user to link a brand name or dose-form drugs with its generic ingredient concept (81).

The NDF-RT is updated regularly and new versions are released every six weeks. Each new release of the UMLS Metathesaurus contains the most recent version of the NDF-RT, as a part of RxNorm. We used version NDFRT_2008_03_11, the version contained in the UMLS 2009AA release used throughout this project.

Structured Product Labels

The Food and Drug Administration's Structured Product Labeling (SPL) is a XML (Extensible Markup Language) document standard for the labeling of human prescription drugs in the United States. Structured Product Labels contain information such as product names, generic names, ingredients, strengths, dosages, dose forms, and route of administration. In addition, they contain the full product labels from drugs sold in the United States (85). While the document is encoded using XML, the content of the sections is in unstructured natural language.

SIDER Side Effect Resource

The SIDER Side Effect Resource is a database that connects 925 drugs to 1450 side effect terms (86). The information contained in SIDER was extracted from the FDA SPLs using text-mining methods. The indication and adverse effects used by SIDER are from the COSTART (Coding Symbols for Thesaurus of Adverse Reaction Terms) (87) vocabulary, and are thus already in the UMLS with CUI representation. Drugs in the SIDER database are identified using a STITCH ID from PubChem (88). SIDER is available freely on the web and for download (89).

CHAPTER III

DRUG INDICATIONS AND ADVERSE EFFECTS EVIDENCE BASE

Overview

We constructed the Drug Indications and Adverse Effects Evidence Base (DEB, for Drug Evidence Base) to enable an automated system to identify potential explanations for why specific drug-finding pairs appear in clinical notes. To build the drug evidence base, we used information from the UMLS Co-Occurrence of Concepts (MRCOC), the National Drug File-Reference Terminology (NDF-RT), and the Food and Drug Administration's Structured Product Labels (SPL) for human prescription drugs. Based on data extracted from these resources, we algorithmically classified a drug-finding pair as an adverse effect (AE) – that is, the drug causes the finding – or an indication (IND) – the drug treats or prevents the finding.

The development of the DEB involved operational definition of a drug-finding pair; extraction of the relevant information from each of the knowledge sources; and reconciliation of this information into the evidence base. We compared the evidence base to the existing widely-used SIDER Side-Effect Resource to evaluate the DEB's comprehensiveness and accuracy.

Materials

This work was performed on a MacBook Pro with a 2.8 GHz Intel Core 2 Duo processor and 8 GB of RAM and a Linux server with eight 2.0 GHz Intel Xeon cores and

16 GB RAM. All data processing scripts were written in Perl. 5.10.0. The project used MySQL version 5.5.16.

Methods

Definition of Drug-Finding Pairs

Throughout this work, we defined a drug-finding pair as the co-occurrence of a particular Unified Medical Language System (UMLS) drug identifier (i.e., its name) and a particular UMLS clinical finding name that appeared together in a relevant clinical resource (e.g., SPL, NDF-RT, or a patient's clinical note). A drug refers to a single-ingredient medication. A finding can describe indications for drug therapy (e.g., a disease treated or prevented by the drug) or adverse effects of therapy (e.g., a physical examination finding, such as maculopapular rash). We designated all concepts using specific UMLS Concept Unique Identifiers (CUIs). We did not constrain concept origins to any specific UMLS source vocabularies. We used the UMLS 2009AA release because project tools, KMCI and MedEx, used that version.

We operationally constrained our definition of "drug" to include a UMLS concept that had at least one of the following UMLS semantic types:

- Antibiotic
- Pharmacologic Substance
- Clinical Drug

Based on review of several hundred de-identified patient charts by an experienced project clinician (RAM), we operationally constrained our definition of "finding" to UMLS concepts having at least one of the following semantic types:

- Anatomical Abnormality
- Injury or Poisoning
- Congenital Abnormality
- Finding
- Sign or Symptom
- Acquired Abnormality
- Clinical Attribute
- Disease or Syndrome
- Mental or Behavioral Dysfunction
- Neoplastic Process
- Pathologic Function

Extracting Pairs from UMLS Co-Occurring Concepts (MRCOC)

The "MRCOC component" of the DEB was derived from co-occurring concepts from articles indexed in MEDLINE (using both MED and MBD intervals, discussed above). To do so, we extracted from the UMLS MRCOC table those drug-finding pairs that potentially denoted an AE or IND. Each entry in MRCOC consists of two concepts, their source information, the number of co-occurrences in the literature during a specified time period, and the MeSH subheadings qualifying the relationship between the two

concepts as they appeared in the literature. The MRCOC table contains separate entries for each direction of the relationship between two distinct concepts (see Figure 2).

Procedurally, we extracted all MRCOC entries in which the first concept (cui1 in Figure 2) met our definition as a drug and the second concept (cui2 in Figure 2) qualified as a finding (per definitions above), or vice versa. Next, we combined entries from the UMLS time intervals MED and MBD entries (mentioned above, shown below in "source" column in Figure 2) and summed the co-occurrence counts ("count" column in Figure 2) and the individual subheading counts (rightmost column in Figure 2). We only retained subheading information for the following relevant MeSH subheadings: AE – Adverse Effect; DT – Drug Therapy; ET – Etiology; and TU – Therapeutic Use. Lastly, to ensure well-established relationships for a given pair, we discarded drug-finding pairs with an overall co-occurrence count less than 4.

As illustrated in Figure 2, MeSH subheadings in the rightmost column of the MRCOC table only apply to the first of the paired two concepts when they appear together. We utilized the MRCOC subheading information to infer whether the relationship between drug and finding was most likely IND or AE. When the subheading TU qualifies the drug concept of the drug-finding pair, this suggests the finding is an IND for the "therapeutic use" of the drug. When the subheading DT qualifies the finding concept of the drug-finding pair, this further supports that the finding was an IND for the "drug therapy". When the subheading AE ("adverse effects") qualifies the drug concept of the drug-finding pair, this directly suggests that the finding was an AE. Similarly, when the subheading ET qualifies the finding concept of the drug-finding pair, this indicates that administration of the drug may have played a role in the etiology of the

finding (i.e., the finding was an AE). Thus, the combination of the subheadings is important; drug/TU+finding and drug+finding/DT implies an indication and drug/AE+finding and drug+finding/ET implies an adverse effect. Nevertheless, the actual indexing of articles in the literature (which may describe concurrent use of multiple drugs for multiple conditions with multiple possible interrelationships) can produce conflicting information. We describe how we addressed such conflicts below.

We stored the "MRCOC component" of the DEB in a temporary MySQL table before combining this information with the other DEB sources (see below). For each MRCOC drug-finding pair, the temporary table contained two entries, one for each direction of the relationship, along with subheadings and co-occurrence counts, as described above and illustrated in Figure 2.

UMLS MRCOC Table

cui1	cui2	source	count	mesh subheadings of cui2 when it appears with cui1
C0043031	C0038454	MBD	91	TU=74,AD=22,AE=20,CT=6,EC=3...
C0043031	C0038454	MED	121	TU=89,AD=41,AE=29,CT=4,EC=4...
...				
C0038454	C0043031	MBD	91	PC=78,ET=42,EP=10,DT=9,MO=6...
C0038454	C0043031	MED	121	PC=100,ET=55,DT=21,EP=20,MO=6...

Combining sources and discarding extra subheadings...

C0043031	C0038454	MED+MBD	212	TU=163,AE=49
C0038454	C0043031	MED+MBD	212	ET=97,DT=30

Figure 2. Sample MRCOC data for Warfarin (C0043031) and Stroke (C0038454)

Extracting Pairs from National Drug File Reference Terminology (NDF-RT)

To create the "NDF-RT component" of the DEB, we extracted drug-finding pairs from the NDF-RT information included in the UMLS MRREL table. As indicated in

Figure 3, we extracted all MRREL entries from NDF-RT that contained a drug concept and a finding concept that also had one of the following NDF-RT relationships: "has physiologic effect" and "induces" (indicating a potential AE), and "may prevent" and "may treat" (indicating a likely IND). The final form of the data extracted from NDF-RT included, for each entry, a drug CUI, a finding CUI, and all of the relevant relationships that appear between the two CUIs.

drugCUI	findingCUI	relationship(s)
C2267047	C0035235	may_prevent,may_treat
C1096766	C0041657	induces
C0913469	C1371635	has_physiologic_effect
C2587204	C0040136	may_treat
...		

Figure 3. Sample rows from the temporary table for the NDF-RT component.

Extracting Pairs from FDA Structured Product Labels (SPLs)

To create the "SPL component" of the DEB, we extracted drug-finding pairs using NLP of the FDA's product labels for prescription drugs.

We downloaded all available human prescription drug product labels from DailyMed (54), as well as the SPL Downloadable Data Elements file (90). This file acts as in index for the SPLs, containing information such as National Drug Code (NDC), proprietary name, ingredient(s), product type, marketing category, and a link to the appropriate SPL files. Based on NDC, this contained information on 56,854 drugs associated with approximately 2400 unique sets of ingredients, often containing more than one active ingredient.

Using the index, we selected all ingredient and proprietary names for drugs containing a single active ingredient and with a "product type" indicating the name applied to a Human Prescription Drug. We then automated a process whereby these ingredients were mapped to CUIs by exact match with UMLS strings of drug concepts available in the MRCONSO table of the UMLS Metathesaurus. We manually reviewed these matches to confirm accuracy and correct any mismatches or unmatched ingredients where possible.

For every drug that we were able to match to a CUI, we then parsed the respective SPL. Using the XML structure, we extracted the Adverse Reactions Section (section 34084-4) and the Indications and Usage Section (section 34067-9) when they were present (91). We then used the Knowledge Map Concept Identifier (KMCI) to extract all of the finding concepts in each section for each label.

Since the SPLs contain entries for every drug marketed in the US, there are many duplicates. For example, there is a label for every package quantity, dose form, and brand name of acetaminophen. Therefore, a single CUI often mapped to many different SPLs.

For each SPL, we extracted a set of all unique finding concepts identified in the AE section, another of all unique finding concepts in the IND section. This data was then transformed into a table of drug-finding pairs by linking the drug CUI with the CUI of every distinct concept identified in the Adverse Reactions Section (classified as an AE) and the CUI of every distinct concept identified in the Indications and Usage Section (classified as an IND).

The final form of the data we extracted from the SPLs includes, for each entry, a drug CUI, a finding CUI, whether or not this was mentioned as an AE, and whether or

not this was mentioned as an IND (see Figure 4). It is possible for a concept to appear as both an AE and an IND. We discuss this issue in the following section on construction of the final combined evidence base.

drugCUI	findingCUI	AE	IND
C0059985	C1514463	AE	
C0059985	C1517205	AE	
C0060135	C0002874	AE	IND
C0248719	C0262926		IND
C0249529	C0018418	AE	
...			

Figure 4. Sample rows from the temporary table for the SPL component.

Mapping to Uniform Drug Concepts

A wide variety of medication-related identifiers from the three DEB knowledge sources met our UMLS-semantic-type-based definition of a drug. For instance, C0000970 – Acetaminophen, C0699142 – Tylenol, and C1640784 – Tylenol 160 mg were all unique UMLS CUIs that appeared in the DEB knowledge sources, and might also appear in patient notes. While these CUIs all represent the same drug ingredient, they all correspond to different UMLS concepts. The first is a generic drug, the second is a brand name drug, and the third is a dispensable form of the drug that includes a specific dose. The DEB requires, whenever possible, that a single identifier/name be assigned to all single-component drugs that involve the same generic ingredient. To combine and condense the DEB drug names from multiple sources, we mapped each identified drug CUI to a drug generic ingredient using the relationships in MRREL, predominantly using mappings from RxNorm. In particular, we used the relationships "ingredient of" and "has ingredient" to map dose forms of a drug to the drug ingredient only (C1640784 – Tylenol

160 MG to C0699142 – Tylenol). We also used the "has tradename" and "tradename of" relationships to map a brand name drug back to its generic ingredient name (C0699142 – Tylenol to C0000970 – Acetaminophen). We mapped all specific instances of each drug term to its generic ingredient, whenever mappings existed.

Integrating Source Information into Combined Evidence Base

We combined the data extracted from MRCOC, NDF-RT, and SPL components of the DEB into a single DEB table. This table contained a single entry for each generic-drug--finding pair. To combine data regarding whether the pair comprised an IND or an AE, we developed a scoring system that assigned weights to the information from the three disparate sources. We created scores indicating the level of support for a pair being an AE versus an IND (the INDscore and the AEscore). We ranked the three component sources based on our own interpretation of their authoritativeness to determine the maximum number of points each could contribute to the overall score.

We rated the NDF-RT highest because it is a manually curated, trusted knowledge source and because the NDF-RT relationships that we used state directly whether a drug-finding pairing represents an IND or an AE. The NDF-RT component contributed 10 points to the AEscore if a relationship indicating an AE was present, and 10 points to the INDscore if a relationship indicating an IND was present. If NDF-RT indicated that both an AE and IND relationship were present, we added 10 points to each score.

We rated the MRCOC component as second most authoritative because it was based on NLM indexers reviewing articles published in the peer-reviewed literature. It could contribute a maximum of 10 points, but the exact score was based upon the fraction

of co-occurrences that indicated either AE or IND. More specifically, if there were X subheading counts that supported an AE determination, Y subheading counts that supported an IND determination, and N total co-occurrences, (X/N)*10 points were added to the AEscore and (Y/N)*10 points were added to the INDscore.

We gave the SPL component the lowest weighted proportion of the overall score, not because of the authoritativeness of SPL per se, but because extraction of information via NLP was deemed to be potentially unreliable. This process involved the least amount of manual review – concepts were extracted automatically using NLP, unlike the structured expert-derived NDF-RT or the structured manual coding of MeSH terms. The SPL component contributed a maximum of 5 points to the scores. If a drug-finding pair was designated only as an AE or an IND in SPL, 5 points was added to appropriate DEB score. If, however, there was contradictory evidence – that is, a finding appeared in both the Indications section and Adverse Effects section of the SPL, we added 2 points to AEscore and 4 points to INDscore. We assigned the slight advantage to IND because we observed that mentions of indications often appeared in the text of the Adverse Effects sections of the SPL, as in "When treating for the indication XYZ, adverse effects might include…". Our NLP processing typically would label XYZ as an IND and an AE in such circumstances. Therefore, if a concept appeared in both SPL sections, we assigned a higher value to IND.

Finally, we algorithmically compared the overall AEscore and INDscore as determined by the combined three sources, and used the larger of the two scores to determine the final DEB classification of a drug-finding pair. In the case of a tie, we assigned IND because we observed that when there was evidence for a drug being both

an IND and an AE, it was often when a treatment when withdrawn, might exacerbate the treated condition, such as when withdrawal of clonidine exacerbates the hypertension it was previously treating. The DEB only classifies a drug-finding pair as AE or IND; there is no "both." A large difference between the AE score and IND score implies higher certainly, whereas very similar scores imply less certainty.

drugCUI	findingCUI	Sources	INDscore	AEscore	Determination
C0020740	C0029408	mrcoc,ndfrt,spl	25.00	5.00	IND
C0020740	C0029882	mrcoc	5.00	5.00	IND
C0020740	C0038358	mrcoc	4.29	5.71	AE
C0020740	C0038454	mrcoc,spl	3.33	11.67	AE
C0020740	C0021368	mrcoc,ndfrt	18.00	0.00	IND
...					

Figure 5. Sample rows from the combined DEB for *Ibuprofen (C0020740)* and *Degenerative Polyarthritis (C0029408), Otitis Media (C0029882), Gastric Ulcer (C0038358), Cerebrovascular Accident (C0038454),* and *Inflammation (C0021368),* respectively.

Evaluation of DEB: Comparison with Expert Opinion and SIDER as "Gold Standard"

SIDER is an open-source database of medication indications and adverse drug effects (described in more detail above) (86). We did not integrate SIDER content into DEB, but instead used it as a "gold standard" external reference for evaluation as described below.

We downloaded the SIDER Side Effect Resource files "adverse_effects_raw", "indications_raw", and "label_mapping" from the SIDER website (89). The adverse_effects_raw and indications_raw files contained AE and IND concepts, respectively, and the drug label ID from which they were extracted. The label_mapping files contained the label IDs, as well as the brand name, generic name, and PUBCHEM

STITCH ID for the drug. The data was in tab separated value format that we loaded into a MySQL database.

To use SIDER for evaluation purposes, we mapped SIDER concepts to UMLS CUIs so we could directly compare the drug-finding pairs in SIDER with those in the DEB. Finding concepts in SIDER were already designated by CUI, but drug names in the SIDER database were identified using STITCH ID, brand name, and generic name. We used simple string matching to compare the generic names (or the brand name if there was no match with the generic) with UMLS strings from the MRCONSO file. When there was a match, we were able to link the STITCH ID to the appropriate UMLS CUI for the drug concept. We manually reviewed these matches to ensure correct matching whenever possible.

The final format of the extracted SIDER information included a drug CUI, a finding CUI, and a field indicating whether the finding was an adverse effect or an indication for that drug. It is possible for SIDER to classify a drug-finding pair as both an AE and an IND; in this case, we considered the SIDER pair an IND, as was done for the DEB.

Expert Reviewer Evaluation of DEB

To evaluate the accuracy and potential utility of DEB, we used a two-step approach. First, we compared DEB to SIDER, a commonly used resource to identify drug adverse effects and indications. We first identified which drug-finding pairs were present in both the DEB and SIDER or absent in one or the other. For drug-finding pairs

present in both SIDER and the DEB, we determined if both resources categorized the relationship in the same manner as an IND or AE, or if they differed.

Four Vanderbilt faculty physician reviewers, each board-certified in internal medicine and with over 10 years of clinical experience, each rated 200 drug-finding pairs from the DEB. Reviewers were blinded to the DEB and SIDER categorizations of the pairs. We provided to each reviewer a Microsoft Excel spreadsheet containing their unique set 200 drug-finding pairs and, for each pair, links that enabled them, with one click, to search either PubMed or Google for information regarding the pair. The experts received instructions to mark the relationship of the pair as either AE, IND, Both (see explanation below), or Neither. They also could fill out an optional field for comments about the relationship or the information they found. Of the 200 pairs given to each reviewer, 25 were the same across all reviewers (to calculate inter-rater agreement), 75 were randomly chosen from drug-finding pairs where the DEB categorization differed from SIDER, and 100 were chosen at random from the overall DEB segment that did not overlap with SIDER (i.e., SIDER did not contain the pairing). To find inter-rater agreement, we calculated both Cohen's Kappa for each pairwise combination of reviewers and Fleiss' Kappa for the entire group.

Due to the nature of indications and adverse effects in clinical practice, a finding concept can represent both an AE and an IND. As previously discussed, this was often the case in the MEDLINE classifications in MRCOC. For example, warfarin might be used for stroke prophylaxis in a patient with atrial fibrillation, but when given in excessive dosages, warfarin can cause a stroke through intra-cerebral hemorrhage. Even though both DEB and SIDER classify a drug-finding pair as one or the other, reviewers

were allowed to classify pairs as “both” or “neither”. The “neither” category implied that the pair was an incorrect association or, in some cases, one that was too broad or nonsensical.

We compiled and compared the expert reviewers’ comments to identify any common themes regarding problems with DEB. Additionally, one reviewer empirically analyzed some of the drug-finding pairs that occurred in DEB alone, SIDER alone, and both DEB and SIDER to assess the similarities and differences among the categorizations.

Results

MRCOC Component of DEB

From the MRCOC table, our DEB construction algorithms extracted 423,776 entries that contained paired drug and finding concepts matching our semantic type criteria. These entries represented approximately 100,000 drug-finding pairs since there were typically four UMLS entries for each pair: an entry in each of two directions from MBD and from MED sources). After combining entries into drug-finding pairs, generalizing and combining drugs concepts, and rejecting pairs that did not have our required MeSH subheadings, the algorithms retained 65,930 distinct pairs consisting of 1825 unique drugs and 3121 unique findings.

NDF-RT Component of DEB

From the NDF-RT, the DEB construction algorithms extracted 51,132 entries in which one of the relevant relationships was present between a drug concept and a finding concept. These entries represented ~25,500 drug-finding pairs (one entry for each direction of the relationship). After generalizing and combining drug concepts, the algorithms retained 7870 pairs consisting of 2084 unique drugs and 1033 unique findings. This included 273 "has physiological effect" relationships, 622 "induces" relationships, 5483 "may prevent" relationships, and 44,934 "may treat" relationships.

SPL Component of DEB

From the SPLs, the DEB construction algorithms identified 958 single active ingredient drugs. Using UMLS strings to map these drugs to CUIs, the algorithms were able to match 888 drugs. After generalizing and combining drug concepts, the algorithms retained 758 distinct drugs. Overall, the algorithms extracted 6980 unique finding concepts from the SPLs as both AEs and INDs. This resulted in 81,223 distinct drug-finding pairs.

Combined DEB

After combining information from the three DEB data sources, the resulting DEB evidence base contained 137,194 drug-finding pairs consisting of 3242 unique drugs and 8266 unique findings. There were 132,629 pairs (97%) with data from only one source, 4086 pairs (~3%) where data came from two knowledge sources, and 479 pairs (< 1%)

where data came from all three sources. Overall, the DEB classified 79,284 pairs as AEs and 57,854 pairs as INDs.

SIDER

After mapping SIDER information to UMLS CUIs and generalizing and combining drug concepts (in the manner done for the DEB), we retained 63,857 SIDER drug-finding pairs consisting of 871 unique drugs and 1688 unique findings. This included 58,024 pairs as AEs and 5833 pairs as INDs.

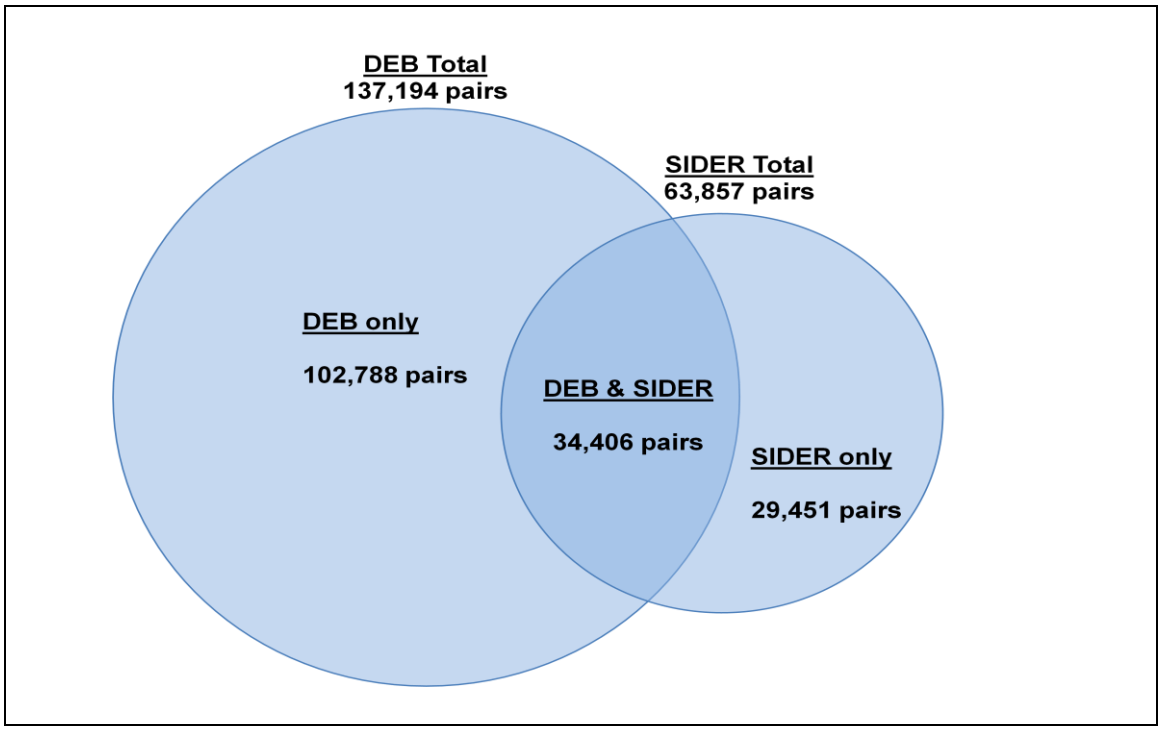

Figure 6. Venn diagram illustrating drug-finding pairs in DEB and SIDER.

SIDER Evaluation

In comparing the DEB with the information in SIDER, we determined that 34,406 pairs were in both SIDER and DEB (see Figure 6). This represented 54% of extracted SIDER findings and 25% of DEB. The other 29,451 pairs in SIDER did not appear in the DEB (see below). Of the pairs in both sources, SIDER and the DEB agreed on 33,398 (97%) categorizations (as IND or AE) and disagreed on 1008 (3%).

Expert Reviewer Evaluations

Of the 200 drug-finding pairs reviewed by each reviewer, 25 were the same for each reviewer. Using their ratings of the common drug-finding pairs, we calculated inter-rater agreements. We used Cohen's Kappa for each pairwise combination between each of the four reviewers, and we used Fleiss' Kappa to measure overall agreement. Since determining whether a finding is an AE or IND can be subjective, we also calculated Kappa separately for those drug-finding pairs on which no reviewer had indicated a "both" relationship, in order to estimate the degree of agreement on the less ambiguous pairs. Results are shown in Table 1 below.

Table 1. Agreement between Reviewers.

	N=25 (all)			**N=21 ("Both" removed)**		
Reviewers	**Kappa**	**p-value**	**95% CI**	**Kappa**	**p-value**	**95% CI**
1 and 2	0.11	0.374	(-0.13, 0.36)	0.17	0.312	(-0.16, 0.49)
1 and 3	0.58	<0.001	(0.28, 0.87)	0.59	<0.001	(0.27, 0.90)
1 and 4	0.43	<0.001	(0.18, 0.68)	0.59	<0.001	(0.27, 0.90)
2 and 3	0.13	0.291	(-0.11, 0.38)	0.18	0.273	(-0.14, 0.49)
2 and 4	0.47	<0.001	(0.23, 0.71)	0.34	0.034	(0.03, 0.66)
3 and 4	0.50	<0.001	(0.25, 0.75)	0.67	<0.001	(0.36, 0.99)
1, 2, 3, & 4	0.36	<0.001	(0.26, 0.47)	0.42	<0.001	(0.29, 0.55)
1, 3, & 4	0.50	<0.001	(0.34, 0.65)	0.62	<0.001	(0.43, 0.80)

Of the 200 drug-finding pairs reviewed by each of the four reviewers, 75 were from a random sample of those pairs where the DEB categorization of IND vs. AE disagreed with that of SIDER (this sample included 300 of the total of 1008 overall DEB-SIDER classification disagreements). Table 2 below shows ratings on these 300 discrepant pairs by our reviewers. All the four reviewers agreed with DEB more often than SIDER ($P \leq 0.01$ for the group as a whole). On average, reviewers agreed with DEB 30% more of the time when the disagreement occurred between DEB and SIDER (95% CI was 20% to 40%).

Table 2. Reviews on Disagreements between DEB and SIDER.

Rev	Agreed w/ DEB (Pr_1)	Agreed w/ SIDER (Pr_2)	Both (Pr_3)	Neither (Pr_4)	$Pr_1 - Pr_2$		$H_0 : Pr_1 = Pr_2$
					Est.	95% CI	P-value
1	0.64 (48/75)	0.28 (21/75)	0.03 (2/75)	0.05 (4/75)	0.36	(0.16, 0.56)	<0.001
2	0.55 (41/75)	0.24 (18/75)	0.09 (7/75)	0.12 (9/75)	0.31	(0.12, 0.50)	0.001
3	0.53 (40/75	0.28 (21/75)	0.01 (1/75)	0.17 (13/75)	0.25	(0.06, 0.45)	0.011
4	0.52 (39/75)	0.25 (19/75)	0.07 (5/75)	0.16 (12/75)	0.27	(0.08, 0.46)	0.006
Avg	0.56	0.26	0.05	0.13	0.30	(0.20, 0.40)	<0.001

Of the 200 drug-finding pairs each expert reviewed, 100 were from a random sample from DEB that did not overlap with SIDER. Results of categorizations for each reviewer are shown in the below in Tables 3 and 4. All the reviewers significantly agreed with DEB on more than 42% of the pairs and disagreed on less than 25% of the pairs. On average, reviewers were 9-fold more likely to agree with the DEB categorization than to disagree with it (95% CI was 5.6 to 20.9 fold). Note that DEB did not include "both" or "neither" options in its internal categorizations.

Table 3. Reviewers' categorizations of random selection from DEB/SIDER overlap.

Reviewer	Agreed w/ DEB (Pr_1)	Disagreed w/ DEB (Pr_2)	Both (Pr_3)	Neither (Pr_4)
1	0.78 (78/100)	0.16 (16/100)	0.03 (3/100)	0.03 (3/100)
2	0.52 (52/100)	0.11 (11/100)	0.08 (8/100)	0.29 (29/100)
3	0.58 (58/100)	0.08 (8/100)	0.00 (0/100)	0.34 (34/100)
4	0.57 (57/100)	0.03 (3/100)	0.02 (2/100)	0.38 (38/100)
On avg.	0.61	0.10	.0.3	0.26

Table 4. Reviewers' agreements (continued from Table 3).

Reviewer	Pr_1		Pr_2		Pr_1 / Pr_2	
	Estimate	**95% CI**	**Estimate**	**95% CI**	**Estimate**	**95% CI**
1	0.78	(0.70, 0.86)	0.16	(0.09, 0.24)	4.88	(3.00, 9.33)
2	0.52	(0.42, 0.62)	0.11	(0.05, 0.17)	4.73	(2.71, 10.60)
3	0.58	(0.48, 0.67)	0.08	(0.03, 0.14)	7.25	(3.79, 20.36)
4	0.57	(0.48, 0.66)	0.03	(0.00, 0.07)	19.00	(7.86, 63.00)
On avg.	0.61	(0.57, 0.66)	0.10	(0.07, 0.12)	8.96	(5.60, 20.86)

Table 5 shows an expert clinician's categorizations of the DEB and SIDER drug-finding pair classifications for one drug, abacavir. The reviewer, in the comments field, noted how often apparent differences might be due to use of differing terms for similar findings.

Table 5. DEB/SIDER classifications for the drug abacavir.

Finding	SIDER	DEB	Comments	Discrepancy Count
Musculoskeletal pain	AE	AE	See "myalgia" below	
Abdominal Pain	AE	AE		
Acidosis	AE		See below	1
Acidosis, Lactic		AE	See above	2
Acquired Immunodeficiency Syndrome	IND	IND		
Adverse event		AE	Bad category	3
Alanine aminotransferase increased		AE	See "Liver function …" and "Increased liver function …" below	4
Allergy Severity - Severe		AE	See below	5
anaphylaxis	AE		See above	6

Table 5 (continued). DEB/SIDER classifications for the drug abacavir.

Anemia	AE	AE		
Anorexia	AE			7
Anxiety	AE			8
Arthralgia	AE		See "Musculoskeletal Pain" above	9
Blind Vision		AE		10
Bronchitis	AE	AE		
Chills	AE	AE		
Conjunctivitis	AE			11
Coughing	AE		See "Bronchitis" above	12
Creatine phosphokinase increased		AE		13
Depressive disorder		AE		14
Diarrhea	AE	AE	See also "Severe Diarrhea" below	
Dizziness	AE	AE	bad category	
Dream disorder		AE	See "Sleep Disorders" below	15
Dyspnea	AE		See "Shortness of breath" below	16
Edema	AE			17
Enlargement of lymph nodes	AE			18
Erythema Multiforme	AE	AE		
Exanthema	AE	AE	See multiple skin disorders listed elsewhere	
Fatigue	AE	AE	bad category	
Fatty Liver	AE	AE		
Fever	AE	AE		
Gastritis	AE	AE	See below	
Gastroenteritis	AE		See above	19
gastrointestinal sign		AE	See above	20
Gastrointestinal symptoms NOS		AE	See above	21
Headache	AE	AE		
HIV Infections	AE	IND	MAJOR error in SIDER; See "Acquired Immunodeficiency …" above	
Hyperamylasemia	AE	AE	See pancreatitis below	
Hyperglycemia	AE	AE		
Hypersensitivity	IND	IND	Looks like ERROR in both unless better explained	
Hypertriglyceridemia	AE	AE		
Hypotension	AE		?? Part of "Anaphylaxis" above ?? Or independent ??	22
Infection	IND		bad category	23
Infective pharyngitis		AE	See "pharyngitis" below	24
Influenza	AE		See "viral respiratory infection" below	25
Kidney Failure	AE			26
Leukopenia	AE	AE	See low individual WBC type descriptors also	

Table 5 (continued). DEB/SIDER classifications for the drug abacavir.

Lipid Metabolism Disorders	AE		See "Hypertriglyceridemia" above	27
Liver Failure	AE		See below	28
Liver function tests abnormal finding		AE	See above AND below "Raised liver …"	29
Lymphopenia	AE		See "Leukopenia" above	30
Malaise	AE	AE	bad category	
Migraine Disorders	AE	AE		
Morular Metaplasia of the Endometrium		AE	?? Bad category ??	31
Myalgia	AE		See "Musculoskeletal Pain" above	32
Myocardial Infarction		AE		33
Nasal infection		AE		34
Nausea	AE	AE		
Neutropenia	AE	AE	See "Leukopenia" above	
Oral Ulcer	AE			35
Pain	AE	AE	bad category	
Pancreatitis	AE	AE		
Paresthesia	AE			36
Pharyngitis	AE		See "infective pharyngitis" above	37
Pneumonia	AE	AE		
Raised liver function tests	AE		See "Liver function …" and "Alanine Amino…" above	38
Respiratory Distress Syndrome, Adult	AE		See below	39
Respiratory Failure	AE		See above	40
Severe diarrhea		AE	See also "Diarrhea" above	41
Shortness of Breath	AE		See "Dyspnea" above	42
Sleep Disorders	AE	AE	See below	
Sleeplessness	AE		See above	43
Sore Throat	AE		See "Pharyngitis" above and "Infective Pharyngitis" above	44
Spondylolisthesis, grade 2		AE	?? Bad category ??	45
Stevens-Johnson Syndrome	AE	AE		
Therapy naive		AE	bad category	46
Thrombocytopenia	AE	AE		
Toxic Epidermal Necrolysis	AE	AE		
Urticaria	AE		See "Allergy" above	47
Viral respiratory infection		AE	See "Influenza" above	48
Vomiting	AE	AE		
White blood cell count increased		AE		49

Of the 49 apparent discrepancies (enumerated in Table 5) where only DEB or only SIDER listed a finding associated with the drug abacavir, the expert marked four as "bad category" errors in for finding concepts. Of the remaining 45 discrepancies, the expert identified 16 where collapse of findings into homogenous terminology would eliminate discrepancies. These 16 collapsible discrepancies plus 4 bad categories represent 44% of apparent discrepancies.

Table 6, below, shows a sample of reviewer comments. Note that most reviewers only commented on those drug-finding pairs that were more ambiguous. Matching determinations between DEB and the reviewer are in bold.

Table 6. Sample of reviewer comments (including Reviewer and DEB determinations.)

Drug	Finding	Review	DEB	Comments
Acetylcysteine	Heart Diseases	**IND**	**IND**	"heart Diseases" too vague to use. N-Acetlycysteine used to prevent damage due to myocardial ischemia, mostly in research
almotriptan	Nausea	Both	IND	Either both, or side effect only . Can cause nausea, and indicated for migraine which has nausea as a symptom often
Anti-Bacterial Agents	Theileriasis	Neither	IND	"anti-bacterial agents" too broad; disease only affects CATTLE
Anticoagulants	Compartment syndromes	**AE**	**AE**	a stretch, sort of
Anticonvulsants	Ketogenic Diet	Neither	IND	both are tx for seizures
Antioxidants	Pathologic Neovascularization	**IND**	**IND**	"Antioxidants" too general, and pathologic neovasc is not much better;
Bupropion	Weight Gain	**IND**	**IND**	Indirect association, helps with smoking but prevents wt gain experienced during smok. Cessation
Cardiovascular Agents	Atrial Fibrillation	Both	IND	"CV agents" category too general; mostly treat; digoxin does both
Cardiovascular Agents	Atrial Fibrillation	**IND**	**IND**	med too general
Chlormethiazole	Alcoholic Intoxication, Chronic	AE	IND	Fatal in alcoholics
Cisplatin	Horse Diseases	Neither	IND	"horse diseases" not a relevant term -- drop it.
Corticotropin	Contracture	Neither	AE	Corticotropin is a natural substance in humans -- its deficiency can lead to flexion contractures

Table 6 (continued). Sample of reviewer comments.

Dantrolene	Tachypnea	Neither	IND	Dantrolene treats malignant hyperthermia, itself a very rare cause of tachypnea
Dantrolene	Tachypnea	Neither	IND	Doesn't treat tachypnea, rather malig hyerthermia which presents with tachypnea
Dantrolene	Tachypnea	**IND**	**IND**	for MH
Dextrothyroxine Sodium	Coronary Arteriosclerosis	Neither	IND	Ancient form of lipid-lowering therapy no longer used, especially for CAD
Estrogens	Cerebrovascular accident	AE	IND	Very weak association in the literature.
Excitatory Amino Acid Antagonists	Tobacco Use Disorder	**IND**	**IND**	Mostly research studies; terms both too vague
Goserelin	Neoplasms	IND	AE	"Neoplasms" too broad-used in treating breast/prostate cancers
Haloperidol	Vomiting	Neither	IND	could be AE but rare
Heparin	Asthma	**IND**	**IND**	Weak association
Hydrocortisone	Erythema	Neither	AE	Could say this is an indication, or an effect, but I think neither is best
Iloprost	Personal Satisfaction	Neither	AE	"Personal satisfaction" is a "junk" term in UMLS for our purposes
imiquimod	Carcinoma	**IND**	**IND**	"Carcinoma" too general - this is a topical agent used in various forms of skin cancer
Lactulose	Diarrhea	AE	IND	Both desired and adverse effect
Levalbuterol	Adverse event	Neither	AE	Adverse Event is too nonspecific, ignore this term
Lidocaine Hydrochloride	Drug toxicity	Neither	IND	too broad
Metformin	Hepatitis	AE	IND	rare
Nicotine	Pain	Both	IND	Too complex -- "nicotinic receptors" involved in neural pain pathways, smoking interacts with pain, etc.
Nitric Oxide	Lung diseases	**IND**	**IND**	"Lung diseases" too nonspecific -- it only treats pulmonary hypertension and rare other disorders
Omeprazole	Vomiting	IND	AE	If GERD/PUD is causing vomiting
Phenylalanine	Cognition Disorders	IND	AE	unsure
Praziquantel	Sheep Diseases	Neither	IND	huh?
Psychotropic Drugs	Substance-Related Disorders	IND	AE	"Substance-related disorders" and psychotropic drugs -- both too broad as categories
Raloxifene	Breast tenderness	Neither	AE	Breast tenderness only occurs when this drug is combined with estrogen therapy
repaglinide	CARDIAC EVENT	**AE**	**AE**	"CARDIAC EVENT" too nonspecific of a term
Zinc	Skin Neoplasms	Neither	IND	unsure
Zinc	Skin Neoplasms	**IND**	**IND**	Zinc oxide sunblock creams used in preventing later skin cancer

Discussion & Limitations

Summary

The DEB automatically combines drug indication and adverse effect information from multiple sources. These sources are frequently updated, and since DEB is constructed algorithmically without manual intervention, it can be regenerated in a fully automated manner to take advantage of updates in the knowledge sources. We have shown that the DEB is comparable in several ways to SIDER, a popular resource for drug-finding information, for the drug-finding pairs that they have in common. Additionally, our expert review suggests that the DEB may be more accurate when SIDER and DEB disagree on IND/AE determination. While there are disparities (e.g., use of different synonyms to denote the same findings) among many of the UMLS concepts in DEB and SIDER, there appears to be general agreement on broad concepts. It appears that future work on condensing clinical synonyms into a canonical list of concepts should be able to address these disparities. We discuss our specific results below.

Direct Comparison with SIDER

For the drug-finding pairs present in both DEB and SIDER (34,406), there was 97% agreement (33,398) on whether the pair represented an AE or IND. These results suggest that DEB is comparable to SIDER for the drug-finding pairs they have in common. We cannot draw conclusions regarding the accuracy or reliability of the

remaining portions of DEB from these results alone. We discuss experts' judgments regarding the disagreements below.

DEB Evaluation: Inter-rater reliability

Landis and Koch (92) gave one of the most popularly used interpretations of Kappa, however it is not universally accepted (93). They reported that a Kappa less than or equal to zero indicates poor agreement; a Kappa of 0.01-0.20, slight agreement; 0.21-0.40, fair agreement; 0.41-0.60, moderate agreement; 0.61-0.80, substantial agreement; and almost perfect agreement at 0.81-1.00.

For the 25 pairs that the reviewers rated in common, the pairwise Kappas between reviewers range from slight agreement to moderate agreement, with values implying moderate agreement. However, if one does not consider the pairs that reviewers rated as both – that is, removing those pairs that were ambiguous – some of the pairwise Kappas move into the range of substantial agreement. Similarly, for Fleiss' Kappa over all reviewers, agreement ranges from fair (Kappa=0.36, $p < 0.001$) to moderate (Kappa=0.42, $p < 0.001$). We believe that this illustrates a general agreement between reviewers, but further illustrates the difficulty of determining the nature of a drug-finding pair relationship. We believe that further research might take advantage of the information used in constructing the DEB to examine the differences in the INDscore and AEscore for each drug pair, along with the number of sources from which the DEB derived a drug-finding pair relationship, to incorporate a certainty metric into DEB.

DEB Evaluation: Expert Review of DEB/SIDER Disagreements

Our expert reviewers examined a random sample of 300 pairs (75 for each reviewer) taken from the overall set of 1008 drug-finding pairs where the DEB and SIDER categorizations (IND vs. AE) disagreed. While the reviewers agreed with DEB between 53% and 72% of the time, they only agreed with SIDER between 24% and 28% of the time. On average, reviewers agreed with DEB 30% more of the time when DEB and SIDER disagreed ($P \leq 0.01$). We believe this suggests that a drug-finding database compiled from multiple sources of varying reliability might provide better information than using only a single, potentially unreliable source whenever information – especially when data regarding a specific effect are ambiguous. Additionally, improved mapping of similar finding concepts to canonical "consensus" terms (discussed below) could potentially resolve many discrepancies within the DEB and within SIDER, as well as between the two knowledge sources.

Expert Review of Random Selection from DEB

The expert reviewers also examined a random sample of 400 pairs (100 for each reviewer) from DEB that were not in SIDER. On average, they agreed with the DEB categorizations of the drug-finding pairs 61% of the time. All the reviewers significantly agreed with DEB on more than 42% of the pairs and only disagreed with DEB on 25% of the pairs. Reviewers only selected the opposite determination (AE when DEB had IND, or vice versa) 10% of the time on average. While reviewers agreed with DEB a majority of the time (they were 9-fold more likely to agree with DEB than disagree), nearly a

quarter of the pairs were marked as "neither," often due to problems with vague terminology.

DEB/SIDER Content Comparison

As illustrated in Table 4, there are many specific CUIs that appear only in SIDER or DEB, but many general concepts are addressed by both sources. Sometimes the two sources use similar but different CUIs for the same findings, and neither uses the CUI the other used. Sometimes one source lists a specific concept and the other a more general concept. Occasionally, one source will list a high-level disease process and the other will only list the disease process and findings of that disease, or sometimes only findings. Similar to how drug-concepts were mapped to a single generic ingredient, a method to combine similar finding concepts would be beneficial. Table 5 suggests that up to 44% of apparent discrepancies in whether a finding was listed in SIDER but not in DEB or vice-versa might be due to using different synonyms for the same concept, or one using a parent concept and the other using a child concept. Future research should attempt to combine finding concepts in a manner analogous to the way in which we combined different drug concepts into a "single ingredient generic drug concept" in constructing the DEB. Automatically relating disease and findings concepts derived from UMLS vocabularies is nontrivial: not all UMLS vocabularies have robust conceptual relationships defined, and related concepts between different vocabularies often cannot be algorithmically linked.

Reviewer Comments

Reviewer comments from the drug-finding pair reviews denoted that many pairs were "too broad" (e.g., Lidocaine and *Drug toxicity*), "not relevant" (e.g., Praziquantel and *Sheep Diseases*), "weak associations" (e.g., omeprazole and *tachycardia*) or "did not make sense" (e.g., *Protease Inhibitor* and *Occupational Diseases*). We believe these comments indicate that metrics in addition to UMLS semantic type should be considered to ensure the relevance of a finding. Concepts such as "adverse effect" are too broad, even though the concept has one of the relevant semantic types. Additionally, comments revealed that reviewers had differing mindsets; for instance, some reviewers would indicate that "both" was a valid choice in their comment, but would select the predominant explanation instead (marking either AE or IND instead of "BOTH" when rating the pair). This inconsistency likely decreased the inter-rater agreement.

Limitations

There was surprisingly little overlap in the drug-finding pairs extracted from MRCOC, NDF-RT, and the SPLs. Nearly 97% of drug-finding pairs came from one of the three knowledge sources (i.e., the other sources did not mention the pair). Thus, in the majority of cases, no corroboration existed to ensure the reliability of the drug-finding pairs entered in the evidence base. Using additional knowledge sources to "back up" DEB entries could potentially increase its reliability. One explanation, previously noted, for why the overlap between sources was low relates to our observation that similar finding concepts did not have the same UMLS CUI representations in different sources, leading to entry of multiple finding concepts for what was essentially the same concept.

The UMLS contains over 140 source terminologies, each containing differing concepts at different levels of specification and meant for different purposes. The resulting discrepancies lead to imprecise matching, imprecise categorization by experts, and incomplete mapping among concepts within the UMLS. One potential solution might be to limit the vocabulary of allowed finding concepts, possibly by using a terminology designed for adverse effects concepts, such as COSTART or MedDRA – after eliminating their own internal inconsistencies.

While we mapped multiple similar drug concepts to a single specific generic ingredient, we did not do the same with finding concepts. Concepts such as "diabetes," "diabetes mellitus," "diabetes mellitus Type I," and "diabetes mellitus Type II" are all related, and could represent the same finding. It was straightforward to map drug concepts such as "Tylenol" and "Acetaminophen 500mg," to the generic ingredient "Acetaminophen," but it is difficult to conclude if "diabetes" the be mapped to the generic "diabetes mellitus" concept, or to the more specific Type I or Type II concepts. Having multiple finding concepts for what is actually one finding results in multiple drug-finding pairs from disparate sources, instead of a single pair with evidence from multiple sources. Future work should focus on ways to correctly generalize and combine these finding concepts.

There were also several drug concepts that we could not automatically map to their generic ingredient concepts. This was due to missing relationships (has_tradename, has_ingredient) in either RxNorm and/or the UMLS MRREL file. While these relationships are present for the majority of drugs, they were not present for several obscure drugs or and for some recently added drug concepts. As newer versions of

RxNorm and UMLS correct these problems, automated re-generation of the DEB content could take advantage of these improvements.

As discussed above, the reviewers considered a number of drug-finding pairs "too broad" or "too vague". A more extensive manual review to eliminate irrelevant identified finding concepts is one option. Another possible solution would be to include more knowledge sources, such as MicroMedex or SemMED (58), and require pairs to be in at least two sources. From our analysis, it seems that using only the UMLS semantic type to define the finding concepts is not restrictive enough. Other UMLS information, such as the hierarchy of concepts, might be useful in restricting finding concepts to more specific terms. We plan to investigate this in future work.

Our expert physician reviewers rated a total 725 drug-finding pairs from the DEB total of over 100,000 pairs. A larger expert review sample and more detailed analysis of DEB might give a more complete picture of the accuracy of the DEB drug-finding pair categorizations. As reviewers marked many of the drug-finding pairs in the analysis sample as "both" or "neither," a larger sample might enable the Kappa values to indicate additional areas of significant inter-rater agreement.

Having one quarter of reviewed drug-finding pairs in the DEB marked as irrelevant is a concern, but the primary purpose of the evidence base is for the classification of drug-finding pairs extracted from clinical notes. When using the DEB for that purpose, we are unlikely to identify "irrelevant" concepts such as "sheep diseases." Therefore, while the "irrelevant" drug-finding pairs constitute noise in the DEB that should ideally be eliminated, we do not believe that their current presence would have a major adverse effect when applying the DEB to clinical notes.

Another limitation of the current study was use of NLP methods to extract information from the SPLs. While NLP is an excellent tool for extracting certain types of discrete information from text, it can also misidentify concepts or identify concepts out of context. Manual review of the SPLs discovered that indications for prescribing a medication were sometimes mentioned in the text of the adverse effects section. This might occur, for example, when withdrawal of a medication (e.g., clonidine) exacerbates the underlying condition that it is used to treat (e.g., hypertension). This phenomenon might cause findings to be improperly classified as adverse effects when they should be classified as indications. We believe we could compensate for this type of confounding in the DEB by using additional knowledge sources to overcome "false" signals generated during NLP. Additionally, generalizing symptom concepts (as discussed above) could help to identify when similar concepts appear in both the AE and IND sections, but are tagged with different CUIs. Improvements in NLP methods may also help alleviate this problem in the future.

CHAPTER IV

DRUG-FINDING CORRELATION AND ADVERSE EFFECT DISCOVERY

Overview

We analyzed Vanderbilt de-identified clinical notes to determine the feasibility of discovering new, unreported adverse drug effects. We did so through analysis of drug-finding pairs mentioned in clinical notes. This exploratory evaluation asked four general questions: (1) Do drug-finding pair classifications derived by applying the DEB to clinical notes appear at face value reasonable, i.e., do DEB categorizations match known drug INDs or AEs most of the time that NLP detects a pairing known to DEB? (2) For instances when DEB classifications appear to be incorrect, is there at least a plausible reason that explains the misclassification (e.g., a common diagnosis appears as false positive confounder, or a diagnosis causally linked to an actual indication appears as false positive confounder)? (3) Does there appear to be, at face value, a known clinical reason that can be cited to explain those clinical-note-derived drug-finding pairs that are highly statistically correlated (whether it be AEs, INDs, or confounding due to comorbid conditions)? (4) Can the NLP-based clinical note correlation algorithm rank recently discovered drug-AE pairs high enough to suggest future potential for more careful side effect discovery?

From the Vanderbilt SD described above, we extracted 500,000 notes that had tags indicating that they might possibly be History and Physical Examination (H&P) notes. Based on identified subsections present in the notes (as determined by SecTag), we

algorithmically selected a subset of the original 500,000 sample notes that were most likely to represent H&Ps. From those "more definite" H&Ps, we extracted finding and drug concepts using the Vanderbilt-developed NLP tools KMCI and MedEx, respectively. We applied the same definitional requirements (based on UMLS semantic types) that we used in DEB construction, to refine extracted finding and drug concepts from the H&Ps. We also manually created a filter to remove many of the less-relevant concepts from consideration. We then carried out statistical correlation analysis of drug-finding pairs appearing in the H&P notes, and determined which pairs were represented in the DEB and SIDER.

Materials

This work was performed on a MacBook Pro with a 2.8 GHz Intel Core 2 Duo processor and 8 GB of RAM and a Linux server with forty-eight 2.2 GHz AMD Opteron cores and 256 GB RAM. All data processing scripts were written in Perl. 5.10.0. The project used MySQL version 5.5.16 for the DEB and to store all results.

Methods

Source of Clinical Notes Used in Analysis

The de-identified Vanderbilt SD repository does not contain specific type identifiers for each type of clinical note entry it contains. In the source EMR system at Vanderbilt, multiple types of clinical notes are all tagged with the same subtype "HP." Therefore, the project had to determine an automated method to identifying those notes

that were highly likely to represent full, thorough H&Ps of the type done on admission to the hospital or generated during a complete evaluation during a new patient clinic visit. Using a set of approximately 5000 known H&Ps taken from the SD (and not included in the sample we later analyzed), we identified criteria to extract likely H&Ps from the SD.

We first obtained a large group of "candidate H&P" notes. Using an SQL query, we extracted notes only from the HP, HPH, and HPL tables (HP, HP Hidden, and HP Large) in the SD, as we believed these were more likely to contain H&P notes. We limited our query based on the subtype field in the SD. We limited the search to note subtypes with the words "history" and "physical", "HP", "Admission Note" or some combination. For this study, we focused only on adult patients and excluded any notes where the subtype contained the words "pediatric," "pccu," or "nicu." We also specified a minimum note length of 3000 characters, based on the average lengths of the known H&Ps we studied.

```
where ( ( sub_type LIKE '%HISTORY%' and sub_type LIKE '%PHYSICAL%' )
  or ( sub_type like '%HISTORY \& PHYSICAL%' )
  or ( sub_type like '%HISTORY AND PHYSICAL%' )
  or ( sub_type like '%ADMISSION NOTE%' )
  or ( sub_type like '%H\&P%' )
  or ( sub_type = 'HP' )
  )
  and
  ( sub_type NOT LIKE '%PEDIATRIC%' )
  and
  ( sub_type NOT LIKE '%PCCU%' )
  and
  ( sub_type NOT LIKE '%NICU%' )
  and length(content) > 3000;
```

Figure 7. Partial MySQL code used to extract notes from the SD.

As the IRB approval restricted our study to include only 500,000 notes, we retrieved the 500,000 candidate H&P notes with the greatest length from our results set

(since we assumed longer notes were more likely to be actual H&Ps). Programmers from the Vanderbilt Informatics Center responsible for the SD helped write and perform the query.

The candidate H&P notes were written between February 15, 1984, and February 28, 2011. Due to the de-identification process, we did not have dates for each individual notes.

Running KM/SecTag to Identify Note Sections and Finding Concepts

The Vanderbilt Informatics Center provided the 500,000 candidate H&P notes in the form of plain text files. Files were pre-processed to remove extraneous line breaks and HTML/XML tags. We ran KMCI (with the negation option on) and SecTag concurrently to both index the concepts in the notes and identify those sections present in the note. SecTag produces as output an XML document containing tags identifying where H&P sections begin and end throughout the note. KMCI produces multiple output files; we used the "detailedcuis.txt" output file. It contains one entry per line for each concept identified in the note, along with information such as CUI, semantic type, whether or not the concept was negated, and (when run with SecTag) in what H&P section the concept appeared.

Identifying "Adequately Extensive" H&Ps from the Candidate H&P Set

To identify adequately extensive, more certain H&Ps from among the 500,000 candidate H&Ps, we utilized the SecTag output indicating which sections were present in each candidate note. Using expert-guided empirical analyses of 5000 notes not part of

the 500,000 samples, we explored multiple criteria to determine which sections should be present in a high-quality (as defined above) H&P note. The expert examined notes using differing empirical criteria over multiple iterations. We arrived at three potential criteria that could potentially identify an "adequately extensive H&P note" (according to our criteria, shown below in Table 7). To further ensure that a note was an H&P, we required it to satisfy at least two of the three criteria.

Table 7. Three section criteria to identify likely H&P notes.

Criterion	Sections Present in the note
1	(History of Present Illness OR Past Medical History) AND Physical Exam
2	At least 4 of (Review of Systems | Chief Complaint | Assessment | Family Medical History | Medications | (History of Present Illness | Past Medical History))
3	At least 5 of (Vital Signs | Pulmonary Exam | Cardiovascular Exam | Neurological Exam | HEENT Exam | Abdominal Exam | Lymphatic Exam | Extremity Exam | General Exam)

Running MedEx to Identify Drug Concepts

From the 366,545 H&Ps notes that appeared to be "adequately extensive" H&Ps as determined by applying our criteria, we extracted H&P sub-sections that were deemed to contain the patient's current or recent medications (i.e., past medications, as opposed to medications that the clinician planned to prescribe as a result of the examination). Thus, the medication list identified drugs the patient was taking, or had very recently take, at the time of generation of the note (corresponding to admission for inpatient H&Ps). We identified the following H&P sections as likely sites for past or current medications:

- History of Present Illness
- Past Medical History
- Chief Complaint
- Problem List
- Medications
- Current Medications
- Admission Medications
- Medication History
- Medications Outside Hospital
- Medications at Transfer
- Inpatient Medications
- Outpatient Medications
- Current Antibiotics
- Oncologic History
- Personal and Social History
- Ethanol Use
- Tobacco Use

For each note, we exported the text of these relevant H&P sections into a separate file that we then used as input for MedEx. MedEx produces output comprised of CUIs of identified drug concepts, along with any associated drug signature information, such as dose, route, and strength. To ensure that a MedEx-identified drug concept actually represented a medication taken by the patient, we required that the H&P note from which

the drug name was extracted had to mention in addition at least one item of signature information.

Identifying and counting drug-finding pairs within H&P notes

For each H&P, we constructed a patient's *finding list* from the KMCI "detailedcuis.txt" output and a patient's *medication list* from the MedEx output, as follows:

We added a finding to the *finding list* if 1) it was of a relevant finding semantic type, as previously defined, 2) it was not in an "excluded" H&P section, and 3) it had not been seen before in the current note. We excluded family history sections since they are likely to contain findings not directly related to the patient. Specifically, we excluded the SecTag sections *Family Medical History*, *Family History*, and *Mother-, Father-, Brother-,* and *Sister-Medical History*.

For each drug, we first generalized the drug concept to its generic ingredient (as described earlier). Next, we added a drug to the *medication list* if 1) it was of the appropriate drug semantic type, as previously defined, 2) at least one of the MedEx signature fields (dose, strength, route, etc.) was present, and 3) it had not been seen before in the current note.

Next, for each pairwise combination of a drug and a finding from the two sets extracted from a given H&P, the overall drug-finding count for that specific pair was incremented by one. We did this over the entire set of H&P notes to generate counts for every drug-finding pair present in the corpus of H&Ps.

Since mentions of drugs and findings may occur at random as well as for a reason, we empirically set a threshold to only consider a drug-finding pair if it occurred in at least 100 of our several hundred thousand "adequately extensive" H&P notes. For every drug-finding pair that appeared greater than this minimum, we reported the *drug-count* – the number of notes in which the drug concept was present, the *finding-count* – the number of notes in which the finding concept was present, and the *co-occurrence count* – the number of notes in which the pair was present. For each drug-finding pair, we calculated the odds ratio and Pearson's chi-squared test statistic.

Removal of Unsuitable Drug and Finding Concepts

After calculating the drug-finding co-occurrences for all H&Ps, I generated the list of all distinct drug concepts identified, including those that occurred less than 100 times. Because some drugs do not have RxNorm mappings back to their generic ingredient, I manually reviewed this list to identify drugs not properly generalized. I then coded manual mappings for these drugs into the drug-generalization procedure and re-calculated the co-occurrences.

Similarly, we generated a list of all distinct finding concepts identified by KMCI. We then sorted this list by descending count of occurrence and manually reviewed this list to remove finding concepts that were non-descriptive or too broad, for example "blood for culture," "non-specific positive culture finding," "able," "date of admission," and "annual exam."

Applying the DEB to Highlight Known and Unknown Drug-Finding Pairs

We combined the DEB described above with the SIDER information (that we previously used for DEB comparison) into a combined DEB/SIDER table stored in a MySQL database. This combination added an additional 29,451 drug-finding pairs to the existing DEB of 137,194 pairs. During the merger, in the case of a disagreement between SIDER and DEB, we used the categorization of a pair in the DEB (i.e., as an IND or an AE). When we applied the combined drug effects database, we reported classifications of H&P drug-finding pairs from H&Ps using SIDER and DEB individually.

For each of the drug-finding pairs (with a co-occurrence count greater than 100 out of the set of 366,545 H&Ps), we checked for a matching pair in the combined DEB/SIDER table. For visual analysis, we exported the pairs to a Microsoft Excel spreadsheet, along with the above-mentioned counts and statistics. Each pair listed in DEB/SIDER as an AE was highlighted in red, while each pair listed as an IND was highlighted in green. By identifying known pairs, we were better able to identify possibly unknown correlations.

We used the chi-squared test statistic as the primary means of sorting pairs based on strength of correlation, using a target p-value cutoff of 0.001. Since we performed the chi-square test for every drug finding pair, we used a Bonferroni correction to correct for multiple testing and reduce false positives. This resulted in an absolute threshold of $p \leq$ 1x10-8, corresponding to a chi-square value greater than or equal to 32.84. As a secondary means of corroborating significant correlation, we also used an odds ratio of greater than two based on empirical examination of the results and confirmation from a statistician that an odds ratio of two was a reasonable cutoff.

Analysis of Drug-Finding Pair Correlations

We examined the top (highest chi-square values) 100 drug-finding pairs ranked in descending (highest to lowest) order to empirically determine if our most common drug-finding pairs fit with known clinical information. A senior project clinician reviewed the list and made comments. We also manually reviewed the automated AE/IND tagging from DEB for accurate classification of selected drugs. For the selected drugs, we manually examined co-occurrences looking for known INDs and AEs to determine if it might be feasible to identify novel AEs using our methodology. In particular, we identified several recent FDA drug recalls and other recently reported FDA warnings regarding possible adverse associations. While DEB data focus almost exclusively on individual drugs, we also analyzed combined co-occurrence data for one drug class – HMG-CoA reductase inhibitors, popularly known as statins. To define the statin class, we combined the data from the following drugs: lovastatin, simvastatin, fluvastatin, pravastatin, atorvastatin, cerivastatin, and rosuvastatin. We empirically examined the H&P drug-finding extraction results looking for interesting pairs, or pairs that illustrated certain limitations.

We examined correlation results for the specific drugs rofecoxib, rosiglitazone, and risperidone, as well as the statin drug class and the top 100 correlations, according to chi-square and odds ratio. This analysis did not use scientific metrics; instead we performed an informal analysis in order to identify any obvious flaws and to ascertain the potential for using our approach to identify novel AEs, not to make scientific conclusions. In particular, we manually identified known adverse effects and indications

that significantly correlate with the given drugs, as well as potential confounders and trivial or irrelevant associations.

Results

Our original, more general SD query to identify "candidate" H&P notes, performed by Vanderbilt Informatics Center programmers, returned 570,845 candidate H&P notes. Due to IRB restrictions, we only obtained and examined the 500,000 notes from that set with the greatest length. We thus received the text of 496,229 "candidate" H&P notes. Of these, we successfully processed 494,661 notes using KMCI and SecTag. Not all could be processed, due to problems with the encoding of some of the notes. After applying our criteria for what constituted an "adequately extensive" H&P note, we further reduced the original number to 366,600 H&Ps. After running MedEx on their extracted relevant sections, we derived a final set of 366,545 H&Ps from which we obtained drug-finding pairs. Some of the 366,600 H&Ps did not mention any drugs and were excluded from further analysis..

From the set of 366,545 H&P notes, we extracted a total of 809,478 drug-finding pairs composed of 1755 distinct drugs and 10,723 distinct findings. After requiring a minimum co-occurrence count of 100 for further retention of a given drug-finding pair in the dataset, 75,749 drug-finding pairs remained with 666 distinct drugs and 2182 distinct findings.

When we applied the combined DEB/SIDER database to classify extracted pairs, we identified 10,500 known AEs (8066 from DEB and 2434 only in SIDER) and 3417 known INDs (3232 from DEB and 185 only in SIDER).

As noted in Methods, to compensate for the approximately 100,000 chi-square tests performed, the Bonferroni correction for our original threshold p-value of 0.001 produced a new p-value threshold of 1x10-8. For the chi-square test corresponding to 1x10-8, a significant correlation required chi-square value of greater than 32.84. Of the drug finding pairs that occurred at least 100 times, 39,304 pairs had a significant chi-square value above our threshold and of those, 20,004 also had an odds ratio greater than two.

Tables below show the highest-ranked correlations and selected results for rofecoxib, rosiglitazone, and risperidone, as well as for the statin drug class and the top 40 correlations, according to chi-square and odds ratio.

We applied DEB from Part 1 (above) to highlight the known drug-finding pairs. Entries highlighted in green have been tagged by DEB/SIDER as indications, entries highlighted in red have been tagged as AEs, and entries absent from the DEB/SIDER knowledge base (unknown) have white backgrounds. Determinations that came solely from SIDER are listed with an asterisk. Full results are available electronically.

Although the chi-square values of the co-occurrences exceeded our threshold, indicating a statistically significant correlation between given drug and finding concepts in the H&Ps, the chi-square values do not imply that the pair is a "true drug-finding pair" – that is, "drug causes finding" or "drug treats finding." For example, the findings are often not independent, both due to synonymy and due to "hidden" interrelationships among diseases and manifestations of those diseases that are concurrently listed as findings. While correlations between concepts may be statistically significant, future work must determine better methods or criteria to separate "true drug-finding pairs" from

those due to known confounders (e.g., drug co-occurs often with a comorbid condition or related findings for the actual indicated condition).

As stated above, our exploratory analysis addressed the following four questions below for the top correlations, each of our selected drugs, and the statin drug class:

1. Do drug-finding pair classifications using DEB appear at face value reasonable, i.e., either matching known drug INDs or AEs most of the time that DEB detects a pairing?
2. For instances when DEB classifications appear to be incorrect, is there at least a plausible reason that explains the misclassification (e.g., common diagnosis appears as false positive confounder, or diagnosis linked to actual indication appears as false positive confounder)?
3. Does there appear to be, at face value, a reason behind those drug-finding pairs that are highly statistically correlated (whether it be AEs, INDs, or confounding due to comorbid conditions)?
4. Can the correlation algorithm rank recently discovered drug-AE adverse effects high enough to suggest future potential for more careful side effect discovery?

Table 8. Top overall chi-square correlations (in decreasing order of chi-square).

Drug	Finding	cocount	odds	chisq	det	Expert Review
Thyroxine	Hypothyroidism	13422	59.93	**122517.76**	IND	OK
Dornase Alfa	Pancreatic Insufficiency	773	637.71	**105067.22**		Confounder, due to CF
Dornase Alfa	Cystic Fibrosis	1418	1658.53	**90518.37**	IND	OK
Tobramycin	Pancreatic Insufficiency	647	368.44	**72462.81**		Confounder, due to CF
Tobramycin	Cystic Fibrosis	1212	346.65	**64923.85**	IND	OK
Allopurinol	Gout	2778	79.57	**61419.85**	IND	OK
Insulin	Diabetes Mellitus, Insulin-Dependent	6179	32.76	**55082.08**	IND	OK
Furosemide	Congestive heart failure	11955	12.04	**44120.11**	IND	OK
Nitroglycerin	Coronary Arteriosclerosis	10379	17	**42400.06**	IND	OK
Colchicine	Gout	1650	90.31	**40544.75**	IND	OK
Insulin	Diabetes Mellitus	11478	10.59	**36228.16**	IND	OK
Lactulose	Hepatic Encephalopathy	747	116.26	**35601.03**	IND	OK
Aspirin	Coronary Arteriosclerosis	19026	6.83	**35539.91**		Prophylaxis and early RX; IND
Statins	Hyperlipidemia	15536	7.73	**35356.23**	IND	OK
valacyclovir	Graft-vs-Host Disease	765	96.44	**33656.09**		Confounder
Albuterol	Asthma	9549	10.01	**32429.05**	IND	OK
donepezil	Dementia	901	96.29	**31183.81**	IND	OK
Cyclosporine	Graft-vs-Host Disease	875	80.65	**31032.96**	IND	OK
Nitroglycerin	Chest Pain	9501	11.42	**29787.4**	IND	OK
clopidogrel	Coronary Arteriosclerosis	7289	14.41	**28112.3**		IND
Illicit Drugs	abnormal bruising	728	87.24	**28061.35**		Too broad
Digoxin	Congestive heart failure	4728	15.16	**26264.48**	IND	OK
Sinemet	Parkinson Disease	756	115.75	**25794.43**		Multi-component drug; IND
latanoprost	Glaucoma	663	97.64	**24977.36**	IND*	OK
Statins	Coronary Arteriosclerosis	15692	5.34	**24296.85**		IND
mesalamine	Crohn's disease	610	101.51	**23912.73**	IND	OK
Cocaine	Cocaine Abuse	552	98.09	**23906.61**		Trivial
Albuterol	Exacerbation of asthma	2553	30.84	**23675.4**	AE	**Incorrect – IND**
Illicit Drugs	No pain	729	65.77	**23650.24**		Trivial
Aspirin	Hypertensive disease	33022	4.51	**23593.69**	IND	**Confounder**, stroke/MI prophylaxis
Hydroxy-chloroquine	Lupus Erythematosus, Systemic	572	86.91	**23029.24**	IND	OK
mesalamine	Ulcerative Colitis	423	105.89	**22653.9**	IND	OK
Levetiracetam	Seizures	2804	37.2	**22565.16**	IND	OK
Insulin	Diabetes Mellitus, Non-Insulin-Dependent	7624	7.81	**22347.89**	IND	OK
Statins	Hypertensive disease	30117	4.65	**22289.33**	IND	**Confounder**, used in stroke/MI prophylaxis
tamsulosin	Benign prostatic hypertrophy	1430	31.73	**22267.01**	IND	OK
Insulin	Diabetic Ketoacidosis	2014	47.45	**22213.8**	IND	OK

Top Overall Correlations

For those drug-finding pairs (in Table 8) tagged by DEB/SIDER as IND or AE, 24 out of 27 pairs were correctly categorized according to our expert review. The DEB/SIDER classification was incorrect on 3 pairs for two obvious reasons. The pair Albuterol-Exacerbation from Asthma was incorrectly categorized as an AE because the finding concept Exacerbation from Asthma was extracted via NLP from the Adverse Reactions section of the SPL, but only the concept Asthma was present in the Indications section. The pairs Aspirin-Hypertensive Disease and Statins-Hypertensive Disease were both incorrectly categorized as IND because the pairs were extracted from the MRCOC as indications, likely due hypertension being a comorbidity of the conditions the drugs were intended to treat (Coronary Artery Disease and hyperlipidemia, respectively).

For the drugs categorized by DEB/SIDER, the reason for their correlation is indicated by the IND/AE determination and the reviewer comments. Five of the uncategorized pairs were highly statistically correlated due to confounding by comorbid conditions; for example, Dornase Alfa and tobramycin are correlated with pancreatic insufficiency likely because pancreatic insufficiency is a result of cystic fibrosis. Four of the uncategorized pairs were indications, one of which was likely missed because it was a multi-ingredient drug. Three of the uncategorized pairs were too trivial or broad to be of use, but seemed reasonable. For these top correlations, there were no serious AEs ranked. Since indications should always co-occur with medications but adverse effects only rarely co-occur, it is not surprising that no AEs would be in these highest-ranked correlations.

Table 9. Top chi-square correlation results for the drug Rofecoxib.

Drug	Finding	cocount	odds	chisq	det
rofecoxib	Degenerative polyarthritis	250	3.35	**318.07**	IND
rofecoxib	Obesity	253	2.58	**188.01**	
rofecoxib	Hypertensive disease	598	2	**138.74**	AE*
rofecoxib	Arthritis	157	2.63	**135.18**	IND*
rofecoxib	Prothrombin time increased	101	3.1	**129.33**	
rofecoxib	Rheumatoid Arthritis	212	2.21	**113.16**	IND*
rofecoxib	Congestive heart failure	170	2.32	**107.98**	AE*
rofecoxib	Metabolic Diseases	216	2.1	**100.06**	
rofecoxib	Myocardial Infarction	189	2.17	**98.77**	AE*
rofecoxib	Chest Pain	267	1.95	**94.2**	AE*
rofecoxib	Coronary Arteriosclerosis	248	1.98	**92.85**	
rofecoxib	White blood cell count increased	233	1.96	**86.54**	
rofecoxib	Mental Depression	238	1.9	**80.1**	
rofecoxib	Shortness of Breath	260	1.77	**66.08**	
rofecoxib	Lupus Erythematosus, Discoid	145	1.99	**61.54**	
rofecoxib	Gastroesophageal reflux disease	212	1.8	**60.93**	AE*
rofecoxib	Adverse Event Associated with the Gastrointestinal System	107	2.1	**55.35**	
rofecoxib	Back Pain	119	2.02	**54.62**	AE*
rofecoxib	Swelling	113	1.93	**44.95**	
rofecoxib	Pain	521	1.49	**44.48**	IND*
rofecoxib	Hypothyroidism	129	1.83	**42.89**	
rofecoxib	Osteoporosis	114	1.87	**41.58**	
rofecoxib	Asthenia	137	1.76	**39.41**	AE*
rofecoxib	Gastrointestinal tract finding	112	1.85	**39.09**	
rofecoxib	Diabetes Mellitus	198	1.6	**36.66**	
rofecoxib	Chronic Obstructive Airway Disease	126	1.67	29.89	
rofecoxib	Urinary tract infection	121	1.67	28.81	AE*
rofecoxib	Anemia	135	1.61	27.52	
rofecoxib	Lesion	273	1.44	27.43	
rofecoxib	Cerebrovascular accident	136	1.57	24.86	AE*
. . .					

Top Correlations for Rofecoxib

For those drug-finding pairs in Table 9 tagged by DEB/SIDER as IND or AE, all are correct (94) except for two: back pain is an indication for rofecoxib, not an AE, and asthenia was considered too broad. Both of these incorrect classifications came from SIDER, not DEB.

For the drugs correctly categorized by DEB/SIDER, the reason for their correlation is indicated by the IND/AE determination. The uncategorized findings highly statistically correlated with rofecoxib seem to be mostly due to confounders, such as obesity, mental depression, hypothyroidism, diabetes mellitus, and COPD. Some of these findings, such as coronary arteriosclerosis and shortness of breath, may be associated with cardiovascular findings tagged by DEB/SIDER as AEs.

As is shown in the Table 9, rofecoxib use is highly correlated with the known AE myocardial infarction and other cardiac affects including congestive heart failure, chest pain, and coronary arteriosclerosis. The known AE cerebrovascular accident, or stroke, also occurs often with rofecoxib use, too, but not above our current threshold for significance. In this case, known AEs are present in the top correlations with rofecoxib, but do not stand out due to confounding and incomplete tagging of known drug-finding pairs.

Table 10. Top chi-square correlation results for the drug rosiglitazone.

Drug	Finding	cocount	odds	chisq	det
rosiglitazone	Diabetes Mellitus, Non-Insulin-Dependent	608	9.11	**2416.6**	IND
rosiglitazone	Diabetes Mellitus	745	8.77	**2334.6**	IND
rosiglitazone	Hypertensive disease	1028	5.02	**849.05**	AE
rosiglitazone	Obesity	420	3.85	**611.06**	
rosiglitazone	Hyperlipidemia	384	3.44	**475.98**	
rosiglitazone	Coronary Arteriosclerosis	396	2.78	**320.68**	
rosiglitazone	Gastroesophageal reflux disease	300	2.13	**139.92**	
rosiglitazone	Lupus Erythematosus, Discoid	209	2.37	**139.78**	
rosiglitazone	hypercholesterolemia	164	2.54	**133.66**	
rosiglitazone	Anicteric	808	1.82	**123.49**	
rosiglitazone	Arthritis	177	2.36	**119.52**	
rosiglitazone	Angina Pectoris	116	2.71	**113.74**	
rosiglitazone	Chronic Obstructive Airway Disease	197	2.18	**107.52**	
rosiglitazone	Dyspnea on exertion	153	2.21	**89.42**	
rosiglitazone	Congestive heart failure	190	2.06	**88.6**	AE
rosiglitazone	Shortness of Breath	326	1.79	**87.12**	
rosiglitazone	Orthopnea	136	2.27	**86.62**	
rosiglitazone	Myocardial Infarction	214	1.95	**83.16**	AE
rosiglitazone	Paroxysmal atrial tachycardia	296	1.79	**81.46**	
rosiglitazone	Anemia	193	1.9	**70.68**	AE
rosiglitazone	Visual impairment	111	2.23	**68.43**	
rosiglitazone	Stenosis	103	2.27	**67**	
rosiglitazone	Mental Depression	273	1.73	**66.77**	
rosiglitazone	Deep vein thrombosis of lower limb	141	2.04	**66.68**	
rosiglitazone	Pulmonary Embolism	139	1.94	**56.41**	
rosiglitazone	Nausea	436	1.47	**44.97**	AE
rosiglitazone	Cerebrovascular accident	180	1.7	**44.6**	
. . .					

Top Correlations for Rosiglitazone

As shown in Table 10, of the seven highest correlated findings with the drug rosiglitazone categorized by DEB/SIDER, two are correctly categorized as indications and five are AEs (myocardial infarction and congestive heart failure are

suspected AEs and are included in the SPL warnings). It is also important to note that MI and CHF could also be due to confounding; patients with diabetes mellitus are much more likely to have heart disease.

For the other drug-finding pairs with statistically correlated with rosiglitazone, most appear to be confounders, including obesity, hyperlipidemia, and hypercholesterolemia – comorbidities that often occur with Type II diabetes mellitus. Suspected AEs of rosiglitazone include heart disease, stroke, bone fractures, eye damage, and liver damage (95). There are statistically significant correlations with *coronary arteriosclerosis*, *angina pectoris*, *congestive heart failure*, *myocardial infarction*, and *chest pain* (chi-square 33.04, odds ratio 1.48). There is also a correlation with stroke (cerebrovascular accident) with a chi-square value of 44.6, but an odds ratio of only 1.7, correlation with two known complications of diabetes, *visual impairment* (chi-square 68.43, odds ratio 2.23) and *diplopia* (chi-square 42.44, odds ratio 1.9), and co-occurrences below our current significance threshold with liver-related findings (*jaundice* with chi-square of 23.44, odds ratio 1.59, and *liver cirrhosis* with chi-square 10.84, odds ratio 1.3). As before, these AE signals are present in the data, but are lost among the confounders.

Table 11. Top chi-square correlation results for the drug Risperidone.

Drug	Finding	cocount	odds	chisq	det
Risperidone	Schizophrenia	308	42.2	**8710.9**	IND
Risperidone	Iron deficiency anemia	240	17.13	**2914.91**	
Risperidone	Mental disorders	121	25.9	**2422.33**	IND
Risperidone	Dementia	269	13.01	**2372.91**	IND
Risperidone	HYPOKINESIS GLOBAL	164	16.7	**2039.94**	
Risperidone	Poor historian	136	12.5	**1253.73**	
Risperidone	Hypothyroidism	411	4.84	**919.79**	AE*
Risperidone	Epilepsy	497	4.2	**829.72**	IND
Risperidone	Bipolar Disorder	178	6.12	**661.89**	IND
Risperidone	Abnormal mental state	191	5.42	**594.93**	
Risperidone	Diabetes Mellitus, Insulin-Dependent	202	5.06	**564.57**	
Risperidone	Agitation	210	4.83	**544.8**	IND
Risperidone	Congestive heart failure	327	3.3	**413.96**	
Risperidone	Diabetes Mellitus	444	2.83	**378.33**	AE
Risperidone	Hypovolemia	156	4.48	**374.39**	
Risperidone	Psychiatric problem	169	4.27	**372.97**	
Risperidone	Coronary Arteriosclerosis	455	2.73	**354.89**	
Risperidone	Obesity	393	2.84	**351.31**	AE
Risperidone	Diabetes Mellitus, Non-Insulin-Dependent	329	2.95	**333.44**	IND
Risperidone	Hypertensive disease	937	2.41	**321.51**	AE*
Risperidone	Rhonchi	115	4.62	**297.31**	
Risperidone	Anicteric	1022	2.24	**256.69**	
Risperidone	Alzheimer's Disease	157	3.23	**215.55**	IND
Risperidone	Hypoxia	138	3.39	**210.24**	
Risperidone	Chest Pain	423	2.21	**205.94**	AE*
Risperidone	Eosinophilia-Myalgia Syndrome	103	3.91	**205.85**	
Risperidone	Myocardial Infarction	297	2.41	**198.11**	AE*
Risperidone	Urinary hesitation	105	3.65	**186.25**	
Risperidone	Deep vein thrombosis of lower limb	198	2.52	**157.91**	
Risperidone	Mental Depression	358	2.01	**139.78**	IND
Risperidone	Confusion	145	2.67	**136.59**	IND
Risperidone	Tobacco use	154	2.59	**134.66**	
Risperidone	Wheezing	162	2.48	**127.85**	
Risperidone	Cerebrovascular accident	250	2.09	**118.38**	AE
Risperidone	Deglutition Disorders	169	2.32	**112.72**	AE
Risperidone	Liver Cirrhosis	284	1.9	**99.54**	
. . .					

Top Correlations for Risperidone

At face value, many of the DEB/SIDER categorizations of findings correlated with risperidone (Table 11) appear to be valid; 9/10 INDs are correct, but *Diabetes Mellitus, Non-Insulin-Dependent* is incorrect. This categorization came from the MRCOC and is likely due to research in MEDLINE related to risperidone use in diabetics. Of the 8 pairs classified as AEs, the 4 from DEB are correct (diabetes mellitus, obesity, stroke, and deglutition disorders) while the 4 from SIDER do not appear to be true (95).

Risperidone's known AEs include weight gain, diabetes, and stroke. The concepts *obesity* and *weight gain* (chi-square 36.12, odds ratio 1.71), *diabetes mellitus*, *diabetes mellitus, non-insulin-dependent*, and *cerebrovascular accident* are all highly statistically significant, but are lost among many other "significant" but confounding drug-finding pairs.

Table 12. Top chi-square correlation results for the statins, with select findings.

Drug	Finding	cocount	odds	chisq	det
Statins	Hyperlipidemia	15105	9.12	40606.57	IND
Statins	Coronary Arteriosclerosis	15198	6.29	28375.31	
Statins	Hypertensive disease	27734	5.48	23711.55	IND
Statins	hypercholesterolemia	7640	7.28	19811.4	IND
Statins	Myocardial Infarction	8130	3.52	8697.75	IND
Statins	Stenosis	4229	4.75	7460.27	
Statins	Diabetes Mellitus	9362	2.79	6471.34	IND
Statins	Peripheral Vascular Diseases	3601	4.52	6048.25	IND
Statins	Diabetes Mellitus, Non-Insulin-Dependent	6909	2.98	5787.67	IND
Statins	Angina Pectoris	3632	4	5192.13	IND
Statins	Congestive heart failure	6180	2.94	5128.04	
Statins	Cerebrovascular accident	6428	2.64	4328.46	IND
Statins	Epilepsy	7264	2.5	4260.15	
Statins	Ischemic cardiomyopathy	1776	5.95	4198.61	
Statins	Retina-normal	1438	6.7	3834.55	
Statins	Chest Pain	9109	2.14	3520.67	
Statins	Ischemia	3179	3.34	3519.73	IND
Statins	Arthritis	4769	2.64	3341.17	
Statins	Dyslipidemias	1783	4.73	3278.79	IND
Statins	Congenital leukocyte adherence deficiency	1612	4.82	3041.16	
Statins	Obesity	6953	2.16	2928.76	IND
Statins	Gastroesophageal reflux disease	7576	2.09	2861.92	
Statins	Mental Depression	8030	2.05	2834.82	AE
. . .					
Statins	Memory impairment	281	1.82	**84.34**	
Statins	Memory observations	104	1.5	**14.92**	
Statins	Memory loss	331	1.23	**12.3**	
. . .					

Top Correlations for the Statins

Unlike the other exploratory analyses presented for single drugs, Table 12 shows correlations for the statins drug class (as defined in Methods section). The DEB categorization of findings correlated with statin use appear reasonable; of the top 13 drug-finding pairs categorized as indications, all seem to be correct (95). The

one tagged AE, *mental depression,* appears to be incorrect as statins have actually been linked to lower risk of depression (96).

Many of the top correlations with statins appear to be reasonable, as well; most are comorbid conditions or findings of those on statins. Recently discovered AEs for statins include memory loss and Type II diabetes mellitus, and were the subject of a March 2012 FDA warning. Memory-related AEs co-occur, including *memory impairment*, *memory loss*, and *memory observations*, but not to a statistically significant degree. If these concepts were merged, however, they would likely reach our study's threshold for statistical significance. Multiple diabetes concepts appear to have a significant correlation, including *Diabetes Mellitus*, *Diabetes Mellitus, Non-Insulin-Dependent, diabetic retinopathy, diabetic neuropathies, diabetes mellitus, insulin-dependent, diabetic, Diabetic nephropathy, proliferative diabetic retinopathy*, *nonproliferative diabetic retinopathy, diabetic foot ulcer,* and *diabetic gastroparesis*. Merging these concepts would increase statistical significance, but since DEB has correctly tagged it as an indication, this alone might not cause the signal to stand out. Again, the signals of known AEs are present, suggesting future potential for adverse effect discovery, but they are hidden among confounders.

Discussion & Limitations

Summary

While we found many "statistically significant correlations" between drugs concepts and finding concepts in the corpus of H&P notes, many of the drug-finding

pairs were not truly significant in terms of representing meaningful drug knowledge – that is, "drug causes finding" or "drug treats finding." While many known AEs were highly ranked in our results, we also found far too many "statistically significant correlations" with confounders and other non-relevant concepts for this to be a viable method to detect novel AEs at this time. However, as our project intended an exploratory analysis to assess the feasibility of this approach, we believe we were successful, and the future applicability of the approach will improve if one can filter the noise in the data.

The findings in our analyses were often not independent, both due to synonymy and due to "hidden" interrelationships among diseases and manifestations of those diseases that were concurrently listed as findings. While correlations between concepts may have been statistically significant, future work must determine better methods or criteria to separate "true drug-finding pairs" from those due to known confounders (e.g., drug co-occurs often with a comorbid condition or related findings for the actual indicated condition).

Among the many true and false drug-finding pairs in our results, we identified serious AEs associated with rofecoxib, rosiglitazone, risperidone, and the statins that were only discovered during post-marketing surveillance at a time years after each drug had been on the market. Our results illustrate that drug-AE relationships are present in H&P data, if only one can find exact and reliable methods to identify them – and to verify such correlations through subsequent independent studies.

We have extended the NLP-base approach use by Wang, et al., (4) on Discharge Summaries to show that NLP processing can identify potentially useful drug-finding relationships in H&P notes. We performed our analyses using a much larger number of

notes than Wang, et al. Using H&P notes instead of discharge summaries enabled us to consider a more complete range of finding concepts as potential AEs because the H&Ps done at admission focus on presenting findings in detail, whereas discharge summaries tend to focus on the clinical course of an inpatient stay, where findings play a less critical role than therapies and outcomes. Our mining of H&P finding concepts, however, also drastically increased the number of drug-finding pairs that derived from comorbidities, and potentially reduced the precision of our approach. Additionally, by incorporating the retrieval of drug and finding concepts only from particular sections in the notes (5) and using a combined knowledge base to automatically identify adverse effects and indications, similar to (59), we developed the basic principles behind a method that may eventually automatically identify correlations and require little manual review to generate potential drug-AE hypotheses.

Limitations

Due to the exploratory nature of our study, the biggest limitation of our project was its lack of formal methods to make definitive conclusions about discovered drug-finding pair correlations during the search for novel AEs. As discussed above, confounding related to comorbidities was a significant source of noise in the data. We address possible methods to reduce or eliminate such confounding in the Future Work section below. We further understand that indicating a potential correlation is not the same as showing causation. Any statistical discovery approach requires subsequent independent confirmation of "interesting" results. Nevertheless, our exploratory analysis

has been successful in indicating the potential of approach to identify drug-AE hypotheses from EMR-derived correlation data.

Another limitation of our study involved the ambiguous nature of the finding (diseases and manifestations of diseases) concepts. There are often multiple UMLS CUIs with very similar meanings that appeared in the H&Ps. For example, the concepts "asthma" and "exacerbation of asthma" are obviously related, but they were treated as completely separate concepts in both the DEB and in terms of the co-occurrence calculations. We describe a potential approach to this problem below in "Future Work".

Finally, an important limitation of our project involved the inexact nature of NLP processing of unconstrained text documents. While NLP approaches allowed us to successfully extract discrete and computable data from clinical text, it limits our conclusions to what NLP tools can "discover". While KMCI is very effective at correctly identifying UMLS finding concepts in clinical notes, there are occasional misidentifications. For example, NLP systems can recognize most forms of negation, but not all. When we noticed that the concept "Exposure to HIV infection" (C0262514) seemed to occur too often in our dataset, we examined the KMCI output files. Often, KMCI would identify the concept in H&Ps that state "HIV exposure negative" or "HIV: negative." No NLP system is perfect, and such limitations must be recognized. An NLP system typically requires pre-specification of "target concepts" to identify; it is likely that a more constrained set of potential findings than that specified in the UMLS might have produced different results.

CHAPTER V

CONCLUSION

The goal of this project was to explore the feasibility of adverse drug effect discovery from the data mining of clinical notes. We have shown that a large corpus of H&Ps has embedded signals that are potentially useful for AE discovery. Further statistical and biomedical research must conclusively prove any adverse effects discovered through any retrospective correlation analysis. One cannot draw conclusions solely from disproportionality analysis or other methods; definitive proof of an adverse effect can only come from a convergence of experimental, clinical, and statistical research (33). Our approach, while promising, cannot be used "as is" to detect new adverse effects.

The project has also demonstrated the potential value of using H&P notes as a source of medication and finding-related information. This extends more than a decade of previous NLP work involving analysis of clinical documents. Unlike spontaneous reporting systems (SRS), EHR data results in better estimates for both the "numerator" and "denominator." Additionally, one can compare associated findings from multiple medications. These both serve to compliment traditional randomized controlled trials and observational studies to provide large sample sizes.

We have learned several lessons during our project. First, identifying known correlations to remove them from consideration as novel AEs, such as indications and known AEs, is a difficult task. Second, while using NLP and UMLS finding concepts is

an effective way to discretize data in natural language text, many of the concepts are too specific, too broad, or too vague. Additionally, there are often too many concepts that refer to the same condition ("diabetes," "diabetes mellitus," "diabetes mellitus, non-insulin-dependent," etc.). Third, false signals appear frequently since a drug often co-occurs not only with the finding that it is used to treat, but also with other findings that co-occur with the finding in questions (for example, insulin treats diabetes, but co-occurs frequently with heart disease since heart disease often accompanies diabetes). Fourth, using a hard cutoff for significance based upon chi-square and odds ratios causes many potentially meaningless correlations to appear statistically significant; other measures of meaningful significance should be explored. We discuss possible ways to address these problems in the following section.

Future Work

Known Correlations

For novel adverse effect signals to better stand out, it would be useful to be able to better identify known reasons for drug-finding correlation. We believe the addition of more knowledge sources would improve the DEB. Sources such as the AERS and SemMed, as used by (58), could enhance the accuracy of DEB. Additionally, a certainty metric, based upon the number of sources from which a drug-finding pair has been extracted, would also improve the usefulness of tagging known correlations. The greatest improvement in the DEB, however, would come from the development of a method to generalize and combine finding concepts (discussed below).

Finding Concept Generalization

To reduce the number of confounding concepts, it would be useful to develop a method for generalizing finding concepts similar to how we mapped drug concepts to their generic ingredient. This would both reduce the number of distinct finding concepts and appropriately increase the counts for remaining relevant finding concepts. Finding concepts, however, are more complex than drug concepts. The symptoms and findings associated with a disease are often variable, and could be attributed to more than one disease concept or even to the drug as an adverse effect. Methods for how and when to combine symptoms must be explored. We plan to explore the use of the UMLS concept hierarchy as a possible source determining related CUIs and condensing them into a single CUI. We are considering the development of expert-derived finding synonyms to condense finding concepts as well.

Comorbidities

Co-morbid conditions, another major source of confounding, must also be addressed. Knowledge-based tools such as QMR include information on common co-morbidities for given diseases, as well as symptoms and findings that occur with these diseases. We believe this information can be applied to reduce confounding or identify inappropriate correlations, such as those between insulin and heart disease. Additionally, data mining techniques, such as frequent itemset mining of symptoms and diseases, might be useful in determining which symptoms are likely due to which diseases, similar to (52).

Comparing Drug-Finding pairs of similar-indication drugs to identify AEs

One potential method to overcome confounding using the current results (or similar results) might be to compare the findings associated with a given drug to the findings associated with a different drug (or drugs) with similar indications and similar target population. By identifying which findings occur with all of the drugs in question, one should be able to distinguish indications or common comorbid conditions of the target patient population. If certain findings occur more often with one drug than the others, that finding is possibly due to the drug in questions – and is a potential adverse effect. We believe this technique has the potential to compensate for confounding from comorbid conditions, but it needs to be both formalized and automated.

Statistical Significance of Drug-Finding Pairs

Finally, there were often too many pairs deemed significant for a given drugs. We believe it is necessary to explore the use of other statistics, or a combination of statistics and other methods, to improve the recognition of meaningful drug-finding pairs. The chi-square test, while appropriate for this type of data, often rejects the null hypothesis (no association between drug and finding) too often for very large samples. Significance cutoffs identified with volume tests developed by (97) and used in (98) have been shown to be more correct when dealing with clinical data. Other data mining techniques and standard data mining significance measures, such as cosine similarity, lift, and support, should be further explored on our H&P co-occurrence data, as well (99).

Improved NLP

Finally, we believe that improved NLP could also enhance the success of this approach towards the discovery of novel adverse effects. We used NLP to identify known AEs and indications from SPLs for the construction of DEB, and to identify UMLS concepts present in H&P notes to calculate the co-occurrence. Improved NLP could enhance the accuracy of this data. We would like to further NLP tools that take advantage of the structure of clinical notes to help disambiguate concepts. For example, if the word "rub" is mentioned in the H&P section "Cardiovascular Exam," it is likely referring to a friction rub of the heart; if the same word is mentioned in the "Orthopedic Exam" section, it is likely referring to a joint rub. Improved concept generalization (above) could also be used to help develop special vocabularies for NLP tools that can facilitate the generalization similar finding concepts.

REFERENCES

1. Roden DM. An underrecognized challenge in evaluating postmarketing drug safety. Circulation 2005 Jan;111(3):246–248.[cited 2012 Jan 10]

2. Wilson AM, Thabane L, Holbrook A, Wilson AM, Thabane L, Holbrook A. Application of data mining techniques in pharmacovigilance, Application of data mining techniques in pharmacovigilance. British Journal of Clinical Pharmacology, British Journal of Clinical Pharmacology 2004 Feb;57, 57(2, 2):127, 127–134, 134.[cited 2012 Apr 19]

3. Amery WK. Why there is a need for pharmacovigilance. Pharmacoepidemiology and Drug Safety 1999;8(1):61–64.

4. Wang X, Hripcsak G, Markatou M, Friedman C. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. Journal of the American Medical Informatics Association 2009 Mar;16(3):328–337.[cited 2012 Jan 10]

5. Wang X, Chase H, Markatou M, Hripcsak G, Friedman C. Selecting information in electronic health records for knowledge acquisition. Journal of Biomedical Informatics 2010 Aug;43(4):595–601.[cited 2012 Jan 5]

6. 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 Feb;15(1):87–98.[cited 2012 Jan 5]

7. Understanding Clinical Trials - ClinicalTrials.gov [Internet]. [date unknown];[cited 2012 Mar 23] Available from: http://clinicaltrials.gov/ct2/info/understand

8. Research C for DE and. Postmarket Drug Safety Information for Patients and Providers - Vioxx (rofecoxib) Questions and Answers [Internet]. [date unknown];[cited 2012 Mar 27] Available from: http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm106290.htm

9. Topol EJ. Failing the Public Health — Rofecoxib, Merck, and the FDA. N Engl J Med 2004;351(17):1707–1709.

10. McGuire S. FDA toughens Avandia warnings [Internet]. MMM 2007;[cited 2012 Apr 1] Available from: http://www.mmm-online.com/fda-toughens-avandia-warnings/article/96354/

11. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N. Engl. J. Med. 2007 Jun;356(24):2457–2471.[cited 2012 Apr 1]

12. Cheung BM. Behind the rosiglitazone controversy. Expert Rev Clin Pharmacol 2010 Nov;3(6):723–725.[cited 2012 Apr 1]

13. Food and Drug Administration. Decision on continued marketing of rosiglitazone (Avandia, Avandamet, Avandaryl) [Internet]. 2010;Available from: http://www.fda.gov/downloads/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/UCM226959.pdf

14. European Medicines Agency. European Medicines Agency recommends suspension of Avandia, Avandamet and Avaglim [Internet]. [date unknown];[cited 2012 Mar 27] Available from: http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2010/09/news_detail_001119.jsp

15. Center for Drug Evaluation and Research. Drug Safety and Availability - FDA Drug Safety Communication: Update to ongoing safety review of Actos (pioglitazone) and increased risk of bladder cancer [Internet]. [date unknown];[cited 2012 Apr 1] Available from: http://www.fda.gov/Drugs/DrugSafety/ucm259150.htm

16. Topol EJ. The Diabetes Dilemma for Statin Users [Internet]. The New York Times 2012;[cited 2012 Mar 27] Available from: http://www.nytimes.com/2012/03/05/opinion/the-diabetes-dilemma-for-statin-users.html

17. Food and Drug Administration. FDA announces safety changes in labeling for some cholesterol-lowering drugs [Internet]. [date unknown];[cited 2012 Apr 1] Available from: http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm293623.htm

18. Thomas K. Drug Giant Is Fined $1.2 Billion in Risperdal Case [Internet]. The New York Times 2012;[cited 2012 Apr 22] Available from: http://www.nytimes.com/2012/04/12/business/drug-giant-is-fined-1-2-billion-in-arkansas.html

19. Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, Farrar K, Park BK, Breckenridge AM. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. BMJ 2004 Jul;329(7456):15–19.[cited 2012 Mar 31]

20. Kongkaew C, Noyce PR, Ashcroft DM. Hospital Admissions Associated with Adverse Drug Reactions: A Systematic Review of Prospective Observational Studies. Ann Pharmacother 2008 Jul;42(7/8):1017–1025.[cited 2012 Mar 31]

21. Atiqi R, van Bommel E, Cleophas TJ, Zwinderman AH. Prevalence of iatrogenic admissions to the Departments of Medicine/Cardiology/ Pulmonology in a 1,250 bed general hospital. Int J Clin Pharmacol Ther 2010 Aug;48(8):517–524.[cited 2012 Apr 2]

22. WHO | Pharmacovigilance [Internet]. WHO [date unknown];[cited 2012 Apr 2] Available from: http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/index.html

23. US Department of Health and Human Services. Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment [Internet]. 2005;Available from: http://www.fda.gov/downloads/regulatoryinformation/guidances/ucm126834.pdf

24. Office of the Commissioner. Safety of Specific Products - FDA Uses A Number Of Approaches To Assess Postmarketing Risk [Internet]. [date unknown];[cited 2012 Apr 3] Available from: http://www.fda.gov/Safety/SafetyofSpecificProducts/ucm180551.htm

25. Srba J, Descikova V, Vlcek J. Adverse drug reactions: Analysis of spontaneous reporting system in Europe in 2007-2009 [Internet]. European Journal of Clinical Pharmacology 2012 Feb;[cited 2012 Apr 9] Available from: http://www.ncbi.nlm.nih.gov/pubmed/22294060

26. EudraVigilance - Pharmacovigilance in EEA [Internet]. [date unknown];[cited 2012 Apr 3] Available from: http://eudravigilance.ema.europa.eu/human/index.asp

27. Lindquist M. VigiBase, the WHO Global ICSR Database System: Basic Facts. Drug Information Journal 2008;42:409–419.

28. Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases [Internet]. Stat Methods Med Res 2011 Aug;[cited 2012 Apr 13] Available from: http://www.ncbi.nlm.nih.gov/pubmed/21878461

29. Moulis G, Sommet A, Durrieu G, Bagheri H, Lapeyre-Mestre M, Montastruc J-L. Trends of reporting of “serious” versus “non-serious” Adverse Drug Reactions over time: a study in the French PharmacoVigilance Database [Internet]. British Journal of Clinical Pharmacology 2012 Jan;[cited 2012 Apr 9] Available from: http://www.ncbi.nlm.nih.gov/pubmed/22257367

30. Motola D, Piccinni C, Biagi C, Raschi E, Marra A, Marchesini G, Poluzzi E. Cardiovascular, Ocular and Bone Adverse Reactions Associated with Thiazolidinediones: A Disproportionality Analysis of the US FDA Adverse Event Reporting System Database. Drug Saf 2012 Apr;35(4):315–323.[cited 2012 Apr 9]

31. Mozzicato P. The Role of MedDRA in Pharmacovigilance Activities [Internet]. 2011;Available from: http://www.apec-ahc.org/files/tp201105/Patricia_Mozzicato_Plenary_Apr_28_1030am_Revised.pdf

32. Almenoff JS. Innovations for the future of pharmacovigilance. Drug Saf 2007;30(7):631–633.[cited 2012 Apr 11]

33. Montastruc J-L, Sommet A, Bagheri H, Lapeyre-Mestre M. Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database. Br J Clin Pharmacol 2011 Dec;72(6):905–908.[cited 2012 Mar 26]

34. Hauben M, Zhou X. Quantitative methods in pharmacovigilance: focus on signal detection. Drug Saf 2003;26(3):159–186.[cited 2012 Apr 13]

35. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001 Nov;10(6):483–486.[cited 2012 Apr 16]

36. Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. Pharmacoepidemiol Drug Saf 2004 Aug;13(8):519–523.[cited 2012 Apr 23]

37. Alvarez Y, Hidalgo A, Maignen F, Slattery J. Validation of statistical signal detection procedures in eudravigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling. Drug Saf 2010 Jun;33(6):475–487.[cited 2012 Apr 16]

38. van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002 Feb;11(1):3–10.[cited 2012 Apr 13]

39. Hanover J, Hodges A. Reading the Signals [Internet]. Int Clin Trials 2006;Spring 2006Available from: http://www.colibris-doc.com/portfolio/Mar06_ICT_Reading%20the%20Signals_J%20Hanover.pdf

40. Harpaz R, Haerian K, Chase HS, Friedman C. Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. AMIA Annu Symp Proc 2010;2010:281–285.[cited 2012 Mar 26]

41. Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinformatics 2010;11 Suppl 9:S7.[cited 2012 Mar 26]

42. Harpaz R, Perez H, Chase HS, Rabadan R, Hripcsak G, Friedman C. Biclustering of Adverse Drug Events in the FDA's Spontaneous Reporting System. Clin Pharmacol Ther 2011 Feb;89(2):243–250.

43. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med 2012 Mar;4(125):125ra31.[cited 2012 Apr 9]

44. Glasgow JM, Kaboli PJ. Detecting adverse drug events through data mining. Am J Health Syst Pharm 2010 Feb;67(4):317–320.[cited 2012 Mar 31]

45. Honigman B, Lee J, Rothschild J, Light P, Pulling RM, Yu T, Bates DW. Using computerized data to identify adverse drug events in outpatients. J Am Med Inform Assoc 2001 Jun;8(3):254–266.[cited 2012 Mar 28]

46. Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically screening discharge summaries for adverse medical events. J Am Med Inform Assoc 2003 Aug;10(4):339–350.[cited 2012 Mar 28]

47. FitzHenry F, Peterson JF, Arrieta M, Waitman LR, Schildcrout JS, Miller RA. Medication administration discrepancies persist despite electronic ordering. J Am Med Inform Assoc 2007 Dec;14(6):756–764.[cited 2012 Apr 18]

48. Matheny ME, Miller RA, Ikizler TA, Waitman LR, Denny JC, Schildcrout JS, Dittus RS, Peterson JF. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. Med Decis Making 2010 Dec;30(6):639–650.[cited 2012 Apr 18]

49. Schildcrout JS, Haneuse S, Peterson JF, Denny JC, Matheny ME, Waitman LR, Miller RA. Analyses of longitudinal, hospital clinical laboratory data with application to blood glucose concentrations. Stat Med 2011 Nov;30(27):3208–3220.[cited 2012 Apr 18]

50. Field TS, Gurwitz JH, Harrold LR, Rothschild JM, Debellis K, Seger AC, Fish LS, Garber L, Kelleher M, Bates DW. Strategies for detecting adverse drug events among older persons in the ambulatory setting. J Am Med Inform Assoc 2004 Dec;11(6):492–498.[cited 2012 Mar 28]

51. Hazlehurst B, Naleway A, Mullooly J. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. Vaccine 2009 Mar;27(14):2077–2083.[cited 2012 Mar 28]

52. Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical narrative reports. AMIA Annu Symp Proc 2008;783–787.[cited 2012 Mar 26]

53. FDA's Sentinel Initiative [Internet]. [date unknown];[cited 2012 Apr 3] Available from: http://www.fda.gov/Safety/FDAsSentinelInitiative/default.htm

54. DailyMed [Internet]. [date unknown];[cited 2012 Feb 8] Available from: http://dailymed.nlm.nih.gov/dailymed/about.cfm

55. Micromedex [Internet]. [date unknown];[cited 2012 Apr 23] Available from: http://www.micromedex.com/

56. Drug Data | FDB (First Databank) [Internet]. [date unknown];[cited 2012 May 4] Available from: http://www.fdbhealth.com/

57. UpToDate Inc. [Internet]. [date unknown];[cited 2012 May 4] Available from: http://www.uptodate.com/index

58. Wang X, Chase HS, Li J, Hripcsak G, Friedman C. Integrating heterogeneous knowledge sources to acquire executable drug-related knowledge. AMIA Annu Symp Proc 2010;2010:852–856.[cited 2012 Jan 10]

59. Li Y, Salmasian H, Harpaz R, Chase H, Friedman C. Determining the Reasons for Medication Prescriptions in the EHR using Knowledge and Natural Language Processing. AMIA Annu Symp Proc 2011;2011:768–776.[cited 2012 Jan 10]

60. Fact SheetUnified Medical Language System® [Internet]. [date unknown];[cited 2012 Apr 10] Available from: http://www.nlm.nih.gov/pubs/factsheets/umls.html

61. Fact SheetUMLS® Metathesaurus® [Internet]. [date unknown];[cited 2012 Apr 10] Available from: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

62. Fact SheetUMLS® Semantic Network [Internet]. [date unknown];[cited 2012 Apr 10] Available from: http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

63. Fact SheetSPECIALIST Lexicon [Internet]. [date unknown];[cited 2012 Apr 10] Available from: http://www.nlm.nih.gov/pubs/factsheets/umlslex.html

64. Denny JC. "Understanding" Medical School Curriculum Content Using KnowledgeMap. Journal of the American Medical Informatics Association 2003 Mar;10(4):351–362.[cited 2012 Jan 5]

65. Denny JC, Miller RA, Waitman LR, Arrieta MA, Peterson JF. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. Int J Med Inform 2009 Apr;78 Suppl 1:S34–42.[cited 2012 Mar 28]

66. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, Peterson NB. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 2010 Aug;17(4):383–388.[cited 2012 Mar 28]

67. Denny JC, Choma NN, Peterson JF, Miller RA, Bastarache L, Li M, Peterson NB. Natural language processing improves identification of colorectal cancer testing

in the electronic medical record. Med Decis Making 2012 Feb;32(1):188–197.[cited 2012 Mar 28]

68. Xu H, Lu Y, Jiang M, Liu M, Denny JC, Dai Q, Peterson NB. Mining Biomedical Literature for Terms related to Epidemiologic Exposures. AMIA Annu Symp Proc 2010;2010:897–901.

69. Denny JC, Ritchie MD, Crawford DC, Schildcrout JS, Ramirez AH, Pulley JM, Basford MA, Masys DR, Haines JL, Roden DM. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. Circulation 2010 Nov;122(20):2016–2021.[cited 2012 May 23]

70. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, Basford MA, Brown-Gentry K, Balser JR, Masys DR, Haines JL, Roden DM. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am. J. Hum. Genet. 2010 Apr;86(4):560–572.[cited 2012 May 23]

71. Denny JC, Miller RA, Johnson KB, Spickard A 3rd. Development and evaluation of a clinical note section header terminology. AMIA Annu Symp Proc 2008;156–160.[cited 2012 Mar 28]

72. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association 2010 Jan;17(1):19–24.[cited 2012 Jan 5]

73. i2b2: Informatics for Integrating Biology & the Bedside [Internet]. [date unknown];[cited 2012 Apr 20] Available from: https://www.i2b2.org/

74. Doan S, Bastarache L, Klimkowski S, Denny JC, Xu H. Integrating existing natural language processing tools for medication extraction from discharge summaries. J Am Med Inform Assoc 2010 Oct;17(5):528–531.[cited 2012 Apr 10]

75. Xu H, Doan S, Birdwell KA, Cowan JD, Vincz AJ, Haas DW, Basford MA, Denny JC. An automated approach to calculating the daily dose of tacrolimus in electronic health records. AMIA Summits Transl Sci Proc 2010;2010:71–75.[cited 2012 Mar 28]

76. Xu H, Jiang M, Oetjens M, Bowton EA, Ramirez AH, Jeff JM, Basford MA, Pulley JM, Cowan JD, Wang X, Ritchie MD, Masys DR, Roden DM, Crawford DC, Denny JC. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. J Am Med Inform Assoc 2011 Aug;18(4):387–391.[cited 2012 Mar 28]

77. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin. Pharmacol. Ther. 2008 Sep;84(3):362–369.[cited 2012 Apr 20]

78. UMLS® Reference Manual - NCBI Bookshelf [Internet]. National Library of Medicine; 2009.[cited 2012 Feb 8] Available from: http://www.ncbi.nlm.nih.gov/books/NBK9676/

79. MEDLINE Backfiles Source Information [Internet]. [date unknown];[cited 2012 Apr 10] Available from: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MBD/

80. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc 2011 Aug;18(4):441–448.[cited 2012 Apr 10]

81. Pathak J, Chute CG. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. Journal of the American Medical Informatics Association 2010 Jul;17(4):432–439.[cited 2012 Jan 5]

82. Carter JS, Brown SH, Erlbaum MS, Gregg W, Elkin PL, Speroff T, Tuttle MS. Initializing the VA medication reference terminology using UMLS metathesaurus co-occurrences. Proc AMIA Symp 2002;116–120.[cited 2012 Feb 21]

83. 2011AA National Drug File - Reference Terminology Source Information [Internet]. [date unknown];[cited 2012 Feb 8] Available from: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/

84. Chute CG, Carter JS, Tuttle MS, Haber M, Brown SH. Integrating pharmacokinetics knowledge into a drug ontology: as an extension to support pharmacogenomics. AMIA Annu Symp Proc 2003;170–174.[cited 2012 Feb 21]

85. Structured Product Labeling Resources [Internet]. [date unknown];[cited 2012 Feb 8] Available from: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm

86. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol. Syst. Biol. 2010;6:343.[cited 2012 Feb 8]

87. Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART). 5th ed. Rockville, MD: US Food and Drug Administration, Center for Drug Evaluation and Research; 1995.

88. The PubChem Project [Internet]. [date unknown];[cited 2012 Apr 10] Available from: http://pubchem.ncbi.nlm.nih.gov/

89. SIDER Side Effect Resource [Internet]. [date unknown];[cited 2012 Feb 22] Available from: http://sideeffects.embl.de/

90. Structured Product Labeling > SPL - Downloadable Data [Internet]. [date unknown];[cited 2012 Feb 8] Available from: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm240580.htm

91. Structured Product Labeling > Section Headings (LOINC) [Internet]. [date unknown];[cited 2012 Feb 8] Available from: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm162057.htm

92. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159–174.[cited 2012 Apr 21]

93. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005 Mar;85(3):257–268.[cited 2012 Apr 21]

94. Vioxx (rofecoxib) - Drug information from MediLexicon [Internet]. [date unknown];[cited 2012 May 5] Available from: http://www.medilexicon.com/drugs/vioxx.php

95. Lehne RA. Pharmacology for nursing care. St. Louis, Mo.: Saunders/Elsevier; 2010.

96. Otte C, Zhao S, Whooley MA. Statin Use and Risk of Depression in Patients With Coronary Heart Disease [Internet]. The Journal of Clinical Psychiatry 2012 Feb;[cited 2012 May 5] Available from: http://article.psychiatrist.com/dao_1-login.asp?ID=10007753&RSID=86191273553399

97. Cao H, Hripcsak G, Markatou M. A statistical methodology for analyzing cooccurrence data from a large sample. J Biomed Inform 2007 Jun;40(3):343–352.

98. 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;106–110.[cited 2012 Jan 5]

99. Han J, Kamber M. Data mining: concepts and techniques. Amsterdam; Boston; San Francisco, CA: Elsevier; Morgan Kaufmann; 2006.