Mining Electronic Health Records For Adverse Drug Effects Using Regression Based Methods

Rave Harpaz
Department of Biomedical Informatics
Columbia University
New York, NY
rave.harpaz@dbmi.columbia.edu

Krystl Haerian
Department of Biomedical Informatics
Columbia University
New York, NY
krh7003@dbmi.columbia.edu

Herbert S. Chase
Department of Biomedical Informatics
Columbia University
New York, NY
herbert.chase@dbmi.columbia.edu

Carol Friedman
Department of Biomedical Informatics
Columbia University
New York, NY
friedman@dbmi.columbia.edu

ABSTRACT
The identification of post-marketed adverse drug events (ADEs) is paramount to health care. Spontaneous reporting systems (SRS) are currently the mainstay in pharmacovigilance. Recently, electronic health records (EHRs) have emerged as a promising and effective complementary resource to SRS, as they contain a more complete record of the patient, and do not suffer from the reporting biases inherent to SRS. However, mining EHRs for potential ADEs, which typically involves identification of statistical associations between drugs and medical conditions, introduced several other challenges, the main one being the necessity for statistical techniques that account for confounding. The objective of this paper is to present and demonstrate the feasibility of a method based on regression techniques, which is designed for automated large scale mining of ADEs in EHR narratives. To the best of our knowledge this is a first of its kind approach that combines both the use of EHR data, and regression based methods in order to address confounding and identify potential ADEs. Two separate experiments are conducted. The results, which are validated by clinical subject matter experts, demonstrate great promise, as well as highlight additional challenges.

Categories and Subject Descriptors
J.3 LIFE AND MEDICAL SCIENCES: Medical information systems

General Terms
Algorithms

Keywords
Adverse drug events, electronic health records, regression, confounding.

1. INTRODUCTION
Drug safety is one of the main concerns in health care, and it is generally well accepted that not all possible safety issues associated with drugs can be identified during the pre-marketed clinical trial phase. Post-marketed adverse drug events (ADEs) results in significant costs, estimated in several billion dollars annually, and cause unnecessary and often fatal harm to patients[1,2]. The objective in pharmacovigilance is the early detection of novel post-marketed ADEs with minimal patient exposure.

The current mainstay within pharmacovigilance are spontaneous reporting systems, which are database resources containing millions of voluntarily submitted reports of suspected ADEs occurring during regular clinical practice. These reports are typically mined for statistical drug-event associations to screen for unknown potential ADEs that are then clinically validated and flagged for continued monitoring. Among the major SRSs are: the United States Food and Drug Administration's (FDA) Adverse Event Reporting System (AERS)[3], and the World Health Organization (WHO) Programme for International Drug Monitoring [4].

Due to their voluntary nature of reporting, SRS are susceptible to several well recognized limitations[5,6], most notably, data quality issues and reporting biases, which may severely effect or lead to erroneous study conclusions. Among these limitations are: the phenomena of under reporting or over reporting due to media influences, subjective diagnoses by the reporter of the event, uneven levels of granularity used to describe or encode the drugs and events involved, duplicity of reporting for the same patient and event, missing data, and typographical errors.

Recently, electronic health records (EHRs) have emerged as a promising and effective complementary resource to SRS in pharmacovigilance[7,8] . Unlike reports submitted to SRS, EHRs contain a more complete record of the patient's conditions and treatments written as part of the process of care, and do not depend on the health care professional's subjective view as to which medications or events may be related to a given ADE. As a result, the main advantages for the use of EHRs over SRS in pharmacovigilance are: earlier detection of ADEs, potential for active and real time surveillance, and the absence of most of the reporting biases attributed to SRS.

However, mining EHRs for ADEs introduce other challenges. First, unlike reports in SRS, most of the clinical information in a patient's record consists of unstructured narratives, which are unsuitable for a direct application to pharmacovigilance. Nonetheless, clinical NLP systems, such as MedLEE (Medical Language Extraction and Encoding)[9] used in this study, have been proven successful in extracting useful information from

IHI'10, November 11-12, 2010, Arlington, Virginia, USA.

clinical narratives for a wide range of applications[10,11]. MedLEE can be used to extract and encode the clinical entities (medications and conditions) relevant to the application of pharmacovigilance. The second challenge, which is the subject of this work, is that unlike reports submitted to SRS, EHRs do not normally contain information which explicitly expresses suspected ADEs or indications for medications. As a result, mining EHRs requires the examination of a much wider range of possible drug-event associations, the majority of which are completely unrelated or associated only because of confounding, where typically the confounder is another condition attributed to the patient, which is not caused by a drug.

Confounded associations are spurious associations between two entities that are mediated or influenced by a third entity not included in the analysis. Confounding is a serious concern in causality or association studies because it may lead to biased inference, in some cases suggesting a causal relationship when in truth none exists. The types of confounding that are likely to occur with drug-event associations obtained by mining EHRs are associations confounded by co-medication, co-morbidities, associations confounded by indication, and in some cases any combination of the three. The first case may occur when two drugs are frequently reported together, leading to a false association of each drug individually with the adverse event, when it is only one of the drugs that is truly associated with the event. An example is HIV patients on a multi-drug regimen. The second type may occur when two typically co-existing medical conditions are reported with a specific drug, leading to a false association between one of the conditions with the drug, when only the other condition is the true adverse response to the drug. An example is the co-existing medical condition of elevated creatine cinase (CK) and myocardial infarction (MI) associated with the drug atorvastatin, which is only known to cause the elevated CK ADE and not MI. The third type occurs when the suspected event reported is a manifestation of the disease for which the drug has been prescribed. For example, the drug clopidogrel may be strongly associated with the adverse event elevated creatine kinase, however elevated creatine kinase is a manifestation of myocardial infarction, which clopidogrel is designed to treat. This study focuses on the last two. By definition, a confounder must be associated with both variables under consideration (associated with both the drug and event), and does not lie on the causal pathway between the two (i.e., is not the sole cause for the event, and is not only caused by the drug).

Current approaches to quantitative analysis of drug safety databases are mostly based on disproportionality measures (ratios) that attempt to quantify the degree of departure of an observed incidence rate of a drug-event association from a base or control case. The most popular measure is the relative reporting ratio (RRR)[12], which is an observed to expected ratio, where expectation is considered under the assumption of independence between the drug and event. Other common disproportionality measures include: the proportional reporting ratio (PRR)[13], the reporting odds ratio (ROR)[14], and Bayesian versions of RRR, such as the gamma poison shrinker (GPS)[15], and Baysian confidence propagation neural network (BCPNN)[16]. Another class of approaches employ statistical hypothesis tests such as the chi-squared test and Fisher's exact test, which are used to test the hypothesis of independence between a pair of drug and event[12]. Common to these methods, is that they perform bivariate analysis, studying each combination of a drug and event separately. This shortcoming renders them inadequate to analyze confounded associations that involve at least one more variable, and thus require other methods that rely on multivariate as opposed to bivariate statistics.

The traditional statistical approach to confounding was based on stratification and Mantel-Haenszel test statistics[17]. Stratification approaches are effective for addressing confounding in large sample sizes and small number of confounding variables. In other cases they are not as effective. As a result, in recent years stratification has been gradually replaced by regression based methods. Regression models allow for the evaluation of several risk factors simultaneously, and as such are better suited than bivariate methods to study confounded associations. Broadly speaking, in a regression model the value of a dependent variable (e.g., the presence of a disease or adverse event) is explained by a set of predictor variables (e.g., different exposures or drugs), each with its own degree of contribution, which are estimated from the data as the parameters of the model. By incorporating into the model potential confounding variables, the effect or influence of the confounding variables on the predictor variables could be assessed to determine whether or not the relationship between the dependent and predictor variables is influenced by the confounders. In this case, the model is said to be "controlling for possible confounders". In addition, it turns out that the parameters of the regression models used in this study are equivalent to the ROR association measure, which eases the interpretability of the model and provides a direct link to the more common bivariate association analysis methods.

A special set of regression techniques known as regularized or shrinkage regression methods, which have been widely used in various data mining applications, especially "bag-of-words" type modeling[18,19], have recently been proposed to the application of pharmacovigilance[20,21]. These techniques are designed to deal with very large sparse data sets, that contain many more variables (predictors) than observations, and as such provide a natural setting in which the thousands of drugs, conditions, or events that exist in drug safety databases, can be analyzed simultaneously for drug-event associations. Caster et al. [20,21] demonstrated the effectiveness of a shrinkage based regression method to mine the WHO database for ADEs confounded by co-medication.

The overall objective of this paper is to present and examine the feasibility of a method, which is designed to perform automated large scale mining of EHR narratives in order to identify potential ADEs, while at the same time addressing the challenge of confounding. The method is based on a combination of standard and shrinkage based logistic regression. To best of our knowledge this is the first of its kind study that incorporates both the use of EHR data and multivariate regression based techniques to identify unconfounded drug-event associations that represent potential ADEs. The paper describes two experiments. First, pre-selected known cases of confounded and unconfounded drug-event associations are examined in order to assess the ability of standard logistic regression, which is part of the overall method, to correctly identify confounded associations. Second, the feasibility of the proposed method is examined by its application to EHR narratives representing cases of patients with elevated CK. The

results of the experiments are validated by clinical subject matter experts.

2. DATA

Creatine Kinase, also known as creatine phosphokinase, is a metabolic enzyme found in various tissues of the body, specifically in tissues that extract energy such as cardiac and skeletal muscle or brain tissues. When this tissue become damaged, CK is released into the blood resulting in elevated levels of CK. Clinically, CK blood tests are used to diagnose several serious medical conditions such as MI and rhabdomyolysis. CK was selected as our adverse event of choice in this study because it is a result of serious adverse effects such as rhabdomyolysis. In addition, the use of CK as an effective lab signal to monitor for a possible ADE was reported in a recent pharmacovigilance study[22].

After obtaining IRB approval, the clinical data warehouse at the New York Presbyterian Hospital (NYPH) was queried for all abnormal CK laboratory tests (five times above normal). In addition, inpatient discharge summaries from the years 2004-2009, which corresponded to the patients with abnormal CK values, were extracted from the warehouse. In addition to patients with elevated CK, henceforth CK+, a large sample of discharge summaries from 2004-2009, that did not include reports of elevated CK, henceforth CK-, were also extracted from the warehouse to create our final training data set for regression modeling of both positive (presence of CK+) and negative (absence of CK+) training examples.

Next, the unstructured inpatient discharge summaries were processed using MedLEE to identify the relevant clinical entities: medications, diseases, and symptoms, in each note. Due to brevity considerations the full details of this process are omitted from this discussion, but are described in [9].

The final outcome of the data collection and processing stage resulted in a set of records, each containing a list of medications, diseases and symptoms, corresponding to both CK+ and CK- instances. The disease and symptom entity types served as potential confounders and as such were grouped into a single type of clinical entity, which we refer to henceforth as conditions. Using this data model, drug-condition-CK+ associations were then analyzed using regression based methods to determine confounding and potential CK+ ADEs.

3. METHODS & EXPERIMENTS

Throughout this section we will be using standard epidemiological notation[23], where the variable D denotes a binary disease outcome, the variables E and X denote binary exposure or risk factors, and C a potential binary confounding variable. In each of these cases the value of 1 assigned to a variable denotes the presence of the outcome or risk factor, and 0 its absence. In this study D will correspond to CK+, X the risk factor to a drug, and C to a certain medical condition.

3.1 The Odds Ratio

The odds of an outcome D is defined as $\Pr(D)/(1-\Pr(D))$. The odds ratio (OR), also known as ROR (mentioned in the introduction), is a measure of association that compares $\Pr(D|E)$ with $\Pr(D|\bar{E})$ and is defined as the odds ratio between the two, i.e.,

$$OR = \frac{\Pr(D \mid E)}{1-\Pr(D \mid E)} \Big/ \frac{\Pr(D \mid \bar{E})}{1-\Pr(D \mid \bar{E})}$$

OR=1 typically indicates independence of D and E, OR>1 that there is a greater risk of D when E is present, and OR<1 a lower risk when E is present. As will be demonstrated in the following OR is linked to logistic regression.

3.2 Ordinary Logistic Regression

In the simple case of one predictor variable the logistic regression model relates the risk of D, $p_x=\Pr(D|X=x)$ to the risk factor X through the following equation

$$p_x = \frac{1}{1+e^{-(a+bx)}}$$

Alternatively, this relationship can be expressed in terms of log odds as

$$\log\left(\frac{p_x}{1-p_x}\right) = a + bx \qquad (1)$$

where a and b are the coefficients or parameters of the model. Extending to the case of several risk factors $X_1,X_2,\ldots,X_k$, and where $p_{x1,x2,..,xk}=\Pr(D|X_1=x_1,X_2=x_2,\ldots,X_k=x_k)$ the model can be expressed as

$$\log\left(\frac{p_{x1,x2,..xk}}{1-p_{x1,x2,\ldots,xk}}\right) = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

The underlying assumption in the logistic regression model is that the log odds of D changes linearly with changes in the risk factors $X_1,X_2,\ldots,X_k$. Hence, holding k-1 predictors fixed and changing only one of them X_i describes how the log odds of D changes due to changes in X_i. b_i=0 indicates no relationship between D and X_i, b_i >0 reflects an increasing risk of D as X_i increases, and b_i <0 a decreasing risk.

As mentioned, the parameters of the model have a direct and appealing interpretation to our application. It can easily be shown that the intercept a of the model is just the log odds of D in the baseline case of no exposure to any risk factor, i.e., $X_1,=0,X_2=0,\ldots,X_k=0$. Similarly one can show that the slope parameter b_i is log odds ratio associated with a unit increase in X_i (change in X for binary exposures) holding all other variables in the model fixed, for example

$$b_1 = \log\left(\frac{p_{1,x2,..xk}}{1-p_{1,x2,\ldots,xk}} \Big/ \frac{p_{0,x2,..xk}}{1-p_{0,x2,\ldots,xk}}\right)$$

In the simple case of one predictor variable, b is simply equal to log(OR). The parameters of logistic regression models are estimated using maximum likelihood approaches [24](for the sake of brevity we omit the details on how).

3.2.1 Confounding

In the case of confounding one needs to incorporate into the model a confounding variable C and assess its impact on the

predictability of the model, specifically the log odds ratio of the variable it may be confounding. A change in the log odds ratio is typically an indicator of confounding. In the simple case of one predictor and one confounder, the model can be specified by

$$\log\left(\frac{p_x}{1-p_x}\right) = a + b_1 x + b_2 c \quad (2)$$

In this case the log odds ratio given by b_1 in Eq. 2 measures the effect of a unit increase in X holding C fixed, i.e., accounting for possible confounding by C. To test for a confounding effect in the logistic regression setting, typically a combination of tests are considered, each providing different evidence as to the effect of confounding[23]. In this work we used three tests to assess confounding.

Test 1 (*Wald test*). Test the null hypothesis H_0: $b_1=0$, that examines the association between X and D controlling for possible confounding by C. $b_1=0$ indicates X and D are independent allowing for confounding by C. Thus rejection of the null hypothesis (small p-values) implies X and D are still associated taking into account confounding by C.

Test 2 (*likelihood ratio test*). Compare the models of Eq.1 and Eq.2 by means of their respective maximum likelihood values using the *likelihood ratio statistic*. This test essentially examines the effect of C on overall model predictability. Small p-values indicate that the confounding variable C is relevant to the model, and therefore may be a possible confounder.

Test 3 (*The empirical approach*). Compare the values of the two parameters b (not accounting for confounding) and b1 (accounting for confounding) from Eq. 1 and Eq. 2 respectively. If they are very similar (within the scope of random variation) then there is little confounding. If they are significantly different (e.g. beyond 10% different) then the data suggests confounding.

3.3 Shrinkage Logistic Regression

Shrinkage regression methods incorporate the idea of shrinking the regression coefficients towards zero, producing parsimonious models particularly suited to cases where there are thousands of variables to consider as potential predictors, but only small subsets of them are relevant to the model. In regression models, variables whose coefficients have a value of zero cannot affect the outcome. Hence, shrinking the regression coefficients towards zero has the effect of eliminating variables from consideration. From a theoretical standpoint, shrinkage based regression models have the advantage of producing estimates with lower variance, which result in overall better predictability.

Shrinkage is achieved by means of imposing a penalty on large regression coefficients. The L_2 penalty restricts the sum of squared values of the coefficients resulting in a shrinkage method referred to as *ridge* regression[25]. The L_1 penalty restricts the sum of absolute values of the coefficients resulting in a shrinkage method referred to as *lasso* regression[26]. Both methods have a Bayesian interpretation. A prior belief that the coefficients are small is encoded into a prior distribution, and the mode of the posterior distribution is taken as the shrunk coefficient estimate. In ridge regression, the prior is assumed to follow a normal distribution, whereas in lasso, a Laplace distribution is assumed. Under this interpretation, shrinkage regression methods can be viewed as the multivariate counterpart to Bayesian disproportionality analysis. The main difference between ridge and lasso is that while both favor coefficients close to zero, lasso favors coefficients that exactly equal zero whereas ridge doesn't, making lasso more appropriate and the method of choice for this study.

Based on the above one can easily notice the advantage of shrinkage regression to ADE mining in EHRs. Shrinkage logistic regression allows us to select from the thousands of drugs or conditions a much smaller set of the most predictive, while still taking into account confounding as with ordinary logistic regression.

The amount of shrinkage is controlled through a pre-specified parameter, usually denoted by λ. In many cases its appropriate value will be determined by its ability to minimize predication error. In other cases, such as ours, this approach of selecting λ has been found inadequate, and a more ad-hoc approach has been suggested [20].

3.4 Experiment 1

A set of six representative cases of known drug-condition-CK+ confounded and unconfounded associations were selected by clinical subject mater experts for our population of study in order to assess the ability of ordinary logistic regression, with the inclusion of confounding variables (section 3.2), to correctly identify confounding effects. For each of the selected cases the three tests presented in section 3.2 were conducted, and a final conclusion as to whether there is statistical evidence in support of confounding was drawn based on the outcome. Confidence intervals (95%) were computed for b and b1 from Eq. 1 and Eq. 2 respectively, in order to the examine the difference between the two, necessary for Test 3. The difference was computed as the average difference between the limits of each respective interval (more stable than computing the difference between the estimated parameters).

3.5 Experiment 2

A method based on both ordinary and shrinkage logistic regression is proposed and tested in order to assess its feasibility as an automated large scale mining process used to identify unconfounded (non-spurious) ADEs in EHR discharge summaries.

The method consists of three steps, which can be summarized as follows: (1) using the OR measure of association a set of drugs strongly associated with CK+ was selected as an initial set of candidate drugs involved in CK+ ADEs. (2) Using shrinkage logistic regression a set of conditions which are strong predictors of CK+ was identified to serve as potential confounders. (3) Using ordinary logistic regression the set of drugs identified in step 1, are filtered by the set of potential confounders identified in step 2, to identify drug-CK+ associations that are not confounded, and represent the final set of potential CK+ ADEs. The following describes each of the steps in greater detail.

Step 1. Using the OR measure of association, specifically the lower limit of its 95% confidence interval, the data was scanned to identify statistically strong associations between drugs and CK+. The set of drugs identified in this step was used as an initial set of candidate drugs that can potentially be attributed to CK+

ADEs. The reason we chose bivariate association analysis over multivariate analysis is because the latter could have eliminated confounding by co-medication, and our aim was to: (a) to generate a set of strongly associated drugs, which was initially as large as possible, but would later be filtered by the confounding tests. (b) Confounding by co-medication was not the focus of this study, and seems to be less severe as compared to the other types of confounding which occur in our data. The OR measure of association was selected rather than other measures in order to be consistent with the measures and interpretation of the coefficients of regression. It has also been shown[27] that there is little difference between the measures with big enough samples, and it is agreed at this point that no measure is universally better than the other. A strong association was qualified as an association that had an OR >1 and at least 5 occurrences in the data.

Step 2. Using shrinkage logistic regression all the conditions extracted from the full set of discharge summaries were regressed against the CK indicator variable to identify a much smaller set of the strongest predictors of CK+. This set of conditions was then used as the set of potential confounders. By definition a confounder must be associated with the outcome (CK+), hence by identifying the best predictors of CK+ we are essentially selecting a set of strongly associated potential confounders, but with one added advantage of eliminating co-morbidities, and at the same time reducing redundancy of conditions and model complexity. The latter is also advantageous also from a computational and statistical aspect (greater precision in parameter estimation). Only conditions with a positive coefficient such as MI were selected, as they correspond to conditions that increase and not decrease (negative coefficients) the probability of CK+. The shrinkage controlling parameter λ was selected based on the size of the set of conditions included in the model and the sign of their respective coefficients. We conjectured that a reasonable size would be between 20-40 conditions. It also turned out that beyond this size, conditions with negative coefficients started entering the model at a higher rate.

Step 3. Each of the candidate drugs identified in step 1 was regressed separately, first by itself, and then along with the full set of potential confounders, against the CK indicator variable using ordinary logistic regression. Test 3 was then used to determine whether or not a confounding effect was present. If the presence of an effect was detected, the drug was eliminated from consideration as a possible drug associated with a CK+ ADE. The threshold used to qualify a significant difference in the drug coefficient before and after the addition of the confounders was set to a suggested value of 10% [23]. Note that each drug was tested not against one confounder only, but against the joint effect of the full set of potential confounders.

4. RESULTS

4.1 Data statistics

A total of 2,018,073 CK laboratory tests for the years 2004-2009 were identified in the clinical data warehouse at NYPH. Of those 8,172 had elevated CK levels (as per our cutoff), corresponding to 2,840 unique patients. Of these patients only 687 had inpatient discharge summaries, which were extracted from the clinical data warehouse to constitute our CK+ set of narratives. In addition a set of 2,500 unique patients who had discharge summaries and were not reported to have elevated CK levels were extracted from the clinical data warehouse to constitute our CK- set of narratives. Both sets where then combined and processed by MedLEE to extract from each patient record the medications and conditions (diseases and symptoms) used for modeling. The total population of narratives produced 1713 unique drugs and 5,630 unique conditions.

4.2 Experiment 1 - Results

The results of the first experiment are displayed in Table 1. The table shows the six pre-selected drug-condition-CK+ cases and their associated statistics: the value and confidence interval for the drug coefficient *b* before and after the addition of the potential confounder/s into the model, along with the three test statistics.

Examples A-D illustrate cases of drug-CK+ associations known to be confounded. Example A illustrates a classic case of known confounding. The example shows a spurious association between clopidogrel and CK+ which is confounded by MI. Clopidogrel is a drug that treats MI, but not known to cause CK+. However, patients with MI typically have CK+. The statistics show a strong association between clopidogrel and CK+ before the inclusion of MI as a confounder (b-before). However, after the inclusion, the strength of association is reduced by 23% suggesting a confounding effect (test 3). Tests 1 and 2 provide somewhat contradictory evidence, suggesting an association between clopidogrel and CK+ accounting for MI (test 1), and that MI should be included in the model (test 2), i.e. some confounding.

Another advantage of logistic regression is that it allows the examination of confounding not only by one potential confounder, but also by the joint effect of a set of confounders. Examples B-D illustrate cases of confounding by multiple conditions.

Example B illustrates a special case of confounding by co-morbidities, reflecting some of the difficulty associated with mining EHRs for ADEs. CK+ is a known adverse drug effect of atorvastatin, which is used to treat patients with high cholesterol. However, patients with heart related conditions such as MI, which in turn manifests CK+, typically also have high cholesterol and are treated with statins . Consequently, examining the association between atrovastatin and CK+, accounting for the heart related conditions: *stemi*, *st segment elevation*, *MI*, and *chest pain,* revealed statistical evidence in support of this confounding. Prior to the inclusion of the confounders in the model, atorvastatin is shown to be strongly associated with CK+ (b-before). After the inclusion, this association is substantially reduced by 43% (test 3), suggesting a strong confounding effect. Test 2 (extremely small p-values exceeding computer precision) also suggest a strong influence of the joint set of confounders on the atorvastatin alone model performance.

Tramadol (pain reliever) in example C is known in some situations, such as drug abuse or when taken with anti-depressants to cause seizures, which in turn elevate CK levels. On the other hand, drug addicts often receive tramadol as a substitute. Drug addicts are also very often in the state of somnolence which would lead to muscle injury and thus elevated CK. This has been found particularly true in our study population. Consequently, studying the association between tramadol and CK+, accounting for the drug abuse related states: *cocaine abuse, drug abuse, drug*

abuser, opioid abuse, heroin abuse (also considered as clinical conditions in this study), found in our patient records should indicate a confounded association. The statistics indeed confirm this, although not as decisively as in the previous example.

In example D, insulin is not known to cause CK+, but diabetics often have heart related conditions such as MI which are known to be associated with CK+, generating spurious associations between insulin and CK+. Once again the test statistics are in support of confounding. Note however, that in this case the association between insulin and CK+ before and after the addition of the confounders *MI* and *diabetes mellitus* (b-before, b-after) are not as strong as in the previous examples. The difference however is significant (15% test 3).

Examples E-F illustrate cases of associations that are not confounded. In example E, CK+ is a known adverse effect of atorvastatin, which is used to treat hyperlipidemia (high cholesterol). However, CK+ is not an indicator of hyperlipidemia by itself, and therefore should not confound the association between atorvastatin and CK+. The statistics show a strong association between atorvastatin and CK+ (b-before), probably due to its association with heart related conditions, as discussed in example B. However, the strength of association is slightly reduced (9%, test 3) after the addition of hyperlipidemia into the model (b-after), but not enough to support confounding. Test 1 suggests an association between atorvastatin and CK+ taking hyperlipidemia into account, i.e., no or very little confounding, whereas test 2 suggests keeping hyperlipidemia in the model, again slightly contradictory support.

In example F, *sirolimus* (an immunosuppressant) is known to cause CK+, and so does seizures. But the two are not related, hence seizures should not confound sirolimus. The test statistics validate this proposition. While they show a strong association between sirolimus and CK+ (b-before, b-after), they also show no difference (test 3) between the two statistics after the inclusion of seizures, pointing to a conclusion of no confounding. Tests 1 and 2 again do not provide conclusive evidence.

4.3 Experiment 2 - Results

Step 1 of the experiment resulted in a candidate set of 71 drugs that were strongly associated with CK+. Due to space limitations they are not provided in this paper.

Table 2 displays the set of potential confounders identified in Step 2 of the experiment. Out of 26 potential confounders identified, all were either plausible causes of or are associated with known causes of elevated CK. The list of confounders fell into one of several categories: known to cause of an elevated CK (myocardial infarction, cocaine abuse, compartment syndromes); associated with causes of elevated CK (chest pain, akinesia); or define muscle damage elevated CK terms (creatine kinase increased; transaminitis). There were several identified that could either be diseases which cause elevated CK or the terms used by providers to describe muscle damage due to a drug (polymyositis, myositis). The terms which had an identical meaning to CK+, such as creatine kinase increased, were removed from the set of potential confounders before the applying step 3.

Table 3 displays the set of 11 drugs identified in step 3 of the experiment (from the total of 71 drugs identified in step 1), which can possibly be attributed to a CK+ ADE, and which were found not to be confounded by the set of potential confounders from step 2. These drugs fell into three categories: those previously reported to cause elevated CK as an ADE (lamivudine, zidovudine, chlorpromazine); those which treat diseases which can cause elevated CK (carbidopa treats Parkinson's disease); those which, through inference, are plausibly a cause of elevated CK (thrombus, causing myocardial infarct), unconscious state (causing rhabdomyolysis); and those which are unlikely to be a cause of elevated CK (bicarbonates, vasopressin). Also on this list is a drug (fosphenytoin) used to treat one of the confounding conditions (seizures).

5. DISCUSSION

Overall, the results of first experiment demonstrate the potential use and advantages of logistic regression to identify confounded associations, which may be generated by mining EHRs. Sufficient statistical evidence was generated by a combination of three tests to support the hypothesis of confounding when it was truly present, while rejecting it when it was absent. Test 3, while being more qualitative than the other two tests, was the most discriminative and was consistently aligned with the known facts for each of the cases.

At times tests 1 and 2 provided contradictory evidence making them less useful in an automatic large scale mining process. In these specific examples the contradictions could be settled by the notion that while a confounding effect was present, it was not extreme nor did it completely dominate the association between a drug and CK+. It is likely that this situation will repeat itself in the full population of associations in EHRs, and demonstrates subtleties associated with the analysis of confounding, especially associations found by mining EHRs.

Test 3 proved the most suitable for an automatic large scale mining procedure, and for this reason it was the test chosen to filter confounded associations in the second experiment. However, it is more qualitative in nature than the other two tests. Specifically, the threshold selected in test 3 (10%) to qualify associations as confounded or not, was based on a value suggested in reference [23], but a different threshold would have likely generated a different set of conclusions, especially for borderline cases such as in examples C and E. A more quantitative test, such as a test to measure the statistical significance of the difference, would be beneficial. At this point we are not aware of one, and plan to investigate this issue in future research.

Overall, the results of the second experiment show great promise. The method did not generate any false positives for the set of potential confounders, i.e., all were verified to be true confounders. In the set of drugs, five are clinically known to cause the CK+ ADE, which demonstrates the ability of the method to identify true ADEs. The remaining were either unknown or unlikely to cause the CK+ ADE, or confounded by known conditions that did not appear in the set of confounders identified by the method. These false positives highlight additional challenges associated with the identification of ADE associations in EHRs, which can be primarily attributed to the

data. We illustrate these challenges and the data related issues with several examples below.

The drug Tramadol is one of the false positives that survived the filtering by potential confounders. Although there are reports of tramadol causing seizures, which in turn elevate CK, it is very unlikely that this was the case. It is more likely that in our study population the reason for tramadol being associated with CK+ is its association with drug abuse related states, which confound the association as illustrated in Table 1 example C. However, except for one state (cocaine abuse) no other drug abuse related states were included in the set of potential confounders. This was because statistically they are not (except for cocaine abuse) strong predictors of CK+, and therefore did enter the set of potential confounders. It is possible that the granularity of the terms related to drug abuse weakened the association between the general drug abuse state and CK+. Similarly, fosphenytoin was a flase positive because it is used to treat seizures, which in turn cause elevated CK. In this case seizures was not statistically associated with fosphenytoin (in fact there were no cases of the two reported together), which by definition cannot make seizures a confounder.

Another interesting case is the statins which were notably lacking from the list of final drugs, but are known to cause the CK+ ADE. This was very likely due to the fact that every patient in our study population that was admitted with myocardial infarct was likely to be on a statin. Therefore, identifying an unconfounded association between a statin and CK+ would be impossible, as illustrated in Table 1 example B. This also suggests that applying the method to an outpatient population is likely to be more successful. In the outpatient setting, patients with elevated CK will not be experiencing most of the confounding conditions (such as myocardial infarction or cocaine withdrawal). That is, the majority of the confounders identified in our study population are conditions seen in hospitalized patients who are very sick, not patients cared for in an outpatient setting. We predict that the association between statins and CK+ will be observable in this population because the major confounding causes of elevated CK (infarction) will not be present.

The above examples suggest that with a larger sample and a less biased population the statistics would be more reflective of medical expectation and fewer of these cases would appear. Obtaining a larger and less biased sample is achievable and we are looking forward to apply our approach to such a sample.

From a methodological stand point, several competing alternatives were available for each of the modeling decisions, filtering, measures of association, and statistical test used in this study. Although some justification was provided for each of our choices, additional experimentation with some of the alternatives is required in order to determine the most promising approach.

It should also be noted that evaluation in terms of sensitivity and specificity would have required manually inspecting hundreds if not thousands of records to determine all possible known ADEs that are statistically supported in the records, which is a very impractical task. Nonetheless, in future work we plan to create a "gold standard" sample with which our method can better be evaluated and tuned.

6. CONCLUSION

In this paper, we presented and assessed a method that automatically mines EHRs in order to identify potential ADEs. The method is based on regression techniques, and is designed to address the main challenge of confounding. To the best of our knowledge, this approach has never been applied to ADE detection based on data in EHRs. Our results show that a set of bone fide ADEs were identified, suggesting that the proposed method is a promising approach to the problem. Nonetheless, additional challenges were identified, most of which were related to certain biases in the data, and some to the methodology.

7. ACKNOWLEDGMENTS

This research was supported in part by grants 5R01LM008635, 1R01LM010016, 3R01LM010016-01S1, and 3R01LM010016-02S1 from the National Library of Medicine. The authors thank Lyudmila Shagina for assistance with MedLEE.

8. REFERENCES

1. Bates DW, Spell N, Cullen DJ, Burdick E, Laird N, Petersen LA, Small SD, Sweitzer BJ, Leape LL: The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. *JAMA* 1997, 277:307-311.
2. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP: Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. *JAMA* 1997, 277:301-306.
3. Adverse Event Reporting System. http://www.fda.gov/cder/aers/default.htm
4. The Upsala monitoring centre. http://www.who-umc.org
5. Stephenson W, Hauben M: Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiol Drug Saf* 2007, 16:359-365.
6. Bate A, Evans SJ: Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf* 2009, 18:427-436.
7. Wang X, Hripcsak G, Markatou M, Friedman C: Active Computerized Pharmacovigilance using Natural Language Processing, Statistics, and Electronic Health Records: a Feasibility Study. *J Am Med Inform Assoc* 2009.
8. Friedman C: Discovering Novel Adverse Drug Events Using Natural Language Processing and Mining of the Electronic Health Record. *AIME '09: Proceedings of the 12th Conference on Artificial Intelligence in Medicine* 2009,1-5.
9. Friedman C, Shagina L, Lussier Y, Hripcsak G: Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004, 11:392-402.
10. 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, 15:87-98.
11. Wang X, Friedman C, Chused A, Markatou M, Elhadad N: Automated knowledge acquisition from clinical narrative reports. *AMIA Annu Symp Proc* 2008,783-787.
12. Hauben M, Madigan D, Gerrits CM, Walsh L, van Puijenbroek EP: The role of data mining in pharmacovigilance. *Expert Opin Drug Saf* 2005, 4:929-948.
13. 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, 10:483-486.
14. van Puijenbroek EP, Diemont W, van GK: Application of quantitative signal detection in the Dutch spontaneous

reporting system for adverse drug reactions. *Drug Saf* 2003, 26:293-301.
15. DuMouchel W: Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System. *Am Stat* 1999, 53:177-190.
16. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM: A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol* 1998, 54:315-321.
17. McNamee R: Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine* 2005, 62:500-506.
18. Genkin A., Madigan D, Lewis DD.: Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics* 2007, 49:291-304.
19. Ifrim G., Bakir G., Weikum G.: Fast logistic regression for text categorization with variable-length n-grams. *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* 2008,354-362.
20. Caster O.: Mining the WHO Drug Safety Database Using Lasso Logistic Regression. *UUDM Project Report 2007:16 Graduation thesis for Master of Science (Mathematical Statistics) degree at Uppsala University, Sweden,* 2007.
21. Caster O., Norén GN., Madigan D, Bate A: Large-Scale Regression-Based Pattern Discovery in International Adverse Drug Reaction Surveillance. *Proceedings of the KDD-08 Workshop on Mining Medical Data* 2008.
22. Ramirez E, Carcas AJ, Borobia AM, Lei SH, Pinana E, Fudio S, Frias J: A pharmacovigilance program from laboratory signals for the detection and reporting of serious adverse drug reactions in hospitalized patients. *Clin Pharmacol Ther* 2010, 87:74-86.
23. Jewell NP.: *Statistics for Epidemiology.* Chapman and Hall; 2003.
24. Hosmer DW.: *Applied logistic regression.* Wiley-Interscience; 2000.
25. Hoerl AE., Kennard RW.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 1970, 12:55-67.
26. Tibshirani R.: Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* 1994, 58:267-288.
27. Almenoff J, Tonning JM, Gould AL, Szarfman A, Hauben M, Ouellet-Hellstrom R, Ball R, Hornbuckle K, Walsh L, Yee C et al.: Perspectives on the use of data mining in pharmaco-vigilance. *Drug Saf* 2005, 28:981-1007.

Table 1: Analysis of confounding for known cases using logistic regression

	drug	potential confounder	b-before	b-after	Test 1	Test 2	Test 3
A	clopidogrel	myocardial_infarction	1.9 [1.7-2.2]	1.6 [1.3-1.9]	7.3E-29	1.8E-07	23%
B	atorvastatin	Stemi st_segment_elevation myocardial_infarction chest_pain	1.5 [1.3-1.7]	1.0 [0.7-1.2]	2.3E-14	0.0E+00	43%
C	tramadol	cocaine_abuse drug_abuse drug_abuser opioid_abuse heroin_abuse	1.2 [0.4-2.1]	1.1 [0.2-2.0]	1.5E-02	0.0E+00	10%
D	insulin	myocardial_infarction diabetes_mellitus	0.7 [0.3-1.0]	0.5 [0.2-0.9]	5.0E-03	0.0E+00	15%
E	atorvastatin	hyperlipidemia	1.5 [1.3-1.7]	1.4 [1.2-1.6]	1.1E-33	4.1E-06	9%
F	sirolimus	seizures	1.7 [0.7-2.7]	1.7 [0.8-2.7]	4.9E-04	7.7E-07	-2%

Table 2: Potential confounders identified using shrinkage logistic regression

creatine_phosphokinase_increased	akinesia	combative
nausea_and_vomiting	agitation	hyperlipidemia
myositis	polymyositis	intoxication
stemi	chest_pain	myopathy
st_segment_elevation	compartment_syndromes	altered_mental_status
cocaine_abuse	unconscious_state	thrombus
aspartate_transaminase_levels_raised_(plasma_or_serum)	kidney_failure_acute	
myocardial_infarction	unresponsive_behavior	
absent_pulse	falls	
transaminitis	acute_myocardial_infarction	

Table 3: Potential drugs identified by the proposed method that are related to elevated CK ADEs

bicarbonates	fosphenytoin	vasopressins
carbidopa	lamivudine+zidovudine	
chlorpromazine	olmesartan_medoxomil	
doxazosin	sirolimus	
efavirenz	tramadol	