APPLICATION OF KNOWLEDGE DISCOVERY IN DATABASES METHODOLOGIES FOR PREDICTIVE MODELS FOR PREGNANCY ADVERSE EVENTS

by

Laritza M. Taft

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

August 2010

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Laritza M. Taft**

has been approved by the following supervisory committee members:

R. Scott Evans , Chair **5-20-2010**
Date Approved

Joyce A. Mitchell , Member **5-20-2010**
Date Approved

Sidney N. Thornton , Member **5-18-2010**
Date Approved

Bruce E. Bray , Member **5-20-2010**
Date Approved

Mitchael W. Varner , Member **5-18-2010**
Date Approved

and by **Joyce A. Mitchell** , Chair of
the Department of **Biomedical Informatics**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

In its report *To Err is Human,* The Institute of Medicine recommended the implementation of internal and external voluntary and mandatory automatic reporting systems to increase detection of adverse events. Knowledge Discovery in Databases (KDD) allows the detection of patterns and trends that would be hidden or less detectable if analyzed by conventional methods.

The objective of this study was to examine novel KDD techniques used by other disciplines to create predictive models using healthcare data and validate the results through clinical domain expertise and performance measures.

Patient records for the present study were extracted from the enterprise data warehouse (EDW) from Intermountain Healthcare. Patients with reported adverse events were identified from ICD9 codes. A clinical classification of the ICD9 codes was developed, and the clinical categories were analyzed for risk factors for adverse events including adverse drug events. Pharmacy data were categorized and used for detection of drugs administered in temporal sequence with antidote drugs. Data sampling and data boosting algorithms were used as signal amplification techniques. Decision trees, Naïve Bayes, Canonical Correlation Analysis, and Sequence Analysis were used as machine learning algorithms.

Performance measures of the classification algorithms demonstrated statistically significant improvement after the transformation of the dataset through KDD techniques, data boosting and sampling. Domain expertise was applied to validate clinical significance of the results.

KDD methodologies were applied successfully to a complex clinical dataset. The use of these methodologies was empirically proven effective in healthcare data through statistically significant measures and clinical validation. Although more research is required, we demonstrated the usefulness of KDD methodologies in knowledge extraction from complex clinical data.

To my parents
Orlando and Milena
and my husband
Gary

TABLE OF CONTENTS

Chapter

LIST OF TABLES

Table Page

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my gratitude to my committee, R. Scott Evans, Joyce M. Mitchell, Sidney N. Thornton, Bruce E. Bray, and Michael W. Varner, for their continued support and encouragement. To Chi-Ren Shyu, who was instrumental with his continual presence and research guidance, thank you. To Michael W. Varner for his patience and understanding, a deserved and honored thank you. To John Hurdle who believed in my capability from day one and encouraged me all along, thank you. I would also like to thank The National Library of Medicine, Intermountain Healthcare, and the faculty, staff, and fellow students of the University of Utah Department of Biomedical Informatics.

.

CHAPTER 1

INTRODUCTION

Background

In its report *To Err is Human* [1], the Institute of Medicine (IOM) recommended the implementation of internal and external voluntary and mandatory reporting systems to increase the detection of adverse events (AE). A more recent IOM report, *Preventing Medication Errors* [2] states that most medication errors (ME) occur in operating rooms, emergency departments, and intensive care units. Operating rooms, emergency departments, and intensive care units are known to have a high incidence of Adverse Drug Events (ADE) [3]. Labor and Delivery (L&D) areas are considered by quality assurance groups as special care units, and pregnant women are considered by the Federal Drug Administration (FDA) as a vulnerable group for ADE [2]. L&D provides emergency care and therefore should also be treated as a high-risk area.

The IOM defines "medical errors" as the failure of a planned action to be completed as intended or the use of a wrong plan to achieve an aim [1]. A common type of ME is an adverse drug event (ADE). ADEs are harm caused by use of medications including medication errors and adverse drug reactions. Medication errors are preventable and occur in the medication administration process: prescription, dispensing, administration. Errors can be reduced with the implementation of electronic tools in the medication administration process. Electronic tools allow the detection of duplicate prescriptions and incorrect doses and aid in the detection of drugs that can potentially cause adverse drug events or drug-drug interactions [2]. Our work demonstrates the possibility of utilizing electronic algorithms to detect those drugs with higher probability of causing ADE by identifying associations of the use of such drugs and clinical diagnosis. Likewise, drugs with a higher probability of causing ADE can be identified through the detection of the use of antidote drugs, providing additional means to increase detection and accurate reporting.

Sentinel events for adverse event monitoring are defined by JCACHO and by some state Health Departments [4, 5]. The list of sentinel events from the Utah Department of Health was used in this study for the identification process of adverse events in labor and delivery [6]. The JCACHO defines a severe AE as

> an unexpected occurrence involving death or serious physical or psychological injury, or the risk thereof. Serious injury specifically includes loss of limb or function. The phrase, "or the risk thereof" includes any process variation for which a recurrence would carry a significant chance of a serious adverse outcome. Such events are called "sentinel" because they signal the need for immediate investigation and response [4].

Advances in computer technology have made it possible to store large amounts of data. Examination of these data has the potential to detect adverse events. The commercial industry has used available data to understand the relationships among multiple variables and help manage businesses. The most important successes have been in fraud detection, marketing, and customer retention, where millions of dollars are saved by identification of associated events and the promotion of policies derived from these techniques. Healthcare providers and decision makers are faced with vast quantities of data and need an effective way to extract information from relationships and trends. Although advancements have been made, the complexity of clinical data remains a challenge in current research. KDD has dealt successfully with complex data from other fields, and the methods could potentially be applied in healthcare. To apply models from other disciplines, it is necessary to validate the methodologies with successful metrics of prediction and through the application of clinical domain knowledge [7, 8]. Prediction metrics from different methods are evaluated through the comparison of statistical significance, true positive rates, and positive predictive values that each algorithm has to detect the target outcome. The study dataset is sampled utilizing different methodologies to avoid overfitting. The ultimate purpose is to determine if the application of a machine learning algorithm is capable of increasing the predictive value of the outcome [9].

The Joint Commission (JCACHO) [4] requests sentinel event reporting in an effort to standardize and increase the detection of adverse events. In spite of multiple attempts, the reported incidence has been shown to be underestimated. It is necessary to incorporate automatic detection systems that can help detect the occurrence of adverse events without human intervention [1, 2]. To incorporate changes and improve patient safety, the real incidence of adverse events and risk factors needs to be identified [10]. To help solve these problems, we have examined the use of Knowledge Discovery in Databases (KDD) techniques used by other disciplines such as business analysis, industry, and other scientific research [11, 12].

Fayyad defines KDD as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” [13]. The process is comprised of several steps. Data transformation includes feature selection, dimensionality reduction, normalization, and data subsetting. Data mining is the extraction of trends and patterns from data. It includes the application of descriptive and predictive algorithms. Post processing is the final step in the KDD process in which patterns are filtered, analyzed, and interpreted. “Closing the loop” is the phrase used to describe the integration of data mining results into decision support systems [13, 14].

Data transformation is a preliminary step in the KDD process; the objective is to create variables more suitable for analysis. It involves transformation of continuous variables into categories and creating changes in the distribution through sampling techniques. The objective of creating categorical variables is to achieve dimensionality reduction and to generate datasets that can be handled by computerized classification algorithms and interpreted by domain experts.

Techniques used to create new variables are commonly known as categorization or mapping. The use of age categories instead of continuous numbers is an example of categorization. Categorization techniques also allow the representation of data in graphical form that facilitate the interpretation and understanding of the structure of the data. Techniques involved in dimensionality reduction and data cleansing also include preliminary steps that aid in the understanding of the data. In these preliminary steps, a combination of domain expertise and the use of descriptive statistics is necessary. A clear understanding of the structure of the data, the composition of the variables, and the evaluation of the impact of each variable on the object of study are also part of the data transformation phase. Likewise, the use of statistical procedures such as correlation matrices, principal components, and factor analysis allows dimensionality reduction. These techniques are designed to detect variables that are highly correlated and to select only those with a higher significance in explaining the outcome without loss of information.

Dimensionality reduction techniques used in the present study included correlation matrices, principal components, and factor analysis. A correlation matrix measures the correlation coefficient and indicates the strength and direction of a linear relationship between two random variables. Highly correlated independent variables are to be avoided in predictive models since they will yield falsely inflated results of the prediction metrics. Principal components analysis is a method of transforming the original variables into new uncorrelated variables. The new variables are called the principal components; they are a linear combination of the original variables. The variance and the Eigen values are the measures that evaluate the amount of information of each individual variable. Principal components also allow the selection of the components of the dataset that explain the highest variances. A further step in dimensionality reduction is factor analysis; it provides the explanation of the relationships among the original variables and aids in the extraction of those that convey the essential information contained in the original set of variables. Factor analysis aids in the detection of variables that might have similar effect on the variance and therefore helps in selection of the variables to be used in the model and in multicollinearity reduction [15, 16].

The data transformation step also involves creating changes in the distribution of the data through procedures such as sampling, oversampling, bagging, and boosting. These techniques are methods for reutilizing and reorganizing the data in order to optimize the selection of cases used for prediction. The reutilization of data is perhaps the most difficult step in the KDD process. Inadequate handling of this step can create changes in the data structure that optimize the prediction results but cannot be generalized when testing the models in a real-life setting. Inadequate sampling can create models that yield ideal predictive results but are not feasible to apply. Model overfitting is the optimization that yields optimal classification results but cannot be generalized. KDD studies include a validation step of the performance measures of the classification algorithms in training and testing sets. Training sets are those in which the classification model is created and optimized. Testing sets are those in which the resulting model is evaluated to determine its value [9, 14]. Data sampling techniques are valid as long as it can be proven that they are representative of the distribution of the original datasets.

Bagging, also known as bootstrapping, is the iterative process of utilizing the data on different samples always maintaining the original sample size. In contrast, boosting iteratively uses those samples that are difficult to classify; it is a form of signal amplification. In this study, a boosting technique known as Synthetic Minority Oversampling Technique (SMOTE) was used. Unlike other boosting algorithms, SMOTE creates new outcome synthetic cases by computation of the values of the variables from the record with the outcome variable and its k-nearest neighbors. As a result, the synthetic cases share variable values from both sample cases and the rest of the cases in the dataset. The resulting oversampled cases are not simple duplication of the data but an expression of the variable values of the complete dataset [17].

The second step in the KDD process is the application of classification algorithms, also known as data mining. Machine learning algorithms are techniques used to create the prediction models that allow the selection of variables that have the most significant value in the prediction of an outcome. The most common of these techniques is statistical regression analysis in which a set of predictive variables are studied against an outcome variable. Classification algorithms used in KDD are Decision Trees, Naïve Bayes, Canonical Correlation Analysis, Association Rules, Sequence Analysis, and Regressions, among others.

Described by Hotelling (1936), the Canonical Correlation Analysis (CCA) algorithm applies an extension of multiple regression and correlation analysis. CCA is useful in situations where regression techniques are applicable and the outcome set consists of several related dependent variables. The technique calculates the linear functions of the two sets of variables; the linear combinations of both sets of variables are the explanation of the correlations between the two sets. The evaluation is done by determining the statistically significant correlation (canonical correlation) from the two linear functions and by determining if a reasonable interpretation can be made from the correlations [15].

Decision Trees are predictive models that allow the selection of a variable that will serve as the root node for prediction. The leaves, or branching nodes, are created based on the probability distribution of the chance of occurrence and gain or utility of the root nodes. Decision Trees are inductive learners that have proven to perform well in clinical research. As an added advantage, decision trees can be displayed in graphical form; this facilitates the interpretation of the results by domain experts [9].

Naïve Bayes is a simple probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions. Bayes' theorem is based on the conditional probability theory: "the posterior probability is proportional to the product of the prior probability and likelihood." [18] With the independence assumption, the Naïve Bayes classifier oversimplifies the models, avoids the complexity of producing joint probabilities across features, and reduces the number of variables. Large numbers of features can be overwhelming and difficult to analyze. While the assumption of independence is "naïve," it has been shown to perform exceptionally well in the medical field [9, 19].

The performance measures for evaluation of the Decision Trees and Naïve Bayes are True Positive Rate (TPD), Area under the Curve (AUC), and Kappa Statistics for agreement of classification between the different models.

The association rules are a statement of conditional probability X→Y where X is the antecedent (product 1) and Y the consequence (product 2). The significance of the rules are measured based on the number of times each item appears in the dataset and the number of times the two items appear as an item set. Sequence discovery is an extension of association rules. It incorporates a time variable that makes it possible to determine the temporal association between the antecedent and the consequence [20-22].

The most valuable aspect of KDD techniques in medical applications relates to the possibility of discovering hidden relationships among risk factors and the possibility of designing predictive models and hypothesis-generating theories. Predictive models obtained by KDD allow the possibility to predict outcomes of future events. The prediction rules can be used to develop decision support systems applicable to clinical care and patient counseling [23]. The high complexity of clinical data, especially in the obstetrical setting where both maternal and fetal factors should be considered, is an ideal setting to apply machine learning algorithms. Existing electronic healthcare datasets can provide a wealth of information. The challenge is to apply KDD techniques that provide the transformation of raw data into formats that allow analysis from which useful information can be extracted.

Clinical Domain

Clinical data sets are large, complex, and only in rare occasions stored in coded form. Patients in Labor and Delivery (L&D) have physiologic characteristics unique to the situation. The study of maternal perinatal events includes maternal and fetal factors. Such is the case of perineal laceration, a sentinel event [24, 25]. Reported risk factors for perineal laceration can be maternal such as nutrition, prenatal care, and infectious diseases. Fetal factors such as size and presentation also play an important role in the severity and extension of maternal perineal lacerations [26-28]. Likewise, the pharmacopeia utilized in L&D patients is unique and different from other medical disciplines. Off-label use of medications is frequent in pregnant and laboring patients. Drugs such as steroids are often used for fetal lung maturity induction. Maneuvers such as rapid change of position from supine to decubitus to control maternal hypotension are frequent and yet not considered as a traditional antidote to an adverse event [29, 30] but clearly serve this purpose in L&D. The complexity and uniqueness data of L&D patients provided the opportunity to experiment with KDD methods.

Selection of Sentinel Events

Perinatal sentinel events focus on maternal mortality and morbidity. The most common causes of maternal morbidity are infections, hemorrhage, maternal hypertension, perineal laceration, thromboembolic events, and adverse drug events [31]. Multiple studies on risk factors for perineal laceration have been published; however, the studies are divided into maternal and fetal factors and studied separately. Aside from the clinical importance of identifying risk factors and preventive measures for perineal laceration, our task was to use computerized algorithms that would allow incorporation of both maternal and fetal factors in a multivariate model.

As noted above, Adverse Drug Events (ADE) are underreported, and computerized tools offer an opportunity to improve detection. The application of oversampling techniques to existing data may facilitate the improvement of the prediction of data mining models. Oversampling techniques are used in disciplines outside of medicine to create predictive models for risk detection in sparse datasets. These data sampling techniques have been proven useful in the analysis of fraud detection, oil spill prediction, and web crawling. [11, 32].

ADE studies have used the order of antidote drugs as triggering signals in reporting and prevention applications [33-35]. The pharmacological action of antidote drugs is clinically known. For the most part, antidotes are substances with pharmacological actions not exclusive of antidotal purpose. To determine if a drug was used as an antidote or was used for a different clinical indication, one must determine the sequence in which it was administered in relation to the drug that caused the ADE. There are multiple algorithms that allow association analysis; these algorithms have been used in industry to apply marketing strategies. Sequence analysis includes a time sequence variable; it finds frequent associations among items treating them as item sets in the database. Simultaneously the algorithm identifies the order in which the events happened and determines if the association is also that of an antecedent and a consequence. The application of the algorithm would allow the use of antidotes as signal triggers for detection of those drugs that have stronger associations with ADE. The application of the association algorithms helps to avoid false positives; it identifies only those cases in which the antidote drug was used after the drug causing the ADE and not those cases in which the antidote could have been used for other purposes different from the ADE. Existing reports have studied the administration of Vitamin K to detect Coumadin overdose as well as the subsequent use of antidiarrheals, anthihistamines, epinephrine, and steroids to detect ADE from antibiotics and other substances. System performance is variable and customization for medical specialties is necessary [35]. The detection of ADE could be possible with the application of sequence analysis. Sequence analysis allowed the detection of associations between an antecedent and a consequence. The automatic detection of associations between antidotes and drugs administered immediately prior to the antidote could improve ADE reporting and decrease false positive alerts.

Hypothesis

The application of data analysis and data sampling algorithms developed in disciplines outside of medicine can aid in the development of predictive models to detect adverse events in women admitted for Labor and Delivery.

Objective

The goals of this project were to (1) investigate the potential of novel KDD techniques to develop electronic predictive models for reportable adverse events and (2) demonstrate the applicability of the use of these models to healthcare data.

Methods

Patient records analyzed in this study were extracted from the Enterprise Data Warehouse (EDW) of Intermountain Healthcare in Salt Lake City, Utah. The EDW contains clinical care and coded data for billing and reporting. Data from 154,000 individual patients admitted for L&D during years 2002-2005 were extracted. The dataset contained continuous and categorical variables both from clinical and billing records. The variables selected included demographic characteristics, ICD9 discharge diagnosis, maternal and fetal outcomes, and maternal comorbidities.

The Intermountain Data Warehouse data dictionary was used to gain understanding of the variables and structure of the dataset. Descriptive statistics on demographic variables were performed and the results validated through clinical domain knowledge and literature review. Descriptive statistics were done on maternal age, expected fetal weight, incidence of cesarean section, use of forceps, postpartum hemorrhage, pregnancy-induced hypertension, and maternal mortality. At this point, a manual review of the electronic patient records was done to verify the accuracy of the ICD9 coding system. It soon became apparent that the structure of the ICD9 coding system for clinical diagnosis had limited value unless the ICD9 codes could be transformed into a categorical clinical classification. The categorical clinical classification allowed analysis of the discharge diagnosis by groups. This classification would also avoid a dataset with multiple variable values that was subject to create a combinatorial explosion from which no valuable information could be extracted. Moreover, clinical classification of the ICD9 codes would allow the identification of variables that could be used as risk factors such as surgical interventions and fetal characteristics. Through a manual review of the ICD9 codes, a clinical diagnosis classification was created. Simultaneously, a paper on accuracy of discharge data for clinical diagnosis was published [36]. This publication made it possible for us to validate the ICD9 clinical classification that we had created. The result was a comprehensive table that included all the ICD9 codes and the clinical diagnosis found in the dataset subject of study and in the publication mentioned above. A Structured Query Language (SQL) process was then used to create the categorical variables in the dataset records. Values of 1 for present and 0 for not present for each clinical category were assigned to the individual records. With the clinical classification in place, more descriptive statistical analyses were performed and the inclusion and exclusion criteria were created. Records included were those women who gave birth between 20 and 44 weeks gestation and birth weight between 500 and 4800 grams. Two patients' records with maternal age above 55 were excluded as they were confirmed to be data entry errors. To avoid duplicate maternal records in patients with multifetal pregnancies, the outcome data of the first-born infant was selected for inclusion. After these eliminations, the resulting dataset had 104,000 study cases from the original 154,000.

As stated in the introduction, the selection of the sentinel events to be included for analysis was based on clinical relevance, the possibility of clinical intervention based on the results, and the necessity to develop automated systems to aid in early detection and reporting.

The descriptive statistics revealed a low incidence of perineal laceration and ADE which were the sentinel events of interest. In these preliminary stages, predictive data mining algorithms such as decision trees, regression, and Bayes were applied. The results demonstrated very low classification metrics. It was understood that unless further data manipulation, dimensionality reduction, and data transformation algorithms were used, it would be impossible to extract any meaningful information from the dataset.

Results of the Analysis of Risk Factors for Perineal Laceration

Our purpose was to incorporate maternal and fetal variables to construct a model that allowed simultaneous evaluation for both sets of variables. The original data set had

84 variables. The correlation matrix demonstrated a high multicollinearity of the variables. To reduce the multicollinearity in the dataset, the principal components algorithm was applied. This algorithm allowed the identification of highly correlated variables; from those, the most clinically significant variables were preserved. Even with the application of the principal components algorithm, there was still a significant degree of multicollinearity in the dataset. As suggested by the literature, the data were transformed with Factor Analysis, and the resulting transformed values were used for analysis.[15] The outcome variable, perineal laceration was a categorical variable with four degrees of severity. The intent was to do a simultaneous analysis of the four categories of the outcome variable with concomitant inclusion of maternal and fetal factors as predictive variables. Classification algorithms such as Decision Trees, Bayes, Regression Analysis, and Multinomial regression did not allow the multidimensional model that we had intended. On the other hand, Canonical Correlation Analysis (CAA) allowed a multivariate model. The drawback of CAA was that it had only been used in psychological contexts and, at the time, there was only one report in the literature for its application in healthcare [37]. However, the algorithm was designed to take two sets of variables and find the commonalities among the two sets. A literature review of the algorithm, consultation with the graduate committee, and personal communications with the author of the paper mentioned above encouraged the use of the algorithm.

The primary findings of this study identified risk factors for perineal laceration as the use of obstetrical interventions during delivery (forceps and episiotomy) as well as abnormal fetal positions and antecedent of maternal trauma. Detailed findings are discussed on page 33.

Analysis of Risk Factors for Adverse Drug Events

The descriptive statistical study showed a ratio of ADE in the dataset of 0.348% records. Variable selection algorithms were performed as well as preliminary classification algorithms, decision trees, regression analysis, and Naïve Bayes. The evaluation metrics of these classification algorithms were discouraging. The dataset was clearly imbalanced with a low incidence of the outcome variable. Boosting algorithms such as ADA boosting and oversampling techniques were applied in the attempt to increase the number of cases of the outcome variable. The results of the predictive metrics after boosting and oversampling were extremely low and inconclusive. Further review of the literature revealed a new boosting algorithm (SMOTE) published by Chawla [11]. The SMOTE algorithm had been tested on experimental datasets with promising results. Through personal communication, the author authorized and encouraged the use of the algorithm in our dataset. The classification algorithms were then applied to the boosted datasets. The results demonstrated drastic improvement of the predictive metrics. The study was published in the *Journal of Biomedical Informatics* in the paper entitled "Countering Imbalanced Datasets To Improve Adverse Drug Event Predictive Models In Labor And Delivery" [38].

The primary findings of this study were the identification of maternal external trauma, infection, history of previous cesarean and preterm birth as the main risk factors for ADE. Detailed findings are discussed on page 95.

Sequence Analysis for Detection of ADE

Sequence analysis has been used in marketing and web-mining. It has also been used by other authors with healthcare data for detection of gene sequence associations and in public health surveillance systems [39, 40]. The main purpose of the study was to find means for automatic detection of AE in the clinical setting. As in the previous studies, the main obstacle was the sparse number of outcome cases in the dataset. In order to improve the results, random sampling was used to modify the distribution, this time without oversampling techniques. Reduction of the number of cases in the group of records without ADE through random selection and preserving the ADE cases in each sample was used. Descriptive statistics of the number of drugs used in the ADE and non-ADE group showed that the samples from groups are representative of the complete dataset. A statistically significant different number of drugs and drug categories were found between the two groups, as expected. The results validated the use of the algorithm in this clinical setting by generating significant association rules in patients treated with antidote drugs in the ADE group. Likewise, the results generated no, or a fewer number of association rules in patients treated with antidote drugs in the group with no ADE.

Results

The results of the study on the dataset from patients admitted for L&D with different KDD methodologies show statistically and clinically significant conclusions.

The study entitled “The Use of Data Mining to Identify Risk Factors in Perineal Laceration” identified significant risk factors for different degrees of severity of perineal laceration. The risk factors included use of forceps, episiotomy, fetal occipito-posterior position, trauma, and ADE. These clinical findings had been reported in previous studies. We were able to validate the use of the technique through clinically significant conclusions.

The study entitled “Countering Imbalanced Datasets to Improve Adverse Drug Event Predictive Models in Labor and Delivery” identified external trauma, anomalies of the cervix, genito-urinary infections, chorioamnionitis, and history of previous cesarean and preterm birth as the main risk factors for ADE. The results from the boosted datasets, aside from being statistically significant and demonstrating enhanced performance of the classification algorithms, are also validated from the clinical perspective. Patients with the previously noted diseases are subject to receiving a larger number of drugs more likely to cause ADE.

The study entitled “Sequence Discovery Techniques in the Labor and Delivery Setting” successfully identified association rules between drugs used as antidotes and drugs given to patients with pregnancy complications.

The results of our studies demonstrate improved measures of performance of computerized classification algorithms through the application of KDD methodologies. Also, the models obtained from the classification algorithms were proven to have clinical significance.

Discussion

The experimentation with clinical data and algorithms used in fields outside of the clinical setting might have seemed aimless at first, especially the use of algorithms like CCA, which was published in the early 1900s and had not been applied to clinical data. However, recent publications demonstrate the application of CCA to clinical data. Two independent authors used the algorithm for multivariate analysis, and both authors have opening statements in reference to the importance of the possibility to analyze complex multivariate outcomes [41, 42]. Likewise, recent publications note the utilization of boosting algorithms with clinical data [43-46]. The authors agree on the need of signal amplification to improve the predictive models when the outcome variable is infrequent. Sequence analysis associations have been used in the study of epidemiological data in public health surveillance, in gene sequence analysis, and to evaluate therapeutic outcomes [20, 47]. However, we were unable to find recent publications utilizing the technique with clinical data.

The methodologies used in this dissertation can be applied in different clinical settings and with different purposes. One aspect can be the application of boosting techniques to uncover unknown relationships among risk factors and outcomes in current research for diseases of low incidence. As medical science evolves and advances are made in fields such as genetics and pharmacology, it becomes apparent that not all the causative factors of disease are known. Multivariate algorithms can be of value when analyzing complex structures such as the impact of the genetic component and the environment and developmental variables affecting the health of the individual. An example application of multivariate models is the National Children's Study where the study design focuses on a multitude of variables associated with different aspects of the individual from conception through the development of life [48, 49]. The algorithms can also be applied in real-time settings to develop logic for decision support systems. Logic can be applied either to decrease noise in the alerts by isolating those events that have been deemed relevant through application of sequence analysis. Another clinical application could be to detect events of rare occurrence which are likely to be missed. Examples of alerting on events of rare occurrence are adverse drug reactions of rare incidence and drug-drug interactions in patients in multipharmacopeia in which the complexity of the underlying disease can mask the drug effects. Logic can be developed to detect mandatory reporting events in a process that can be semiautomated to help the clinician fill out the required forms. Developing computer logic that can aid the clinician in detecting potential adverse events, either to prevent them or to foster prompt action can improve patient outcomes and reduce the overall healthcare costs.

Limitations

Limitations of the study include the use of ICD9 codes for identification of AE and clinical diagnosis. The ICD9 coding system is used for billing and healthcare reporting, but ICD9 is broad and lacking the specifics of clinical granularity. Unfortunately, ICD-10 coding will not solve these problems as it is not used for reporting. There are multiple codes for similar diseases, signs, symptoms, and interventions. We attempted to overcome this limitation by creating a clinical categorization that allowed us to group similar clinical events into broad categories. A detailed description of the methodologies can be found in the publication “Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery” [38]. The lack of granularity makes it impossible to discriminate among patients with similar diagnosis but with different scales of severity of the disease. An example of discrimination difficulty is patients with pregnancy-induced hypertension in which different scales of severity are possible in clinical practice but impossible to differentiate with ICD9 codes.. Another example is post-partum hemorrhage; several degrees are included in the same group of ICD9 codes. Likewise, ICD9 coding is done *a posteriori* and is based on the overall patient record, making it is impossible to determine the timing for occurrence of an event within the hospitalization period.

References

[1] IOM. To Err is Human:Building a Safer Health System; 1999.

[2] IOM. Preventing Medication Errors: Institute of Medicine; 2006.

[3] Weingart SN, Mc LWR, Gibberd RW, Harrison B. Epidemiology of medical error. West J Med. 2000 Jun;172(6):390-3.

[4] JCAHO. Joint Commission. [cited 2005; Available from: http://www.jcaho.org/

[5] Utah/Missouri Patient Safety Project (AHRQ H. The National Expert Panel of the ICD9-CM Adverse Event Classification Version 2002 Master List of the Panel Selected ICD9-CM Adverse Event Codes.

[6] Complete Indicator Profile of Adverse Events Related to Hospital Inpatient Care: Office of Health Care Statistics, Center for Health Data, Utah Department of Health, Salt Lake City, UT; 2006 11/08/06.

[7] Breton V, Blanquer I, Hernandez V, Legre Y, Solomonides T. Proposing a roadmap for HealthGrids. Studies in Health Technology and Informatics. 2006;120:319-29.

[8] Goodwin LK, Iannacchione MA, Hammond WE, Crockett P, Maher S, Schlitz K. Data mining methods find demographic predictors of preterm birth. Nursing Research. 2001 Nov-Dec;50(6):340-5.

[9] Witten IH, Frank E. Data mining : practical machine learning tools and techniques. 2nd ed. Amsterdam ; Boston, MA: Morgan Kaufman 2005.

[10] Olsen S, Neale G, Schwab K, Psaila B, Patel T, Chapman EJ, et al. Hospital staff should use more than one method to detect adverse events and potential adverse events: incident reporting, pharmacist surveillance and local real-time record review may all have a place. Qual Saf Health Care. 2007 Feb;16(1):40-4.

[11] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2003 2002:341-78.

[12] Liu F, Wets G. A neural network method for prediction of proteolytic cleavage sites in neuropeptide precursors. Conf Proc IEEE Eng Med Biol Soc. 2005;3:2805-8.

[13] Usama Fayyad GP-S, Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence. 1996 1996;0738(4602):37-54.

[14] Tan P-N, Steinbach M, Kumar V. Introduction to Data Mining. 1st ed. Boston: Pearson Addison Wesley 2006.

[15] Afifi A.A CV. Multivariate Analysis, Canonical Correlation Analysis. *Computer-Aided Multivariate Analysis*. Second ed. New York: Van Nostrand Reinhold 2004:252-70.

[16] Zar JH. Biostatistical analysis. Englewood Cliffs, N. J.,: Prentice-Hall 1974.

[17] Chawla N V LA, Hall LO, Bowyer K.,. SMOTEBoost: Improving Prediction of Minority Class in Boosting. Dubrovnik, Croatia 2004.

[18] Woodworth GG. Biostatistics : a Bayesian introduction. Hoboken, N.J.: Wiley-Interscience 2005.

[19] Shortliffe EH, Cimino JJ. Biomedical informatics : computer applications in health care and biomedicine. 3rd ed. New York, NY: Springer 2006.

[20] Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. J Am Med Inform Assoc. 1998 Jul-Aug;5(4):373-81.

[21] Cerrito PB, SAS Institute. Introduction to data mining using SAS Enterprise Miner. Cary, N.C.: SAS Institute 2006.

[22] Creighton C, Hanash S. Mining gene expression databases for association rules. Bioinformatics (Oxford, England). 2003 Jan;19(1):79-86.

[23] Castellani B, Castellani J. Data mining: qualitative analysis with health informatics data. Qual Health Res. 2003 Sep;13(7):1005-18.

[24] JCAHO. Joint Commission on Accreditation of Healthcare Organizations. [cited 2005; Available from: http://www.jcaho.org/

[25] Utah TUo, Gynecology DoOa. Joint OB/urogyn/gyn research meeting. Salt Lake City Utah 2006.

[26] Behringer FR. Cervical and perineal lacerations; studies on the mechanics of labor in relation to cervix, vulva, and perineum. Obstet Gynecol. 1956 May;7(5):557-61.

[27] Albers LL, Sedler KD, Bedrick EJ, Teaf D, Peralta P. Factors related to genital tract trauma in normal spontaneous vaginal births. Birth. 2006 Jun;33(2):94-100.

[28] Angioli R, Gomez-Marin O, Cantuaria G, O'Sullivan M J. Severe perineal lacerations during vaginal delivery: the University of Miami experience. Am J Obstet Gynecol. 2000 May;182(5):1083-5.

[29] Briggs GG, Wan SR. Drug therapy during labor and delivery, part 2. Am J Health Syst Pharm. 2006 Jun 15;63(12):1131-9.

[30] Briggs GG, Wan SR. Drug therapy during labor and delivery, part 1. Am J Health Syst Pharm. 2006 Jun 1;63(11):1038-47.

[31] Geller SE, Rosenberg D, Cox S, Brown M, Simonson L, Kilpatrick S. A scoring system identified near-miss maternal morbidity during pregnancy. Journal of Clinical Epidemiology. 2004 Jul;57(7):716-20.

[32] Fernandez G. Data mining using SAS applications. Boca Raton: Chapman & Hall/CRC 2003.

[33] Gardner RM, Evans RS. Using computer technology to detect, measure, and prevent adverse drug events. J Am Med Inform Assoc. 2004 Nov-Dec;11(6):535-6.

[34] Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc. 2003 Mar-Apr;10(2):115-28.

[35] Handler SM, Altman RL, Perera S, Hanlon JT, Studenski SA, Bost JE, et al. A systematic review of the performance characteristics of clinical event monitor signals used to detect adverse drug events in the hospital setting. J Am Med Inform Assoc. 2007 Jul-Aug;14(4):451-8.

[36] Yasmeen S, Romano PS, Schembri ME, Keyzer JM, Gilbert WM. Accuracy of obstetric diagnoses and procedures in hospital discharge data. Am J Obstet Gynecol. 2006 Apr;194(4):992-1001.

[37] Razavi AR, Gill H, Stal O, Sundquist M, Thorstenson S, Ahlfeldt H, et al. Exploring cancer register data to find risk factors for recurrence of breast cancer--application of Canonical Correlation Analysis. BMC Med Inform Decis Mak. 2005;5:29.

[38] Taft LM, Evans RS, Shyu CR, Egger MJ, Chawla N, Mitchell JA, et al. Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. Journal of Biomedical Informatics. 2008 Sep 14.

[39] Morgan XC, Ni S, Miranker DP, Iyer VR. Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. BMC Bioinformatics. 2007 Nov 15;8(1):445.

[40] Zhou X, Wong STC. Computational systems bioinformatics : methods and biomedical applications. Hackensack, NJ: World Scientific 2008.

[41] Sefton JM, Hicks-Little CA, Hubbard TJ, Clemens MG, Yengo CM, Koceja DM, et al. Sensorimotor function as a predictor of chronic ankle instability. Clinical Biomechanics (Bristol, Avon). 2009 Apr 3.

[42] Ylipaavalniemi J, Savia E, Malinen S, Hari R, Vigario R, Kaski S. Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli. NeuroImage. 2009 Mar 31.

[43] Frank AM. Predicting Intensity Ranks of Peptide Fragment Ions. Journal of Proteome Research. 2009 Apr 2.

[44] Hewett R, Kijsanayothin P. Tumor classification ranking from microarray data. BMC Genomics. 2008;9 Suppl 2:S21.

[45] Carneiro G, Georgescu B, Good S, Comaniciu D. Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree. IEEE Transactions on Medical Imaging. 2008 Sep;27(9):1342-55.

[46] Schmid M, Hothorn T. Flexible boosting of accelerated failure time models. BMC Bioinformatics. 2008;9:269.

[47] Chen Q, Chen YP. Mining frequent patterns for AMP-activated protein kinase regulation on skeletal muscle. BMC Bioinformatics. 2006;7:394.

[48] Landrigan PJ, Trasande L, Thorpe LE, Gwynn C, Lioy PJ, D'Alton ME, et al. The National Children's Study: a 21-year prospective study of 100,000 American Children. Pediatrics. 2006 Nov;118(5):2173-86.

[49] The National Children's Study Is Finally Underway. Adv Neonatal Care. 2009 Apr;9(2):51-2.

CHAPTER 2

USING CANONICAL CORRELATIONS TO IDENTIFY RISK FACTORS IN PERINEAL LACERATION

Taft LM[1], Evans RS[1], Shyu CR[1, 3], Mitchell JA[1], Thornton SN. [1], Bray BE. [1] , Varner M[2],

[1]Department of Biomedical Informatics
[2]Department of Obstetrics and Gynecology

University of Utah Health Sciences Center
30 North 1900 East
Salt Lake City, Utah 84132

[3]Informatics Institute
University of Missouri Columbia
Columbia, MO 65211

Address Correspondence to:

Laritza Taft MD
Department of Biomedical Informatics
University of Utah Health Sciences Center

Salt Lake City, Utah
PH: 801.581.4080
FAX: 801.581.4297
Email: laritza.taft@hsc.utah.edu

Abstract

Objective

The purpose of this study was to investigate the use of canonical correlation methodology to develop a predictive model for severe perineal laceration.

Study Design

We studied 5857 patients with third- or fourth-degree perineal laceration in a retrospective cohort study. Variable reduction was achieved with the use of factor analysis and principal components. A multivariate canonical correlation model was used to identify maternal and fetal risk factors and four degrees of severity outcomes.

Results

The methodology reduced the predictor variables from 127 initial variables to 5 variables in the final model. Significant risk factors included use of forceps, episiotomy, fetal occipito-posterior position, trauma, and adverse drug events.

Conclusion

Canonical correlation analysis allowed simultaneous identification of maternal and fetal risk factors for severe perineal laceration and can aid in the identification of preventive measures at all levels. Simultaneous analysis of numerous risk factors was done without loss of information. The study demonstrates the use of two multivariate sets of data which in turn allows the development of complex models for analysis and interventions. This is one of the first uses of this multivariate method in the obstetrical clinical domain.

Key Words: Perineal laceration, vaginal delivery, canonical correlation analysis

Introduction

Severe perineal damage in a vaginal delivery can have serious lifelong consequences on a woman's health. Immediate post-partum effects include infection, hemorrhage, severe pain, and prolonged hospital stay. Long-term effects include various degrees of urinary and fecal incontinence, chronic pain, dyspareunia, and genital organ prolapse. The Joint Commission [1] includes third- and fourth-degree perineal lacerations as reportable adverse events for evaluation of quality of care [2].

Published studies have analyzed third- and fourth-degree perineal lacerations as a combined outcome and, for the most part, do not include cervico-vaginal lacerations and vaginal hematoma [3-9]. Cervico-vaginal lacerations and vaginal hematomas can also play a role in pelvic floor damage through neurological compromise of the levator ani muscle and the pudendal nerve [10]. In addition, the dynamics of labor and, thus, any perineal damage, can be affected by fetal size and position [11, 12].

A comprehensive analytical predictive model for perineal laceration that could be used for pelvic floor damage evaluation should include maternal and fetal characteristics that influence labor as predicting factors. It should also include cervico-vaginal lacerations and vaginal hematomas as part of the outcome. Using knowledge discovery in databases (KDD) could help provide this model.

Knowledge discovery in databases, commonly known as data mining, allows analysis of large datasets capable of including hundreds of thousands of cases in combination with large numbers of attributes [13]. The most valuable feature of the technique in medical applications is the possibility of discovering hidden relationships among risk factors and the possibility of designing predictive models that include new hypotheses.

Described by Hotelling in 1936, the Canonical Correlation Analysis (CCA) algorithm applies an extension of multiple regression and correlation analysis. CCA is useful in situations where regression techniques are applicable and the outcome set consists of several related dependent variables.

Disciplines outside medicine, such as marketing, agriculture, and human behavior, have used multivariate outcome algorithms for prediction and hypothesis generation. While multiple reports are published using data mining algorithms and data dimensionality reduction techniques such as principal components and factor analysis, few have applied canonical correlation in the clinical field [14]. Canonical correlation analysis had been used primarily in the analysis of behavioral data. In recent years, studies with clinical data have been published in the literature. Its application has been proven useful in the identification of significant associations of multivariate models of phenotypic and genotypic variables in genomic studies [15]. Likewise, CCA has been demonstrated useful in the analysis of multivariate models of findings in images and clinical findings in areas such as ophthalmology, orthopedics, and neurology [16-18].

The complexity of obstetric clinical data provides an ideal setting to apply machine learning algorithms. Obstetrical data as presented in the present study include both maternal and fetal factors that together play a role in the final outcomes of the mother and the newborn. The possibility of simultaneously analyzing maternal and fetal factors and developing combined models increase the opportunity of developing interventions that are likely to have a higher impact on the outcomes.

Commonly used prediction algorithms concentrate on the relationship between multiple predictor variables and a single outcome variable. The model proposed in this report allows a comprehensive evaluation of maternal and fetal characteristics as risk factors for different degrees of perineal laceration as a multivariate outcome variable.

Methods

Subjects

Data used in this study were extracted from the Intermountain Healthcare Labor and Delivery electronic data warehouse. Intermountain Healthcare is a nonprofit healthcare organization serving Utah and southeastern Idaho [19]. Patient selection criteria included discharge years 2002-2005, live singleton vaginal births, gestational age between 20 and 44 weeks, and birth weight between 500 and 4800 grams. De-identified patient data were extracted through a Virtual Private Network connection using Oracle© client version 9.1. The resulting dataset was exported to MySQL database management system for transformation and data preparation.

Risk Factor Identification

Based on the classification of pregnancy-related comorbid diagnoses published by Yasmeen et al. [20], a controlled medical vocabulary was created. Billing and discharge codes from the International Classification of Diseases Coding System version 9 (ICD9) [21] were transformed into dichotomous variables. The outcomes, risk factors, and comorbidities were identified and used as predictor and outcome variables.

Statistical Procedures

A normal distribution for risk factors and comorbidities was assumed based on the Central Limit Theorem [22-24]. Because multidimensional models such as the maternal-fetal unit are difficult to visualize and analyze with mathematical models, and because multicollinearity among the predictor variables generates unstable canonical correlation coefficients [24], the following three dimensionality reduction techniques were used: Pearson correlation, principal components analysis (PCA), and factor analysis (FA). The resulting model included a linear combination of the original variables. The correlation coefficients, the original variable names, the new variables names, and the factors obtained by the procedure are shown in Table 2.1.

To create the CCA model, third- and fourth-degree perineal laceration, vaginal hematoma, and cervico-vaginal laceration were used as the set of outcome variables. The variables included in the outcome set, severe perineal laceration, where identified utilizing the same methodologies as the risk factors by extracting the ICD9 codes from the clinical classification mentioned above and creating dichotomous variables. The coefficients generated by the FA procedure were stored in a new dataset and used as the predictor set.

The final step for the correlation model was the CCA. Canonical correlation coefficients were interpreted as significant if the value was >= 0.3 or <= -0.3 and p values < .0001 as suggested by other studies [23, 24].

The overall significance of the model was assured by executing the procedures in random samples of the dataset. SAS software Release 9.1 was used for the statistical analysis. Institutional Review Board approval was obtained from both Intermountain

Healthcare and the University of Utah. The study was determined to be exempt from review; it included no human interventions.

Results

There were 104,867 patient records of which 85,426 (82%) met the study criteria for vaginal singleton deliveries.

Table 2.2 shows the number of patients in each group of the set outcomes and the most relevant predictor variables. The incidence of severe perineal laceration was distributed as follows: third-degree perineal laceration (3.61%), fourth-degree perineal laceration (0.97%), cervico-vaginal laceration (2.17%), and vaginal hematoma (0.14%).

The CCA results are shown in Table 2.3. The table shows the first three significant canonical correlations with p values < 0.001 along with the canonical coefficients in both predictor and outcome sets of variables V1-V4 and W1-W4. The cumulative proportion of variation explained in the outcomes was 99.23% for the first three canonical correlations.

The results of the standardized canonical coefficients show that for the first canonical correlation the most important outcomes are third- and fourth-degree perineal laceration (correlation coefficients of 0.81 and 0.51, respectively). The most important predictors were the use of forceps, episiotomy, occipito-posterior fetal head position in respect to the pelvis, and maternal anemia (correlation coefficients of 0.68, 0.41, 0.37 and 0.31, respectively). Due to the nature of the data, it is impossible to discern if the anemia was caused by severe bleeding from perineal laceration, or if it was present before delivery and was a contributing factor to the laceration. The second canonical correlation identified the most relevant outcomes to be cervico-vaginal laceration (coefficient 0.88) and vaginal hematoma (coefficient 0.30). The relevant predictor factors were anemia (coefficient 0.39), use of forceps (coefficient 0.34), episiotomy (-0.67), and shoulder dystocia (-0.31). We interpreted the negative values of episiotomy and shoulder dystocia in this canonical correlation as positive significance because these two variables were highly correlated. The third significant canonical correlation had a significant coefficient only for vaginal hematoma (0.94), and the relevant predictors were external trauma and adverse drug events (0.49), deep venous thrombosis (0.37), endometritis (0.33), and ante-partum hemorrhage (0.30).

Discussion

To the best of our knowledge, studies addressing multivariate linear dependencies in labor and delivery data have not been previously reported. This approach provides a better understanding of the causes of variation in the population and the association with adverse pregnancy events. Further implementation of the methodology could allow concomitant identification of predisposing factors to other obstetric complications. Such is the case of diseases in which both genetic and environmental factors are present but the degree in which each contributes is still unclear.

Clinicians can be reluctant to use electronic databases to conduct clinical studies [20]. Although the use of coded administrative and discharge data is not ideal for clinical studies, it can be useful when appropriate statistical and data validation methods are applied. The approach taken in the present study allowed mathematical confirmation of data integrity and validation of the results. We used five different methods of indirect data validation that showed statistically significant agreement between the medical record and the ICD9 coding system.

The current analysis identified areas where changes in screening and practice can influence outcomes. One such example is determination of fetal position during labor. Friedman's initial descriptions of labor progress have been universally accepted in obstetric practice for the past 50 years, albeit often without rigorous attention to some important details [25-27]. For example, the importance of fetal position has again been recently addressed by Wu et al. [28] and again confirmed by this report.

The use of forceps has been reported by other investigators as a predisposition to perineal trauma, and the current analysis confirms this finding [29-32]. Although epidural anesthesia during the active phase of labor has been reported in some studies to decrease labor progress and increase operative deliveries that, in turn, can increase the risk of perineal trauma, it was not one of the significant predictive factors identified in our model. This suggests that epidural anesthesia could be a confounding factor rather than a direct risk factor.

Previous clinical studies have looked for independent risk factors that contribute to perineal laceration such as bone fractures associated with adverse perineal outcomes [33]. Those studies focused on the combined outcome as severe laceration, an approach that simplifies the statistical analysis and allows the use of statistical regression models [34]. The important contribution of the methodology used in the present study is the ability to analyze at once both maternal and fetal factors and understand how each one is a contributing factor rather than an isolated circumstance.

Additional non-intrapartum maternal risk factors such as obesity, hypertension, infection, mental state, alcohol and drug abuse, and history of tobacco use have not been found to have a significant influence on perineal laceration rates. However, our results conclude that the above-mentioned factors have an important impact in the variation of the dataset and should be considered in the study of other adverse outcomes.

Medical research often relies on hypothesis testing experiments to generate evidence-based conclusions. However, important hypotheses and associations can be identified using algorithms not widely applied in medical research. Such approaches can improve the possibility of finding multivariate cause-effect relationships that can be important in clinical care.

In a recent report, El Kady et al found the association of maternal fractures with adverse perinatal outcomes. The study also found a higher incidence of maternal morbidity, including abruption, deep venous thrombosis, and transfusions in women with trauma [35]. Our results also found an important correlation between maternal non-obstetrical related trauma and adverse drug events and perineal outcome. The question that remains to be answered is whether the increased risk of adverse drug events can be explained by an increase in use of medications in patients with medical complications, such as severe perineal laceration, or if there is a different underlying cause that increases the risk in such patients. An increased incidence of nonobstetrical maternal trauma associated with substance and alcohol abuse in association with motor vehicle accidents in a similar manner as in the nonpregnant population. The association with alcohol or substance abuse can also have consequences on maternal nutrition. Under-nourishment has been reported in the literature as a risk factor for severe perineal laceration. More research is required in order to explain and validate our results and understand the association of severe perineal laceration and maternal trauma.

A limitation of the canonical correlation algorithm is that the maximum number of canonical correlations and the set of canonical variables computed need to be equal to the minimum number of outcome or predictor variables. This creates an important constraint when the number of predictor variables to analyze is large. Due to this constraint, there might be other risk factors that were not identified in the present analysis.

Conclusion

This study confirms that canonical correlation algorithms could be a valuable method to analyze large, complex clinical databases. We can conclude from the results that both maternal and fetal factors are significant for severe perineal laceration. Likewise, the combined analysis allows us to foresee the development of maternal adverse events. The significant factors identified were: use of forceps, episiotomy, fetal occipito-posterior position, trauma, and adverse drug events. Complex medical settings such as the maternal-fetal combination, where the number of variables is large and associations difficult to find, can be approached with the use of these analytical tools. Although only applied to severe perineal lacerations in this report, these techniques can be applied to many other obstetric complications.

Acknowledgements

The present study was conducted with data from Intermountain Health Care. It was supported in part by the grant No. LM007124 from the National Library of Medicine.

Table 2.1: Results of Factor Analysis with the corresponding loadings from each variable. Correlation coefficients >= 0.5 were included in the canonical correlation analysis. The 24 factors in the table explain 90% of the variation in the dataset.

	Variables in Factor Analysis	**New Factor Names used for CCA**	**Correlation Coefficient**
Factor1	Adverse Drug Event	Trauma_ADE	0.88
	Trauma		0.89
Factor 2	Substance Abuse	SA_Mental	0.6
	Mental alteration		0.79
	Alcohol use		0.72
Factor 3	Stillbirth		0.88
Factor 4	Forceps		0.76
Factor 5	Endometritis		0.82
Factor 6	Premature rupture of membranes		0.66
Factor 7	Deep venous thrombosis		0.79
Factor 8	Herpes Infection		0.75
Factor 9	Epidural		0.72
Factor 10	Episiotomy		0.72
Factor 11	Intrauterine growth restriction		0.72
Factor 12	Ante partum Hemorrhage	Anemia	0.65
	Ante partum Anemia		0.68
Factor 13	Cephalopelvic Disproportion		0.74
Factor 14	Hypertension	Hypertension	0.65
	Obesity		0.63
Factor 15	Previous Cesarean		0.69
Factor 16	Cardiovascular Disease		0.62
Factor 17	Pregnancy Induced Hypertension		0.69
Factor 18	Occipito Posterior presentation		0.73
Factor 19	Ante partum Hemorrhage		0.52
Factor 20	Amniotomy		0.62
Factor 21	Seizures		0.45
Factor 22	Polyhydramnios		0.57
Factor 23	Shoulder Dystocia		0.53
Factor 24	Congenital Uterine Anomaly	Uterine Anomaly	0.6
	Uterine Anomaly		0.6

Table 2.2: Incidence of comorbidities, risk factors, and outcomes in the study population.

Variable	Not present	Percentage	Present	Percentage
Outcomes				
3rd Degree Perineal Laceration	82342	96.39	3084	3.61
4th Degree Perineal Laceration	84600	99.03	826	0.97
Cervico Vaginal Laceration	83571	97.83	1855	2.17
Vaginal Hematoma	85308	99.86	118	0.14
Predictors				
Adverse Drug Event	85244	99.79	182	0.21
Amniotomy	81978	95.96	3448	4.04
Anemia	84561	98.99	865	1.01
Ante partum Hemorrhage	85233	99.77	193	0.23
Cardiovascular Disease	84825	99.3	601	0.7
Cephalopelvic disproportion	85308	99.86	118	0.14
Deep Venous Thrombosis	85371	99.94	55	0.06
Drug Abuse	85138	99.66	288	0.34
Elective Induction	65340	76.49	20086	23.51
Endometritis	85342	99.9	84	0.1
Epidural Anesthesia	18865	22.08	66561	77.92
Episiotomy	57986	67.88	27440	32.12
Failed Forceps	85332	99.89	94	0.11
Forceps	83658	97.93	1768	2.07
Herpes Infection	85305	99.86	121	0.14
Hypertension	84630	99.07	796	0.93
IUGR	84429	98.83	997	1.17
Occipito Posterior Presentation	84800	99.27	626	0.73
Pregnancy Induced Hypertension	81355	95.23	4071	4.77
Previous Cesarean	82709	96.82	2717	3.18
PROM	83916	98.23	1510	1.77
Seizure	85256	99.8	170	0.2
Shoulder Dystocia	83796	98.09	1630	1.91
Stillbirth	85192	99.73	234	0.27
Uterine Anomaly	85287	99.84	139	0.16
Venous Complication	84655	99.1	771	0.9

Table 2.3: Canonical Correlation Analysis between risk factors and four degrees of perineal laceration. Significant canonical correlation significant coefficients (< or > +/- 0.3) are highlighted. W1-W4 = outcome set V1-V4= predictor set

Canonical Correlation		**R**	**Eigenvalues (R2)**	**Cumulative Proportion**	**Likelihood Ratio**	**F value**	**df**	**P value**
1	0.232515	0.054063	0.0572	0.9067	0.94040007	55.09	96	<.0001
2	0.066628	0.004439	0.0045	0.9774	0.99414698	7.27	69	<.0001
3	0.030626	0.000938	0.0009	0.9923	0.99857999	2.76	44	<.0001
4	0.021966	0.000483	0.0005	1.0000	0.99951748	1.96	21	0.0053

Risk Factor Variables	**W1**	**W2**	**W3**	**W4**
Amniotomy	-0.1473	-0.0670	-0.0121	**0.5266**
Anemia	**0.3375**	**0.3908**	0.1936	-0.2565
Ante-partum Hemorrhage	-0.0783	0.0074	**0.3058**	0.1513
DVT	-0.0084	0.0250	**0.3710**	0.0177
Endometritis	0.0318	0.1736	**0.3372**	0.1154
Episiotomy	**0.4114**	**-0.6695**	0.1676	-0.1606
Forceps	**0.6800**	**0.3436**	-0.2146	0.1423
Occipito Posterior	**0.3707**	-0.0981	0.0568	**0.5186**
Shoulder Dystocia	0.0956	**-0.3130**	0.0694	0.1037
Trauma ADE	0.0024	0.1934	**0.4939**	0.1094

Outcome Variables	**V1**	**V2**	**V3**	**V4**
3rd Degree Laceration	**0.8100**	-0.1741	-0.0119	**-0.5602**
4th Degree Laceration	**0.5198**	-0.2780	0.0146	**0.8079**
Cervico Vaginal Laceration	0.2907	**0.8853**	-0.3367	0.1381
Vaginal Hematoma	0.0921	**0.3035**	**0.9483**	0.0243

References

[1] JCAHO. Joint Commission. [cited 2005; Available from: http://www.jcaho.org/

[2] Jill Rosenthal MB. Defining Reportable Adverse Events, A Guide for States Tracking Medical Errors; 2003.

[3] Aikins Murphy P, Feinland JB. Perineal outcomes in a home birth setting. Birth. 1998 Dec;25(4):226-34.

[4] Albers LL, Sedler KD, Bedrick EJ, Teaf D, Peralta P. Factors related to genital tract trauma in normal spontaneous vaginal births. Birth. 2006 Jun;33(2):94-100.

[5] Altman D, Ragnar I, Ekstrom A, Tyden T, Olsson SE. Anal sphincter lacerations and upright delivery postures-a risk analysis from a randomized controlled trial. Int Urogynecol J Pelvic Floor Dysfunct. 2006 Apr 25.

[6] Angioli R, Gomez-Marin O, Cantuaria G, O'Sullivan M J. Severe perineal lacerations during vaginal delivery: the University of Miami experience. Am J Obstet Gynecol. 2000 May;182(5):1083-5.

[7] Bex PJ, Hofmeyr GJ. Perineal management during childbirth and subsequent dyspareunia. Clin Exp Obstet Gynecol. 1987;14(2):97-100.

[8] Bodner K, Bodner-Adler B, Wagenbichler P, Kaider A, Leodolter S, Husslein P, et al. Perineal lacerations during spontaneous vaginal delivery. Wien Klin Wochenschr. 2001 Oct 15;113(19):743-6.

[9] Bodner-Adler B, Bodner K, Kimberger O, Wagenbichler P, Kaider A, Husslein P, et al. The effect of epidural analgesia on the occurrence of obstetric lacerations and on the neonatal outcome during spontaneous vaginal delivery. Arch Gynecol Obstet. 2002 Dec;267(2):81-4.

[10] Wallner C, Maas CP, Dabhoiwala NF, Lamers WH, DeRuiter MC. Innervation of the pelvic floor muscles: a reappraisal for the levator ani nerve. Obstet Gynecol. 2006 Sep;108(3 Pt 1):529-34.

[11] Christianson LM, Bovbjerg VE, McDavitt EC, Hullfish KL. Risk factors for perineal injury during delivery. Am J Obstet Gynecol. 2003 Jul;189(1):255-60.

[12] Kalis V, Chaloupka P, Turek J, Rokyta Z. [Delivery and anal incontinence: definition, classification, prevalence and pathophysiology]. Ceska gynekologie / Ceska lekarska spolecnost J Ev. 2003 Jul;68(4):283-93.

[13] Usama Fayyad GP-S, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence. 1996 1996;0738(4602):37-54.

[14] Razavi AR, Gill H, Stal O, Sundquist M, Thorstenson S, Ahlfeldt H, et al. Exploring cancer register data to find risk factors for recurrence of breast cancer--application of Canonical Correlation Analysis. BMC Med Inform Decis Mak. 2005;5:29.

[15] Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Statistical applications in genetics and molecular biology. 2009 Jan;8(1):Article 1.

[16] Ylipaavalniemi J, Savia E, Malinen S, Hari R, Vigario R, Kaski S. Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli. NeuroImage. 2009 Mar 31.

[17] Sefton JM, Hicks-Little CA, Hubbard TJ, Clemens MG, Yengo CM, Koceja DM, et al. Sensorimotor function as a predictor of chronic ankle instability. Clinical biomechanics (Bristol, Avon). 2009 Apr 3.

[18] Ehongo A, de Maertelaer V, Cullus P, Pourjavan S. [Correlation between corneal hysteresis, corneal resistance factor, and ocular pulse amplitude in healthy subjects]. Journal francais d'ophtalmologie. 2008 Dec;31(10):999-1005.

[19] IHC. Intermountain Health Care. 2006 [cited 2006 April 20, 2006]; Intermountain Healthcare is a nonprofit healthcare system serving the healthcare needs of Utah and southeastern Idaho residents.]. Available from: http://intermountainhealthcare.org/xp/public/aboutihc/

[20] Yasmeen S, Romano PS, Schembri ME, Keyzer JM, Gilbert WM. Accuracy of obstetric diagnoses and procedures in hospital discharge data. Am J Obstet Gynecol. 2006 Apr;194(4):992-1001.

[21] ICD9-CM. International Statistical Classification of Diseases and Related Health Problems. Tenth Revision. Geneva, World Health Organization,: World Health Organization.; 1992.

[22] Fernandez G. Data mining using SAS applications. Boca Raton: Chapman & Hall/CRC 2003.

[23] Afifi A.A CV. Multivariate Analysis, Canonical Correlation Analysis. *Computer-Aided Multivariate Analysis*. Second ed. New York: Van Nostrand Reinhold 2004:252-70.

[24] Bulach MW. Canonical Auto and Cross Correlations of Multivariate Time Series [Ph.D., Statistics]: Birkbeck College-University of London 1997.

[25] Friedman EA, Sachtleben MR. Dysfunctional labor. I. Prolonged latent phase in the nullipara. Obstet Gynecol. 1961 Feb;17:135-48.

[26] Friedman EA, Sachtleben MR. Dysfunctional labor. III. Secondary arrest of dilatation in the nullipara. Obstet Gynecol. 1962 May;19:576-91.

[27] Friedman EA, Sachtleben MR. Dysfunctional labor. IV. Combined aberrant dilatation patterns in the nullipara. Obstet Gynecol. 1962 Dec;20:761-73.

[28] Wu JM, Williams KS, Hundley AF, Connolly A, Visco AG. Occiput posterior fetal head position increases the risk of anal sphincter injury in vacuum-assisted deliveries. Am J Obstet Gynecol. 2005 Aug;193(2):525-8; discussion 8-9.

[29] Dandolu V, Chatwani A, Harmanli O, Floro C, Gaughan JP, Hernandez E. Risk factors for obstetrical anal sphincter lacerations. Int Urogynecol J Pelvic Floor Dysfunct. 2005 Jul-Aug;16(4):304-7.

[30] Dandolu V, Gaughan JP, Chatwani AJ, Harmanli O, Mabine B, Hernandez E. Risk of recurrence of anal sphincter lacerations. Obstet Gynecol. 2005 Apr;105(4):831-5.

[31] Kok J, Tan KH, Koh S, Cheng PS, Lim WY, Yew ML, et al. Antenatal use of a novel vaginal birth training device by term primiparous women in Singapore. Singapore Med J. 2004 Jul;45(7):318-23.

[32] Johnson JH, Figueroa R, Garry D, Elimian A, Maulik D. Immediate maternal and neonatal effects of forceps and vacuum-assisted deliveries. Obstet Gynecol. 2004 Mar;103(3):513-8.

[33] Carroll TG, Engelken M, Mosier MC, Nazir N. Epidural analgesia and severe perineal laceration in a community-based obstetric practice. J Am Board Fam Pract. 2003 Jan-Feb;16(1):1-6.

[34] Robinson JN, Norwitz ER, Cohen AP, McElrath TF, Lieberman ES. Epidural analgesia and third- or fourth-degree lacerations in nulliparas. Obstet Gynecol. 1999 Aug;94(2):259-62.

[35] El Kady D, Gilbert WM, Xing G, Smith LH. Association of maternal fractures with adverse perinatal outcomes. Am J Obstet Gynecol. 2006 Sep;195(3):711-6.

CHAPTER 3

COUNTERING IMBALANCED DATASETS TO IMPROVE ADVERSE DRUG EVENT PREDICTIVE MODELS IN LABOR AND DELIVERY

Taft LM[1], Evans RS[1,5], Shyu CR[1, 3], Egger MJ[1], Chawla N[4], Mitchell JA[1],
Thornton SN. [1,5],Bray B.[1], Varner M[2]

[1]Department of Biomedical Informatics
University of Utah, School of Medicine
[2]Department of Obstetrics and Gynecology
University of Utah, School of Medicine

[3]Informatics Institute
University of Missouri Columbia
[4]Department of Computer Science & Engg.
University of Notre Dame

[5]Department of Medical Informatics
Intermountain Healthcare

Address Correspondence to:

Laritza Taft MD
Department of Biomedical Informatics
University of Utah Health Sciences Center
30 North 1900 East
Salt Lake City, Utah 84132
PH: 801.581.4080
FAX: 801.581.4297
Email: laritza.Taft@hsc.utah.edu

This chapter is reprinted with permission from the Journal of Biomedical Informatics Volume 42 , Issue 2 (April 2009) Pages:356-364 Year of Publication: 2009 ISSN:1532-046

Abstract

Background

The IOM report, *Preventing Medication Errors,* emphasizes the overall lack of knowledge of the incidence of Adverse Drug Events (ADE). Operating rooms, emergency departments and intensive care units are known to have a higher incidence of ADE . Labor and Delivery (L&D) is an emergency care unit that could have an increased risk of ADE, where reported rates remain low and under-reporting is suspected. Risk factor identification with electronic pattern recognition techniques could improve ADE detection rates.

Objective

The objective of the present study is to apply Synthetic Minority Over Sampling Technique (SMOTE) as an enhanced sampling method in a sparse dataset to generate prediction models to identify ADE in women admitted for Labor and Delivery based on patient risk factors, and comorbidities.

Results

By creating synthetic cases with the SMOTE algorithm and using a 10-Fold Cross validation technique, we demonstrated improved performance of the Naïve Bayes, and the decision tree algorithms. The true positive rate (TPR) of 0.32 in the raw dataset increased to 0.67 in the 800% over-sampled dataset. CONCLUSION: Enhanced performance from classification algorithms can be attained with the use of synthetic minority class oversampling techniques in sparse clinical datasets. Predictive models created in this manner can be used to develop evidence based ADE monitoring systems.

Word Count: 210

Key Words: Adverse Drug Events, Pregnancy, Labor and Delivery, Oversampling, Data mining

Background

The Institute of Medicine (IOM) in the report, *Preventing Medication Errors* [1] recommended the implementation of decision support tools derived from evidence based knowledge and patient information as part of the strategies to prevent medication errors (ME). The report also recommended the active monitoring of medication use to promote prevention strategies. Although medical research has actively pursued these problems, the reported incidence of ME is suspected to be under-estimated[1-3].

These IOM reports [1,2] define ME as avoidable errors occurring in the medication use process. Adverse drug event (ADE) is a more inclusive definition that covers both ME, and adverse drug reactions.

Operating rooms, emergency departments and intensive care units are known to have a higher incidence of ADE [4]. Labor and Delivery (L&D) areas are considered by quality assurance groups as special care units and pregnant women are considered by the FDA as a vulnerable group for ADE[1]. L&D provides emergency care and therefore should also be treated as a high risk area Studies published in the literature focus on specific drugs, and anesthesiology events. [5-9] To the best of our knowledge there are no published studies of ADE as a general category in pregnant women. Our findings indicate an incidence of 0.34% of ADE in women admitted to L&D. This incidence is surprisingly low in a population that includes at least 10% of high risk pregnancies that require poly-pharmacy[10].

One of the most complex tasks in the design and development of automated decision support tools is evidence based rule generation and knowledge extraction from existing data[11]. The task is even more challenging in those cases where the class label of interest or ADE patients as in this case, has an incidence of 1% or less[12]. Datasets with these characteristics are also known as skewed or imbalanced datasets. The class of interest is relatively rare and there are important trade-offs in the decision between false negatives, and/or false positives. Overall, it is more costly to have a false negative versus a false positive . More so in a medical application where the interest is detecting patients with adverse outcomes that can be prevented. Without loss of generality, we will assume that the larger class or the majority class is the negative class and the class of interest is the minority (smaller) or positive class. We will use these terms interchangeably in the paper. The use of machine learning algorithms in sparse datasets with class imbalance causes suboptimal classification performance as these techniques get overwhelmed by the majority class. Recent work has focused on sampling techniques that counter the problem of class imbalance by either oversampling the minority class or under-sampling the majority class [12-15].

In this paper, we focus on the application of the Synthetic Minority Over sampling Technique (SMOTE). SMOTE works by generating new instances from the existing cases. SMOTE effectively counters the imbalance in data by not only solving the problem of high class skew but also the problem of high sparsity. It works in the "feature space" rather than "data space". The synthetic samples are created by taking each minority class sample and the k nearest neighbors. The synthetic sample shares features of both the chosen minority class sample and one or more of the nearest neighbors. This approach effectively forces the decision region of the minority class to become more general. The synthetic cases will not only increase the data space but will also amplify the features of the minority class without duplicating the original data. SMOTE's effectiveness has been shown in a variety of domains and with a variety of classifiers [15, 16].

The objective of the present study was to apply SMOTE as an enhanced sampling method using a sparse dataset and to identify a prediction model for ADE in women admitted for L&D based on patient risk factors, and comorbidities. We would like to note here that we tried other of oversampling methods like replication and random under-sampling but none of them resulted in improvement. Hence, for clarity of presentation in the paper, we only focus our discussion, and results on using SMOTE.

Description of Data Mining Techniques

Machine learning techniques include both data sampling, and learning algorithms. Over sampling techniques are applied to reuse the available data by dividing the dataset into three or more sets. Once the data sampling step is completed, the classification algorithms are applied to the resulting datasets . Subsequently, the performance of the classifiers is evaluated by comparison of the results in the training, testing, and validation datasets.

SMOTE was used to generate new synthetic cases for this study. The computations for the new synthetic sample variables are based on Euclidian distance for continuous variables, and the Value Distance Metric for the nominal features. The continuous variable values are created by taking the difference in distance between two existing minority class samples and multiplying that difference by a random number between 0 and 1. The resulting number is added to the feature value of the original sample and the result will be the value of that variable in the new synthetic sample. For nominal variables, the variable value is assigned by majority vote of the K nearest neighbors. As a result, the synthetic cases will have attributes with values similar to the existing cases and not just replications as provided with oversampling. The objective is to increase the representation of the minority class in the resulting dataset, and reflect the structure of the original cases. By adding new samples of similar characteristics to the originals the decision region is amplified and there should be improvement of the evaluation measures: true positives and the Area Under the Curve (AUC). The newly created cases are appended to the original dataset in 100% increments. Thus the "second" dataset will have 100% more minority class cases, the third 200% more minority class cases, and so forth. This technique has proven to be useful in improving prediction of sparse datasets by other authors [14].

Classification Algorithms

Naïve Bayes is a simple probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions. Bayes' theorem is based on the conditional probability theory; the posterior probability is proportional to the product of the prior probability, and likelihood. With the independence assumption, the Naïve Bayes classifier over-simplifies the models. It avoids the complexity of producing the joint probabilities across features, which quickly becomes overwhelming by the large number of features. While the assumption of independence is "naïve", it has been shown to perform exceptionally well in classification in the medical field[17, 18]

Decision Trees are predictive models that allow the selection of an attribute that will serve as the root node for prediction. Based on the probability distribution chance of occurrence and gain or utility of the root nodes, the leaf nodes (or branching nodes) are created[17]. Decision Trees are inductive learners that have proven to perform well in clinical research. The interpretation is facilitated for domain knowledge experts by the display in graphical form. C4.5 is a popular decision tree learning algorithm used in a multitude of domains. We used the WEKA[17] (Waikato Environment for Knowledge Analysis) Open Source Software implementation of C4.5, namely JR48, in our experiments.

Naïve Bayes and Decision Trees were chosen as the classification algorithms for the experiments because the results are in a format that facilitates interpretation by domain experts. The graphical representation of the Decision Trees and the simplicity of the Naïve Bayes model are easily understood as opposed to the "black box" that other algorithms such as Neural Networks, and Vector Machines generate [19].

Methods

Subjects

Records for the present study came from the Enterprise Data Warehouse (EDW) of Intermountain Healthcare in Salt Lake City, Utah. The EDW contains clinical care, and coded data for billing and reporting. Data from 135,000 individual patients admitted for L&D during years 2002-2005 were extracted. The variables included demographic characteristics and discharge diagnosis as well as maternal and fetal outcomes, and maternal comorbidities.

Inclusion criteria were post partum women with gestational ages between 20 and 44 weeks and birth weight between 500 and 4800 grams. Two patient's records with maternal age above 55 were excluded as they were confirmed to be data entry errors. In patients with multifetal pregnancies, the outcome data of the first-born infant were selected for inclusion.

Data Preprocessing

A classification methodology for outcomes and comorbidities was created based on the clinical classification of ICD9 codes for labor and delivery published by Yasmeen[20] and on the reportable adverse events criteria published by the Joint Commission, and the Utah Department of Health[21, 22]. In interest of clarity we called these tables "published classifications."

The published classifications included ICD9 codes assigned to obstetrical diagnosis, pregnancy related comorbid diagnoses, procedures, and for sentinel events. For example the diagnosis "diabetes mellitus" includes ICD9 codes: 250.xx, 357.2, 362.0, 648.0x. We created an electronic table called "classifications" with one column that included each one of the diagnosis, procedures and sentinel events and another column with the ICD9 code. The original ICD9 table included the ICD9 code, and the description. We then used SQL queries to join both tables on the ICD9 code and selected both the description from the ICD9 table, and the classification from the published classifications. One by one each row was verified to ensure that the ICD9 description matched the classifications. A column in the ICD9 table was added for class variables of diagnosis, procedures and risk factors to use in our study, e.g., 'ADE', 'Cesarean', 'pregnancy induced hypertension', etc. Once the tables were joined by ICD9 and the verification was made, we updated the class variable column assigning a category to each ICD9 code. Table 3.1 shows the resulting clinical classification and categories and the corresponding ICD9 codes. We found some factors not included in the published classifications, since those were of interest for prediction they were added to the table . The factors added by us were: demographic variables such as maternal age, fetal weight and fetal presentation during labor.

The clinical classification attribute was added to the patient dataset as a dichotomous variable. Those records that had an ICD9 code corresponding to each comorbidity, risk factor or procedure were assigned a value of 1 or 0 if not present.

The above procedure was done in order to ensure the accuracy of the classifications and include other codes that were in use at Intermountain Healthcare and were not in the publications. It also allowed us to assign a diagnosis to each patient and use it for the validation with the patient electronic record.

Data Validation

Despite shortcomings, numerous clinical and informatics researchers have proven the usefulness of ICD9 coding systems for clinical research [23]. Table 3.2 describes the different methodologies used to validate the accuracy of the clinical classification.. The patient electronic records were randomly selected and the validation for diagnosis was done on the clinicians interface of the medical record. Kappa statistic for agreement between the free text diagnosis in the clinical notes and the classification created based on the ICD9 codes was used.

From the pharmacy database we extracted values for number of drugs administered to the patients with ADE, and to those with no-ADE. The mean values for number of drugs for each group and the t-statistic for comparison are also included in Table 3.2. As expected from previous reports in the literature, patients with ADE had a statistical significant higher number of drugs [24].

Comparison of disease incidence in the study population and the population disease incidence reported by the Utah Department of Health were performed. Similar incidences were found in the comparison for pregnancy induced hypertension, gestational diabetes, preterm birth, and fetal weight.

Statistical Procedures

Attribute selection or dimensionality reduction

The original dataset consisted of 84 variables including maternal comorbidities, demographic information, fetal outcomes, and surgical procedures. Principal components analysis (PC), and Chi-Square ranking were used to determine the explained variability in the dataset. The methods were also used for variable selection of highly correlated variables and to avoid multicollinearity[1, 3]. We applied Chi-Square ranking and PC to each of the complete datasets after the SMOTE procedure. This approach allowed the comparison of the variance in each of the original, and resulting datasets. The intent was to verify if SMOTE altered the structure of the data. Variables with high collinearity (Eigenvectors > .5) were dropped in favor of those that preserved more specific information, e.g., puerperal fever vs surgical wound infection. After we ensured that the preserved variables had no collinearity, we selected the variables with Eigenvalues that explained 80% of the variability as advised in the literature[25].

Data Sampling

The ratio of ADE to controls in the dataset was 0.348/100 and clearly qualifies as a highly imbalanced data set. We used 10-fold cross-validation as a vehicle to empirically validate the results. Ten-fold cross-validation divides the data into 10 mutually exclusive subsets, and then combines 9 of those at a time and evaluates the 10th left-out subset. Thus, a classifier is identified on ten different, but overlapping training sets, and evaluated on 10 completely unique testing sets. In preliminary experiments (results not included not included in this study), we applied a popular ensemble technique called AdaBoost that provides random oversampling of the minority class, and random under-sampling of the majority class. None of these resulted in an improvement over the performance of the base classifier. The SMOTE algorithm was applied creating new synthetic cases of the class of interest in 100% increments. The first synthetic dataset had 100% more ADE cases than the original one, the second synthetic dataset had 200% more synthetic cases, and so forth.

The suite of classification algorithms were then applied to the datasets modified by SMOTE boosted datasets using the 10-fold cross validation sampling technique. The decision to use 10-fold cross validation sampling technique was based on the small number of cases with class label of interest (ADE). The literature reports risk of overfitting and therefore introducing bias to the evaluation of the performance of the classification algorithms with this technique. However, the standard evaluation technique in situations where a limited number of cases is available is stratified 10-fold cross validation[17, 26]. Stratified 10-fold cross validation implies averaging the results after invoking the algorithm 10 times 10-fold. In other words, each classification algorithm runs 100 times on each dataset. In our experiments, the Naïve Bayes classifier took 2 hours for one instance of 10-fold and 4.5 hours for the Decision Tree. The total time to run the experiments reported was 136.5 hours. The computational expense for 21 datasets was beyond the capacity of our resources. Based on the literature 10 is the suggested number of folds for the best estimate of errors[17]. Likewise, SMOTE does not alter the original distribution of the data, therefore the problem of overfitting is avoided[27].

Performance Measures

The performance measures for evaluation of the classification algorithms were True Positive Rate (TPD), AUC (Area Under the Curve) and Kappa Statistics for agreement of classification between the different models.

Validity of Results and Clinical Interpretation

As previously noted, the justification for utilizing SMOTE as the data boosting algorithm is to increase the availability of cases with the class label of interest; patients with ADE. We decided not to use oversampling techniques that involve exact data replication and favored SMOTE as an alternative that creates new synthetic cases of the original class label of interest. In order to prove that SMOTE did not change the original data structure, we applied PC to compare the variance of the original dataset and that of synthetic datasets through the comparison of the eigenvalues. Likewise, PC is described as an exploratory technique useful to gain a better understanding of the interrelationships among the data[23].

Domain expertise, in this case clinical interpretation of the results is necessary when applying novel techniques for predictive models [17, 25]. In order to determine if the predictive models generated by our experiments can eventually be used to create electronic applications, the results were clinically analyzed by two of the authors both specialists in obstetrics, and gynecology. The purpose was to determine if the risk factors and comorbidities in the predictive models are likely to be associated with a higher risk of ADE.

The statistical comparison for the performance of the classifiers was done with the results of the three tests in the SAS output of the univariate procedure: Student's t test, Wilcoxon, and signed rank test. Although the t test is the most common one found in the data mining literature for this purpose, there is evidence that nonparametric tests are more reliable when the number of datasets to compare is 30 or less and there is no assumption of normal distribution [28]. The statistical reason in favor of nonparametric tests for this purpose is beyond the scope of the present report. We refer the reader to the paper published by Demsar on *Statistical Comparison of Classifiers over Multiple Data Sets [28]* for this purpose.

Software Packages

MySQL V5.0 Open Source database management system was used for data preparation, and transformation. WEKA Machine Learning Tools version 3.5.5. Open Source system and SAS software Release 9.1 and SAS Enterprise Miner Release 4.3 were used for data analysis and construction of the predictive models.

IRB Approval

Institutional Review Board approval was obtained from both Intermountain Health Care and the University of Utah.

Results

There were 106,480 cases that met the inclusion criteria and 371 ADE were identified based on the clinical classification previously described.

The demographic maternal characteristics as well as fetal outcomes showed no significant variation on ADE as indicated by the Eigenvalues of the PC. Surgical procedures (cesarean section and forceps) had the highest variation. Fifty five independent comorbidities were identified and accounted for explaining 80% of the variation in the dataset and were used in the final model.

Performance Measures

Figures 3.1 and 3.2 show the increments in the number of new synthetic ADE cases obtained after each SMOTE procedure. Each time the algorithm was applied 371 new synthetic cases were added to the original dataset. Figure 3.1 shows the improved performance of the evaluation metrics with the minority class boosted datasets on the J48 Decision Tree. The original dataset showed a TPR of .32 and an AUC of .78. In the first synthetic dataset the TPR increased to .59, and the AUC to .81. A small increment of the evaluation metrics was observed as the number of synthetic cases increased. Figure 3.2 shows the results for the evaluation metrics for the Naïve Bayes classification algorithm. With the initial 100% boosting there was a slight decrease in the AUC and the TPR remained unchanged. However, after 200% boosting there was an immediate improvement of the performance measures. After the initial increment, the performance measures slightly improved until the 900% SMOTE point was reached. There was no further increased performance beyond the 1000% increase of the synthetic cases.

Validity of Results and Clinical Interpretation

An analysis of the structure of the synthetic datasets was done by comparison of the principal components. The principal components of the original dataset and of those including synthetic cases remained the same. There was a nonsignificant variation in the eigenvalues and the percentage of variation explained by each principal components did not vary. Thus, we believe that SMOTE was effectively able to counter the highly sparse nature of the data by increasing the density of points that enabled the classifiers to discriminate between the two classes.

The decision trees in all the models were similar in structure. The first split in the decision tree occurred in patients with external trauma followed by anomalies of the cervix, genito-urinary infections, and chorioamnionitis. The next split occurred at severe pregnancy induced hypertension followed by history of previous cesarean, and preterm birth labor. The main difference in the structure of the decision trees is in the number of leaves and granularity of the divisions for each rule. While a greater granularity in the decision trees is not necessarily a sign of improvement in the prediction model and can be attributed to overfitting, the increased number of leaves in the boosted models facilitates the ability of domain experts to determine if the comorbidities and risk factors found could be associated with patients with ADE. Figure 3.3 shows the difference in structure and decision paths obtained with the decision tree classification algorithm in the raw dataset and the 900% boosted dataset.

Table 3.3 shows the results of the test statistics used for comparison of the performance of the two classifiers on the raw dataset, and the SMOTED datasets. The results indicate a statistical significant difference for the Kappa statistics both with parametric, and nonparametric tests. The p value from the t Statistic for the comparison of the AUC shows a level of significance < 0.0321. However, the sign test, and the ranked signed test indicate a p <.0001. The number of datasets for evaluation was 21 and with a t Statistic within levels of significance we conclude that the evaluation metrics are indeed significantly different as confirmed by the nonparametric tests.

Discussion

The importance of developing automatic detection tools for ADE have been widely emphasized [29]. The current low ADE reporting rate creates unbalanced datasets that are very difficult to analyze and use for automatic rule extraction. Electronic methods used for knowledge extraction are likely to fail as demonstrated by the evaluation of the classifiers in the raw dataset. Alternative data manipulation methodologies are a subject of current research in disciplines outside of medicine where it is also necessary to develop knowledge bases to predict rare occurrences of an event [12]. Sparse data sets that would otherwise be useless can be used to create the starting point of evidence based electronic systems. Predictive models created in this manner can be used to develop evidence based ADE monitoring systems with the potential to increase ADE detection.. Increased detection of patients at risk for ADE can lead to changes in patient care protocols and improve patient safety and quality of care. One role of biomedical informatics is to evaluate these methodologies and determine the usability in the clinical arena [20, 30-33].

The use of ICD9 coded data for clinical research has been controversial. However, multiple research studies have demonstrated its usefulness[14, 27]. In addition Yasmeen et al proved the reliability of reports of disease incidence using such classification. It should be kept in mind that the resulting clinical classification is a general classification of risk factors and comorbidities with the limitations and short comings of a system as nonspecific as ICD9 . Nonetheless, it can be used to create useful predictive models to automatically detect those patients at higher risk for ADE and even as an automatic method to detect disease incidence or study populations for further research.

Obstetric indicators report severe pregnancy induced hypertension, embolism and infection as the three leading causes for severe maternal morbidity, and mortality[34, 35]. Our results show severe hypertension and wound infection as two of the leading factors for variability in the dataset. It is unclear to us why “trauma” appears as the leading factor for variability since the incidence of trauma is extremely low. We can only speculate that it is because these patients are at higher risk for obstetrical complications such as embolism, infections and hemorrhage as reported in the literature[36].

As noted in the introduction, existing methodologies for detection of ADE and AE in general are insufficient, underreporting is suspected at all levels. We believe that the introduction of machine learning methods could have a promising future in this arena if we are able to create predictive models that could deal with clinical factors of low incidence like ADE. Machine learning methods are capable of detecting associations that are not evident when the prevalence is low. Clinical data are numerous, complex, can be confounding and noisy, as a consequence datasets of this nature are likely to be sparse, and difficult to analyze. The introduction of boosting algorithms like SMOTE where the original structure of the data is maintained is promising, and future research is necessary. However, for a real time automatic detection method to be reliable, the clinical data of interest would have to be coded in real time. Existing real time reports of Natural Language processing and detection of antidote drugs for ADE are promising [37, 38].

Study Limitations

In the present study, we found important discordance between the coded data and the text reports in the electronic medical record (Table 3.2). ICD9 coding for billing and reporting is done based on both electronic, and paper records. Therefore higher agreement could be expected if the validation of the ICD9 codes were done including both sources. Nonetheless, our data indicated similar disease incidence when comparing the study population to that of the State of Utah. Likewise, based on the validation study published by Yasmeen[20] we can conclude that the ICD9 coding system is accurate for clinical classification of obstetrical diagnosis.

Another limitation of the ICD9 coding system and more so of the way it is used for billing and reporting, is the impossibility to determine the timing of the comorbidity in relation to the time of delivery, and patient admission. The ICD9 codes are included in the electronic record after patient discharge and account for all the events that accompanied the patient during the hospital stay and are not stratified by date or time. This could be a problem if specific comorbidity analysis is done. We can only conclude that patients with certain comorbidities are prone to ADE but we cannot determine the timing of the appearance of the comorbidity in relation to the maternity admission or the ADE. Also, the nature of the data makes it impossible to differentiate among those patients with preventable, and nonpreventable, ADE. The clinical classification used in the present study could be used to classify patients in general categories of comorbidities, procedures, and to identify risk factors. A classification like this could be useful to identify groups of patients with shared clinical trends. However, a real time monitoring system could not be implemented since the ICD9 codes are not assigned until days after the patient is discharge from the hospital.

The disadvantages of using sampling and classification techniques with all types of datasets are overfitting or overtraining. Oversampling leads to overfitting, while random undersampling does not necessarily provide new information. The data are optimized in such a way that the classifiers have an excellent performance in the training and testing sets but can have poor performance in the validation sets. In this case, the normal distribution of the individual variables is altered. Oversampling techniques often involve making exact copies of the majority class, resulting in overfitting and does not solve the problem of sparse data. It can on the other hand increase the computational expense without improving the performance in the validation sets. Undersampling can discard useful information and therefore decrease classifier performance [16, 17]. The SMOTE algorithm creates synthetic cases based on the values of the variables of the nearest neighbors. This approach maintains the original distribution and therefore the overfitting problem is avoided. In the present study, we were able to verify this t by comparison of the eigenvalues of the principal components in the raw dataset with those that included the synthetic cases.

It could be argued that the improvement of the evaluation throughout the experiment is evident but that it does not show dramatic changes. We demonstrated statistical significant differences with the use of both parametric and nonparametric statistics in the evaluation metrics of both classifiers. The differences of the structure of the decision trees does change and shows additional split areas that can be used in practical applications through identification of patients at higher risk for ADE. These models can be used as a starting point in future research to focus attention on factors that might be shared by the cases present in the models.

Although precise clinical conclusions cannot be drawn from the results of the present study, the decision trees allow clinical validation of the results. The decision tree in the raw dataset has one split at the beginning and does not allow discrimination between different groups of patients that may have similar risk for ADE than others. By displaying the risk factors in this manner, it is impossible to discern if there are groups that could share a similar risk for ADE and not the same diagnosis. On the other hand, the tree resulting from the SMOTED datasets allowed the visualization of different groups at the same level of risk for ADE and that do not share diagnosis (Figure 3.3). The left hand side figure (tree resulting from the raw data) shows trauma, severe pregnancy induced hypertension, wound infection in decreasing levels of importance. The right side of the figure (tree resulting from 900% SMOTED dataset) shows trauma, severe pregnancy induced hypertension and wound infection as parent nodes at the same level. Through this graphical display we can see how patients with different diseases receiving completely different set of medication can share a similar risk for ADE.

Future Studies

The ICD9 classification system used in the present study is general and unspecific for the study of individual diseases. We believe that if a similar methodology to the ones used in this report were to be applied by replacing ICD9 codes with clinical events, signs, symptoms, and data from the actual medical record, there would be more success in developing predictive models that could be used in real time electronic systems. It is also of importance to study the types of drugs associated with ADE in the pregnant population. The pharmacopeia in obstetrics is limited and it is likely that a sparse dataset can be encountered when analyzing drugs likely to cause ADE. Further research is necessary in order to determine which drugs are associated with ADE and also to determine which drug combinations are likely to produce ADE and drug-drug interactions.

In addition, it would be desirable to compare the performance of the classifiers among the subsets selected with additional variable selection techniques as advised by Hall[19].

Conclusions

The use of knowledge extraction techniques in clinical applications with sparse data is prone to failure without further data manipulation. Enhanced performance from classification algorithms can be attained with the use of SMOTE in the clinical setting as demonstrated in this study, and previously reported by other clinical specialties[14]. Models obtained through this methodology can be used as starting points to develop prediction models for future experiments that will ultimately aid in the development of automatic reporting tools.

Acknowledgments

The present study was conducted with data from Intermountain Health Care. It was supported in part by the grant No. LM 007124-11 from the National Library of Medicine.

Table 3.1 Clinical Classification and ICD codes of Selected Comorbidities

Comorbidity	**ICD9_DX_CD**
Abnormal Cervix	1808, 1809, 2331, 2333, 6150, 6160, 6168, 6221, 62211, 62212, 6223, 6224, 6225, 6227, 6228, 65450, 65451, 65453, 65461, 65462, 65463, 75240, 75249, 7950, 79500, 79503, 79504, 79505, 79509, 7951, V1041, V6110
Adverse Drug Event	2454, 2865, 4582, 62210, 6923, 6930, 7955, 9623, 9681, 9750, 979, 98982, 995, 9952, 9958, 99589, 9998, E8506, E8552, E8580, E8582, E8586, E876, E8768, E8789, E8798, E8799, E930, E9300, E9301, E9302, E9303, E9304, E9305, E9306, E9307, E9308, E9309, E931, E9310, E9311, E9312, E9313, E9314, E9315, E9316, E9317, E9318, E9319, E932, E9320, E9321, E9322, E9323, E9324, E9325, E9326, E9327, E9328, E9329, E933, E9330, E9331, E9332, E9333, E9334, E9335, E9338, E9339, E934, E9340, E9341, E9342, E9343, E9344, E9345, E9346, E9347, E9348, E9349, E935, E9351, E9352, E9353, E9354, E9355, E9356, E9357, E9358, E9359, E936, E9360, E9361, E9362, E9363, E9364, E937, E9370, E9371, E9372, E9373, E9374, E9375, E9376, E9378, E9379, E938, E9380, E9381, E9382, E9383, E9384, E9385, E9386, E9387, E9389, E939, E9390, E9391, E9392, E9393, E9394, E9395, E9396, E9397, E9398, E9399, E940, E9400, E9401, E9408, E9409, E941, E9410, E9411, E9412, E9413, E9419, E942, E9420, E9421, E9422, E9423, E9424, E9425, E9426, E9427, E9428, E9429, E943, E9430, E9431, E9432, E9433, E9434, E9435, E9436, E9438, E9439, E944, E9440, E9441, E9442, E9443, E9444, E9445, E9446, E9447, E945, E9450, E9451, E9452, E9453, E9454, E9455, E9456, E9457, E9458, E946, E9460, E9461, E9462, E9463, E9464, E9465, E9466, E9467, E9468, E9469, E947, E9470, E9471, E9472, E9473, E9474, E9478, E9479, E948, E9480, E9481, E9482, E9483, E9484, E9485, E9486, E9488, E9489, E949, E9490, E9491, E9492, E9493, E9494, E9495, E9496, E9497, E9499, E9800
Alcohol Abuse	2948, 30390, 30391, 30393, 30500, 30501, 30502, 30503, V113
Amniotic Infection	65840, 65841, 65843, 65931
Asthma	49302, 49381, 49390, 49392
Breech Presentation	65220, 65221, 65223
Prolonged Labor	66201, 66211, 66221, 66223
Complicated Labor	65983

Table 3.1 Continued

Comorbidity	**ICD9_DX_CD**
Congenital Uterine anomaly	65401, 65403, 7522, 7523
Cardiovascular Disease	3004, 3643, 3940, 3941, 3942, 3949, 3963, 3968, 3969, 3970, 3971, 3979, 39890, 39891, 4101, 4102, 4111, 41411, 416, 4168, 4239, 4240, 4241, 4243, 42490, 4254, 4258, 4260, 42613, 4263, 4264, 4266, 4267, 42682, 4270, 4271, 42731, 42732, 42741, 42742, 42761, 42769, 42781, 42789, 4279, 42831, 42971, 42989, 4299, 5300, 64851, 64853, 64861, 64862, 64863, 66811, 67321, 67322, 67323, 67451, 67452, 7454, 7455, 74602, 7463, 7464, 74687, 74689, 7469, 74710, 7473, 7475, 74762, 74763, 7593, 7603, 785, 7851, 7852, 78551, 79431, 99674, 9971, V151, V422, V433, V4501, V4509, V452
Diabetes	25000, 25001, 25002, 25003, 25010, 25011, 25040, 25041, 25051, 25053, 25060, 25061, 25080, 25081, 25083, 25090, 25091, 25092, 25093, 2535, 36201, 36202, 64801, 64802, 64803, 64881, 64882, 64883, 64884, 79029
Maternal age > 35	65951, 65953, 65961, 65963, V2381, V2382
Failed Induction	65901, 65910, 65911, 66061
Fetal Distress	65571, 65631, 65633, 65970, 65971, 65973, 66321, 76381
Uterine Fibroids	2180, 2181, 2182, 2189, 65411, 65412, 65413
Genito Urinary Infection	1121, 1122, 11289, 1129, 13101, 1319, 541, 5411, 59010, 59080, 5909, 6142, 61610, 61611, 6162, 6164, 6169, 64651, 64661, 64662, 64663, 64701, 64711, 64723, 7810, 7811, 794, 7998, 920, 980, 9950, 9953, 9954, 9955, 9959, 999
Hemorrhage	2851, 2879, 4590, 64193, 66602, 66612, 66614, 66624, 99811
Herpex Infection	5410, 5412, 5419, 549
Hypertension	36211, 4010, 4011, 4019, 40599, 4293, 4372, 64201, 64202, 64203, 64211, 64213, 64221, 64222, 64223, 64271, 64273, 64291, 64292, 64293, 7962
Uterine Inertia	66101, 66103, 66121, 66123
Infection	1103, 1105, 1120, 1123, 1125, 1140, 1190, 1309, 1320, 1330, 1398, 3229, 340, 3570, 3682, 38010, 38013, 3810, 3842, 388, 389, 4109, 4119, 412, 413, 414, 4184, 4189, 419, 431, 460, 4619, 462, 4659, 4660, 46619, 4732, 4733, 4739, 4781, 4822, 48230, 48282, 4829, 4830, 4838, 485, 486, 490, 5400, 5401, 5409, 542, 5551, 56722, 5990, 64731, 64733, 64761, 64763, 64781, 64782, 64791, 64792, 65921, 65923, 67202, 67511, 6868, 6869, 71, 73090, 7806, 7819, 78552, 7907, 7988, 845, 9162, 9181, 958, 9951, 99592, 99662, 998, 9993, V0259, V1200, V1209
Intrauterine Death	65641, 65643, V271
IUGR	65651, 65653
Legal Abortion	63591, 63592

Table 3.1 Continued

Comorbidity	**ICD9_DX_CD**
Macrosomia	65661, 65663, 7660
Abnormal Fetal Presentation	65201, 65203,65231, 65233, 65241, 65243, 65271, 65281, 65283, 65291, 65293, 66001, 66003, 66522, 66961, 7617
Benign Tumor	2141, 2158, 2166, 2168, 2169, 217, 220, 221, 326, 61172
Viral Infection	4809, 4871, 4878, 528, 529, 539, 5449, 5479, 548, 5679, 64762, 7030, 7070, 75, 7799, 7989, 7999, 88
Mental Alteration	29383, 29384, 29389, 29534, 29570, 29590, 29620, 29623, 29626, 29630, 29632, 29633, 29634, 29640, 29650, 29653, 29660, 2967, 29680, 29682, 29689, 29690, 29699, 2979, 2989, 30000, 30001, 30002, 30009, 30015, 30021, 30022, 30029, 3003, 3009, 3010, 30113, 3017, 30183, 3019, 3061, 30651, 3069, 3071, 30750, 30751, 3080, 3082, 3083, 3089, 3090, 30921, 30924, 30928, 3094, 30981, 30989, 311, 31400, 317, 319, 64841, 64842, 64843, 66901, 78050, 78052, 78071, 7830, 7992, E9538, V110, V111, V119, V409, V610, V624, V6284, V6289
Multiple Gestation	64513, 65101, 65103, 65111, 65113, 65121, 65131, 65133, 65141, 65143, 65151, 65171, 65261, 65263, 66231, V272, V273, V274, V275, V276, V277
Obesity	27800, 27801, 64611, 64612, 64613, V8535, V854
Obstructed Labor	3314, 65991, 66011, 66021, 66023, 66091, 66191
Occiput Posterior	66031
Oligohydramious	65801, 65803
Severe Pregnancy Induced Hypertension	64261, 64262, 64263
Placenta Previa	64100, 64101, 64103, 66351, 7620
Polyhydramnios	65701, 65703
Postdates	64511, 64521, 64523
Precipitaded labor	66131, 66133
Pregnancy Induced Hypertension	64231, 64232, 64233, 64234, 64241, 64242, 64243, 64244, 64251, 64252, 64253
Preterm Pregnancy	64400, 64403, 64413, 64420, 64421
Previous Cesarean	65421, 65423
Premature Rupture of Membranes (PROM)	65810, 65811, 65813
Prolonged PROM	65821, 65823, 65831
Shoulder Dystocia	66041
Streptococcal Infection	380, 4100, 4104, V0251
Tobacco Use	3051, V1582
Thyroid Disease	2409, 2410, 2419, 24200, 24221, 24280, 24290, 24291, 243, 2441, 2443, 2448, 2449, 2459, 2462, 2468, 2469, 2749, 64811, 64812, 64813, 7945, V1087

Table 3.1 Continued

Comorbidity	**ICD9_DX_CD**
Maternal Trauma	3543, 72210, 7605, 80500, 80505, 80506, 8054, 8088, 81341, 81504, 81601, 8248, 83100, 83650, 8439, 8449, 84500, 8460, 8470, 8472, 8479, 87341, 87364, 8821, 9051, 9070, 9072, 9075, 9100, 9110, 9130, 9221, 92321, 9243, 9331, 94203, 94213, 94423, 94536, 9478, 94800, 9532, 9571, 95901, 95919, 9925, 99581, E8120, E8121, E8129, E8147, E8160, E8161, E8190, E8191, E8198, E8199, E8490, E8495, E8496, E8497, E8498, E8499, E8809, E8844, E8859, E887, E8888, E8889, E9179, E918, E9248, E9288, E9289, E9290, E9293, E9298, E9299, E9600, E9670, V5417, V714
Uterine Anomaly	2198, 6159, 6212, 6215, 6218, 65431, 65441, 65442, 65443, 66143
Venous Thrombotic Disease	4439, 45341, 4538, 4549, 4550, 4552, 4553, 4554, 4555, 4556, 4557, 4558, 4565, 4568, 45981, 67101, 67102, 67103, 67111, 67112, 67113, 67121, 67122, 67131, 67133, 67142, 67151, 67152, 67181, 67182, 67191, 67192, 67193, V1251, V1252

Table 3.2: Methods for validation of ICD9 codes

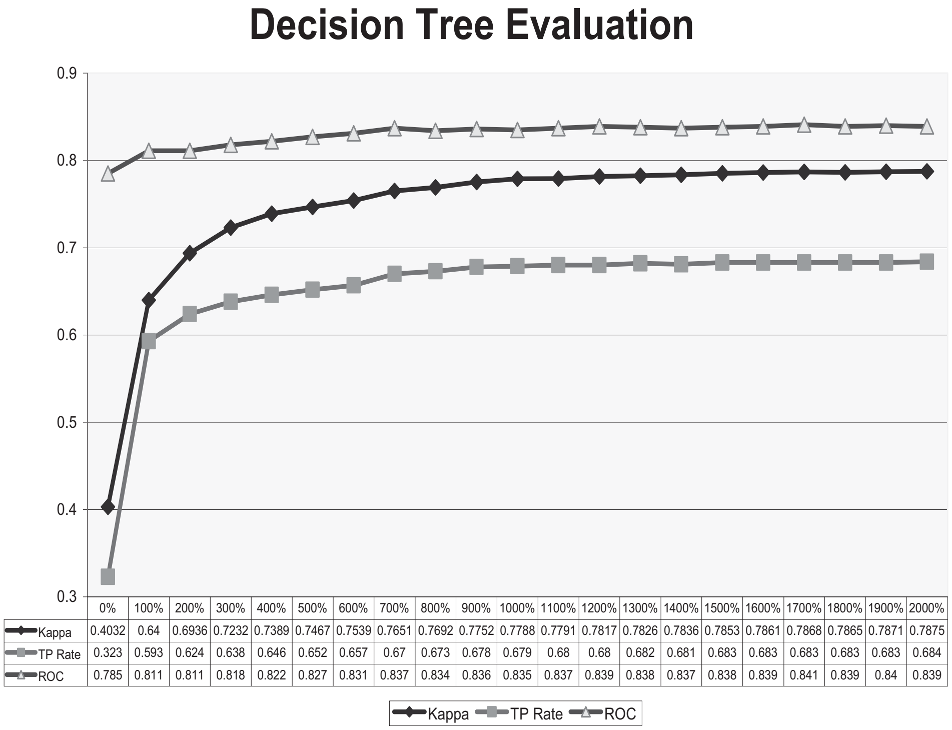
Method	Result
Manual and electronic revision of ICD9 codes included in the clinical classification	All ICD9 codes found in the dataset were included in the clinical classification
Comparison of disease incidence from ICD9 coding system and Utah Department of Health reporting system.	The disease incidence found with both methods was the same
Paired sample t test for comparison of number of drugs used in patients with identified codes for adverse drug events and the control population	ADE group Mean number of drugs used 14 no-ADE Mean 10 P < .001 for number of drugs used in both groups.
Use of Kappa statistic for agreement between ICD9 codes and text from the electronic medical record at the point of care	Kappa statistic: .65-.73 for agreement between free text in the medical record and ICD9 codes
Manual revision of free text notes from the electronic medical record and the ICD9 codes	Kappa statistic: .55-.75 for agreement between free text in the medical record and ICD9 codes for ADE, trauma, hypertension.

Table 3.3 Statistical comparison of Classifiers for Kappa statistic and AUC

Kappa				AUC			
Test	Statistic	Value	P Value	Test	Statistic	Value	p Value
Student's t	T	29.66	<.0001	Student's t	t	-2.3	< 0.0321
Sign	M	10.5	<.0001	Sign	M	-10.5	<.0001
Signed Rank	S	115.5	<.0001	Signed Rank	S	-115.5	<.0001

The results indicate the value for parametric and nonparametric tests. The p value indicates a statistical significant difference between the raw dataset and the SMOTED datasets. The table on the right shows the values for Kappa, the one on the left for AUC. The p values for non-parametric tests show greater significance for the Area Under the Curve Non-parametric tests are statistically safer in samples of 30 or less when the assumption of normal distribution is violated.

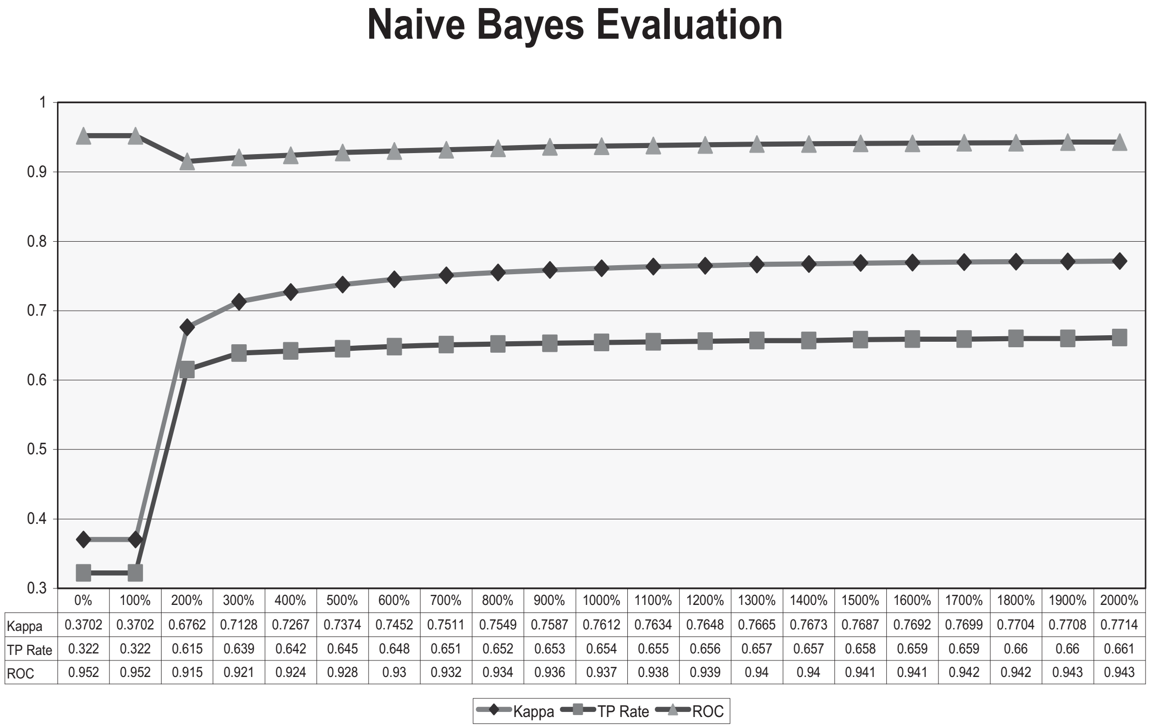
Figure 3.1.: Performance of evaluation metrics in the Decision Tree



	0%	100%	200%	300%	400%	500%	600%	700%	800%	900%	1000%	1100%	1200%	1300%	1400%	1500%	1600%	1700%	1800%	1900%	2000%
Kappa	0.4032	0.64	0.6936	0.7232	0.7389	0.7467	0.7539	0.7651	0.7692	0.7752	0.7788	0.7791	0.7817	0.7826	0.7836	0.7853	0.7861	0.7868	0.7865	0.7871	0.7875
TP Rate	0.323	0.593	0.624	0.638	0.646	0.652	0.657	0.67	0.673	0.678	0.679	0.68	0.68	0.682	0.681	0.683	0.683	0.683	0.683	0.683	0.684
ROC	0.785	0.811	0.811	0.818	0.822	0.827	0.831	0.837	0.834	0.836	0.835	0.837	0.839	0.838	0.837	0.838	0.839	0.841	0.839	0.84	0.839

True positive rate (TP rate) and value of the Receivers Operating Curve (ROC). The Kappa statistic shows the level of agreement for the 10-Fold Cross validation method.

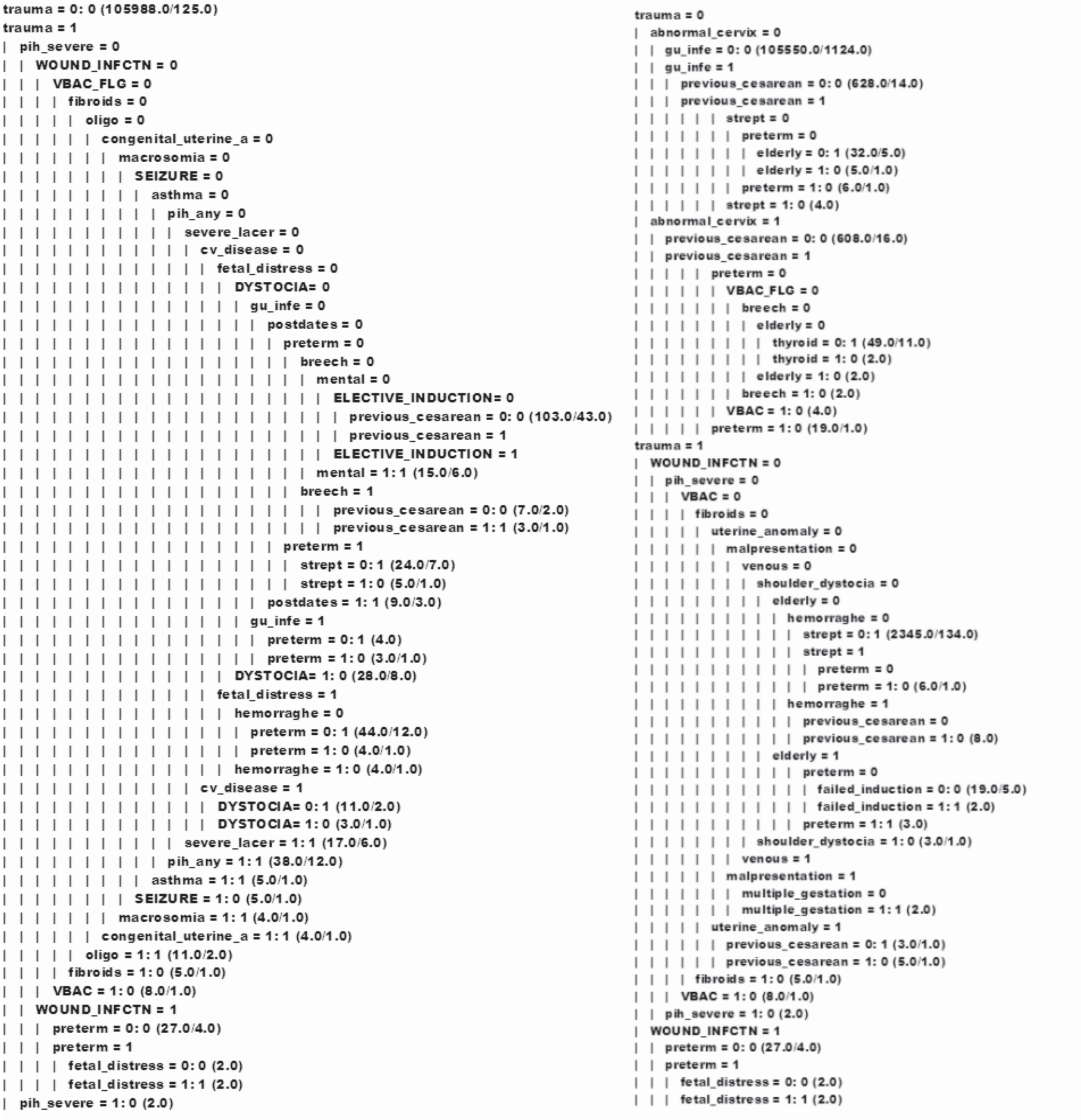
Figure 3.2: Performance of evaluation metrics in the Naive Bayes classification algorithm



	0%	100%	200%	300%	400%	500%	600%	700%	800%	900%	1000%	1100%	1200%	1300%	1400%	1500%	1600%	1700%	1800%	1900%	2000%
Kappa	0.3702	0.3702	0.6762	0.7128	0.7267	0.7374	0.7452	0.7511	0.7549	0.7587	0.7612	0.7634	0.7648	0.7665	0.7673	0.7687	0.7692	0.7699	0.7704	0.7708	0.7714
TP Rate	0.322	0.322	0.615	0.639	0.642	0.645	0.648	0.651	0.652	0.653	0.654	0.655	0.656	0.657	0.657	0.658	0.659	0.659	0.66	0.66	0.661
ROC	0.952	0.952	0.915	0.921	0.924	0.928	0.93	0.932	0.934	0.936	0.937	0.938	0.939	0.94	0.94	0.941	0.941	0.942	0.942	0.943	0.943

True positive rate (TP rate) and value of the Receivers Operating Curve (ROC). The Kappa statistic shows the level of agreement for the 10-Fold Cross validation method.

Figure 3.3 Comparison of the Structure of the Decision Trees before and after the SMOTE process



```
trauma = 0: 0 (105988.0/125.0)
trauma = 1
|  pih_severe = 0
| |  WOUND_INFCTN = 0
| | |  VBAC_FLG = 0
| | | |  fibroids = 0
| | | | |  oligo = 0
| | | | | |  congenital_uterine_a = 0
| | | | | | |  macrosomia = 0
| | | | | | | |  SEIZURE = 0
| | | | | | | | |  asthma = 0
| | | | | | | | | |  pih_any = 0
| | | | | | | | | | |  severe_lacer = 0
| | | | | | | | | | | |  cv_disease = 0
| | | | | | | | | | | | |  fetal_distress = 0
| | | | | | | | | | | | | |  DYSTOCIA= 0
| | | | | | | | | | | | | | |  gu_infe = 0
| | | | | | | | | | | | | | | |  postdates = 0
| | | | | | | | | | | | | | | | |  preterm = 0
| | | | | | | | | | | | | | | | | |  breech = 0
| | | | | | | | | | | | | | | | | | |  mental = 0
| | | | | | | | | | | | | | | | | | | |  ELECTIVE_INDUCTION= 0
| | | | | | | | | | | | | | | | | | | | |  previous_cesarean = 0: 0 (103.0/43.0)
| | | | | | | | | | | | | | | | | | | | |  previous_cesarean = 1
| | | | | | | | | | | | | | | | | | | |  ELECTIVE_INDUCTION = 1
| | | | | | | | | | | | | | | | | | |  mental = 1: 1 (15.0/6.0)
| | | | | | | | | | | | | | | | | |  breech = 1
| | | | | | | | | | | | | | | | | | | |  previous_cesarean = 0: 0 (7.0/2.0)
| | | | | | | | | | | | | | | | | | | |  previous_cesarean = 1: 1 (3.0/1.0)
| | | | | | | | | | | | | | | | |  preterm = 1
| | | | | | | | | | | | | | | | | |  strept = 0: 1 (24.0/7.0)
| | | | | | | | | | | | | | | | | |  strept = 1: 0 (5.0/1.0)
| | | | | | | | | | | | | | | |  postdates = 1: 1 (9.0/3.0)
| | | | | | | | | | | | | | |  gu_infe = 1
| | | | | | | | | | | | | | | |  preterm = 0: 1 (4.0)
| | | | | | | | | | | | | | | |  preterm = 1: 0 (3.0/1.0)
| | | | | | | | | | | | | |  DYSTOCIA= 1: 0 (28.0/8.0)
| | | | | | | | | | | | |  fetal_distress = 1
| | | | | | | | | | | | | |  hemorraghe = 0
| | | | | | | | | | | | | | |  preterm = 0: 1 (44.0/12.0)
| | | | | | | | | | | | | | |  preterm = 1: 0 (4.0/1.0)
| | | | | | | | | | | | | |  hemorraghe = 1: 0 (4.0/1.0)
| | | | | | | | | | | |  cv_disease = 1
| | | | | | | | | | | | |  DYSTOCIA= 0: 1 (11.0/2.0)
| | | | | | | | | | | | |  DYSTOCIA= 1: 0 (3.0/1.0)
| | | | | | | | | | |  severe_lacer = 1: 1 (17.0/6.0)
| | | | | | | | | |  pih_any = 1: 1 (38.0/12.0)
| | | | | | | | |  asthma = 1: 1 (5.0/1.0)
| | | | | | | |  SEIZURE = 1: 0 (5.0/1.0)
| | | | | | |  macrosomia = 1: 1 (4.0/1.0)
| | | | | |  congenital_uterine_a = 1: 1 (4.0/1.0)
| | | | |  oligo = 1: 1 (11.0/2.0)
| | | |  fibroids = 1: 0 (5.0/1.0)
| | |  VBAC = 1: 0 (8.0/1.0)
| |  WOUND_INFCTN = 1
| | |  preterm = 0: 0 (27.0/4.0)
| | |  preterm = 1
| | | |  fetal_distress = 0: 0 (2.0)
| | | |  fetal_distress = 1: 1 (2.0)
|  pih_severe = 1: 0 (2.0)
```

```
trauma = 0
|  abnormal_cervix = 0
| |  gu_infe = 0: 0 (105550.0/1124.0)
| |  gu_infe = 1
| | |  previous_cesarean = 0: 0 (628.0/14.0)
| | |  previous_cesarean = 1
| | | | | |  strept = 0
| | | | | | |  preterm = 0
| | | | | | | |  elderly = 0: 1 (32.0/5.0)
| | | | | | | |  elderly = 1: 0 (5.0/1.0)
| | | | | | |  preterm = 1: 0 (6.0/1.0)
| | | | | |  strept = 1: 0 (4.0)
|  abnormal_cervix = 1
| |  previous_cesarean = 0: 0 (608.0/16.0)
| |  previous_cesarean = 1
| | | | |  preterm = 0
| | | | | |  VBAC_FLG = 0
| | | | | | |  breech = 0
| | | | | | | |  elderly = 0
| | | | | | | | |  thyroid = 0: 1 (49.0/11.0)
| | | | | | | | |  thyroid = 1: 0 (2.0)
| | | | | | | |  elderly = 1: 0 (2.0)
| | | | | | |  breech = 1: 0 (2.0)
| | | | | |  VBAC = 1: 0 (4.0)
| | | | |  preterm = 1: 0 (19.0/1.0)
trauma = 1
|  WOUND_INFCTN = 0
| |  pih_severe = 0
| | |  VBAC = 0
| | | |  fibroids = 0
| | | | |  uterine_anomaly = 0
| | | | | |  malpresentation = 0
| | | | | | |  venous = 0
| | | | | | | |  shoulder_dystocia = 0
| | | | | | | | |  elderly = 0
| | | | | | | | | |  hemorraghe = 0
| | | | | | | | | | |  strept = 0: 1 (2345.0/134.0)
| | | | | | | | | | |  strept = 1
| | | | | | | | | | | |  preterm = 0
| | | | | | | | | | | |  preterm = 1: 0 (6.0/1.0)
| | | | | | | | | |  hemorraghe = 1
| | | | | | | | | | |  previous_cesarean = 0
| | | | | | | | | | |  previous_cesarean = 1: 0 (8.0)
| | | | | | | | |  elderly = 1
| | | | | | | | | | |  preterm = 0
| | | | | | | | | | | |  failed_induction = 0: 0 (19.0/5.0)
| | | | | | | | | | | |  failed_induction = 1: 1 (2.0)
| | | | | | | | | | |  preterm = 1: 1 (3.0)
| | | | | | | |  shoulder_dystocia = 1: 0 (3.0/1.0)
| | | | | | |  venous = 1
| | | | | |  malpresentation = 1
| | | | | | |  multiple_gestation = 0
| | | | | | |  multiple_gestation = 1: 1 (2.0)
| | | | |  uterine_anomaly = 1
| | | | | |  previous_cesarean = 0: 1 (3.0/1.0)
| | | | | |  previous_cesarean = 1: 0 (5.0/1.0)
| | | |  fibroids = 1: 0 (5.0/1.0)
| | |  VBAC = 1: 0 (8.0/1.0)
| |  pih_severe = 1: 0 (2.0)
|  WOUND_INFCTN = 1
| |  preterm = 0: 0 (27.0/4.0)
| |  preterm = 1
| | |  fetal_distress = 0: 0 (2.0)
| | |  fetal_distress = 1: 1 (2.0)
```

The figure on the left hand side shows the decision tree structure on the raw dataset. The figure on the right hand side shows the additional branches in the decision process in the datasets with 900% more synthetic cases. The decision tree on the raw dataset shows a unique decision branch allowing no discrimination for other risk factor variables. The introduction of synthetic cases increases the granularity of the tree and allows the identification of other risk factors and comorbidities that might be equally associated with ADE.

References

[1] IOM. Preventing Medication Errors: Institute of Medicine; 2006.

[2] Rothschild JM, Landrigan CP, Cronin JW, Kaushal R, Lockley SW, Burdick E, et al. The Critical Care Safety Study: The incidence and nature of adverse events and serious medical errors in intensive care. Critical care medicine. 2005 Aug;33(8):1694-700.

[3] IOM. To Err is Human: Building a Safer Health System; 1999.

[4] Weingart SN, Mc LWR, Gibberd RW, Harrison B. Epidemiology of medical error. West J Med. 2000 Jun;172(6):390-3.

[5] Tsai PS, Chen CP, Tsai MS. Perioperative vasovagal syncope with focus on obstetric anesthesia. Taiwan J Obstet Gynecol. 2006 Sep;45(3):208-14.

[6] Cesario SK. Managing the second stage of labor: using evidence to guide practice. Worldviews Evid Based Nurs. 2004;1(4):230.

[7] Shimo T, Nishiike S, Masuoka M, Seki S, Tsuchida H. [Intraoperative anaphylactic shock induced by methylergometrine and oxytocin]. Masui. 2006 Apr;55(4):447-50.

[8] Gaiser RR, McHugh M, Cheek TG, Gutsche BB. Predicting prolonged fetal heart rate deceleration following intrathecal fentanyl/bupivacaine. Int J Obstet Anesth. 2005 Jul;14(3):208-11.

[9] Bolukbasi D, Sener EB, Sarihasan B, Kocamanoglu S, Tur A. Comparison of maternal and neonatal outcomes with epidural bupivacaine plus fentanyl and ropivacaine plus fentanyl for labor analgesia. Int J Obstet Anesth. 2005 Oct;14(4):288-93.

[10] Caughey AB, Bishop JT. Maternal complications of pregnancy increase beyond 40 weeks of gestation in low-risk women. J Perinatol. 2006 Jul 13.

[11] Zorman M, Podgorelec V, Kokol P, Peterson M, Sprogar M, Ojstersek M. Finding the right decision tree's induction strategy for a hard real world problem. International journal of medical informatics. 2001 Sep;63(1-2):109-21.

[12] Weiss G M . Mining with rarity: a unifying framework. SIGKDD Explor Newsl. 2004;6(1):7-19.

[13] Chawla N V LA, Hall LO, Bowyer K,. SMOTEBoost: Improving Prediction of Minority Class in Boosting. 7th European Conference of Priciples and Practice of Knowledge Discovery in Databases (PKDD); 2003; Dubrovnik, Croatia; 2003. p. 107-19.

[14] Liu F, Wets G. A neural network method for prediction of proteolytic cleavage sites in neuropeptide precursors. Conf Proc IEEE Eng Med Biol Soc. 2005;3:2805-8.

[15] Chawla NV BKW, Hall LO ,Kegelmeyer WP, SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2003 2002:341-78.

[16] Chawla N V LA, Hall LO, Bowyer K,. SMOTEBoost: Improving Prediction of Minority Class in Boosting. Dubrovnik, Croatia 2004.

[17] Witten IH, Frank E, Data mining : practical machine learning tools and techniques. 2nd ed. Amsterdam ; Boston, MA: Morgan Kaufman 2005.

[18] Shortliffe EH, Cimino JJ, Biomedical informatics : computer applications in health care and biomedicine. 3rd ed. New York, NY: Springer 2006.

[19] Hall MA HG. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transactions on Knowledge and Data Engineering. 2003 November/December 2003;15:1-16.

[20] Yasmeen S, Romano PS, Schembri ME, Keyzer JM, Gilbert WM. Accuracy of obstetric diagnoses and procedures in hospital discharge data. Am J Obstet Gynecol. 2006 Apr;194(4):992-1001.

[21] JCAHO. Joint Commission on Accreditation of Healthcare Organizations. [cited 2005; Available from: http://www.jcaho.org/

[22] Utah TUo, Gynecology DoOa. Joint OB/urogyn/gyn research meeting. Salt Lake City Utah 2006.

[23] Afifi A.A CV. Multivariate Analysis, Canonical Correlation Analysis. *Computer-Aided Multivariate Analysis*. Second ed. New York: Van Nostrand Reinhold 2004:252-70.

[24] Evans RS, Pestotnik SL, Classen DC, Burke JP. Evaluation of a computer-assisted antibiotic-dose monitor. Ann Pharmacother. 1999 Oct;33(10):1026-31.

[25] Fernandez G. Data mining using SAS applications. Boca Raton: Chapman & Hall/CRC 2003.

[26] Tan P-N, Steinbach M, Kumar V. Introduction to data mining. 1st ed. Boston: Pearson Addison Wesley 2006.

[27] Nitesh V. Chawla AL, Lawrence O. Hall and Kevin W. Bowyer. SMOTEBoost: Improving Prediction of the Minority Class in Boosting 2003.

[28] Demsar J. Statistical Comparisons of Classifiers Over Multiple Data Sets. Journal of Machine Learning Research 2006;7:1-30.

[29] Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc. 2003 Mar-Apr;10(2):115-28.

[30] Cannon-Albright LA, Farnham JM, Thomas A, Camp NJ. Identification and study of Utah pseudo-isolate populations-prospects for gene identification. American journal of medical genetics. 2005 Sep 1;137(3):269-75.

[31] Romano PS, Yasmeen S, Schembri ME, Keyzer JM, Gilbert WM. Coding of perineal lacerations and other complications of obstetric care in hospital discharge data. Obstet Gynecol. 2005 Oct;106(4):717-25.

[32] Geller SE, Rosenberg D, Cox S, Brown M, Simonson L, Kilpatrick S. A scoring system identified near-miss maternal morbidity during pregnancy. Journal of clinical epidemiology. 2004 Jul;57(7):716-20.

[33] Allen-Brady K, Camp NJ, Ward JH, Cannon-Albright LA. Lobular breast cancer: excess familiality observed in the Utah Population Database. International journal of cancer. 2005 Nov 20;117(4):655-61.

[34] Geller SE, Cox SM, Kilpatrick SJ. A descriptive model of preventability in maternal morbidity and mortality. J Perinatol. 2006 Feb;26(2):79-84.

[35] Geller SE, Rosenberg D, Cox SM, Kilpatrick S. Defining a conceptual framework for near-miss maternal morbidity. Journal of the American Medical Women's Association (1972). 2002 Summer;57(3):135-9.

[36] Holden DA, Quin M, Holden DP. Clinical risk management in obstetrics. Curr Opin Obstet Gynecol. 2004 Apr;16(2):137-42.

[37] Evans RS, Pestotnik SL, Classen DC, Bass SB, Menlove RL, Gardner RM, et al. Development of a computerized adverse drug event monitor. Proc Annu Symp Comput Appl Med Care. 1991:23-7.

[38] Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. Jama. 1995 Jul 5;274(1):29-34.

CHAPTER 4

SEQUENCE DISCOVERY TECHNIQUES IN THE LABOR AND DELIVERY SETTING

Taft LM[1], Evans RS[1], Shyu CR[1, 3], Mitchell JA[1], Thornton SN. [1], Bray BE. [1] , Varner M[2],

[1]Department of Biomedical Informatics
[2]Department of Obstetrics and Gynecology

University of Utah Health Sciences Center
30 North 1900 East
Salt Lake City, Utah 84132

[3]Informatics Institute

University of Missouri Columbia
Columbia, MO 65211

Address Correspondence to:

Laritza Taft MD
Department of Biomedical Informatics
University of Utah Health Sciences Center

Salt Lake City, Utah
PH: 801.581.4080
FAX: 801.581.4297
Email: laritza.Taft@hsc.utah.edu

Abstract

Background

Sequence discovery techniques have been utilized to find significant temporal association patterns. This technology has not been applied in the Labor and Delivery (L&D) setting to detect adverse drug events (ADE).

Objective

We assessed the applicability of sequence discovery techniques to clinical healthcare data. The experiments were conducted with records of women admitted to Labor and Delivery with discharge diagnoses of ADE.

Methods

Patient records for this study were extracted from the enterprise data warehouse from Intermountain Healthcare in Utah. Patients with reported ADE and no ADE were identified through ICD9 codes from billing and reporting data. Drug administration data for all patients were extracted from electronic pharmacy records. Clinical domain knowledge was applied to create condensed drug categories and to identify antidotes. SAS Enterprise Miner was used to generate sequence analysis and to generate electronic association rules.

Results

The average number of drugs received by patients with identified ADE was 14 and the average number of drugs received by control patients was 10 ($p < .001$). The number of different drug categories found in the ADE group was 123 compared to 93 in

the non-ADE (p < .001). As expected, a significant difference was found in the use of antidote drugs (antihistamines and ephedrine) between patients with and without ADE.

Conclusion

Episode mining techniques can be used to create electronic rules to detect infrequent healthcare events such as ADE.

Word Count: 223

Key Words: Adverse Drug Events, Pregnancy, Temporal, Episode, Data mining

Adverse Event (AE) reporting has been a predominant focus point for quality assurance and patient safety. As a response to the IOM report *To Err is Human* [1] and through the recommendations of the Quality Interagency Coordination Task Force (QuIC), a series of strategies have been recommended [2]. In addition to other requirements, voluntary and mandatory adverse event reporting using the sentinel event policy is required by the Joint Commission [3]. The Institute of Medicine (IOM) in the report, *Preventing Medication Errors* [4], reviewed existing processes and emphasized the low reported rate of Adverse Drug Events (ADE). The report recommends the need to develop and install automatic detecting systems to improve the detection rate of ADE.

The greatest obstacle in the development of an ADE electronic reporting system is the availability of data in an electronic and coded form. For this reason, computerized ADE monitoring systems have focused on the use of administrative coded data for diagnosis and procedures, laboratory, and pharmacy data [5]. However, coded data are for the most part unavailable in real time and therefore can be used only for reporting.

Natural Language Processing tools have been used to detect events in real time and have been demonstrated to improve patient safety [6, 7]. However, the installation of such systems can be cumbersome and the performance evaluation is difficult. Laboratory data and pharmacy data, on the other hand, are available in real time and have been used successfully as triggers in reporting and prevention applications [5, 8-10].

When pharmacy data have been used for ADE detection, the focus has been on identification of antidotes following specific drug administration as a marker for ADE. Existing reports have studied the administration of Vitamin K to detect Coumadin overdose as well as the subsequent use of antidiarrheals, anthihistamines, epinephrine, and steroids to detect ADE from antibiotics and other substances. System performance is variable and customization for medical specialties is necessary [9, 11, 12].

The unique physiologic characteristics of patients in Labor and Delivery (L&D) require the use of drugs and measures that are not used as antidotes for AE and ADE in other specialties. For example, a pregnant woman may experience acute hypotension following administration of a regional anesthetic and then be instructed to change from the supine to lateral decubitus position. This change in position, or even the administration of an intravenous fluid bolus to counteract this sudden hypotension, is often not considered an antidote. Likewise, off-label use of medications is very common in the L&D setting. Thus, in order to develop an automatic ADE detecting system for L&D patients, creation of domain-specific electronic rules is necessary [13, 14].

Creation of domain specific rules generally involves the creation of association logic and use of sequence discovery techniques. The most common business application is the market basket analysis, where items are grouped into item sets to identify those that are purchased together. With the identification of frequent item association sets, it is possible to launch marketing and advertising campaigns and increase sales [15, 16]. The technique has also been demonstrated useful for detection of gene sequence associations and in public health surveillance systems [17-20].

In the present study, we studied the application of sequence discovery techniques with clinical healthcare data. The experiments were conducted by comparing patients records of women admitted to Labor and Delivery (L&D) with and without a discharge diagnosis of ADE.

Background

The association rule is a statement of conditional probability X→Y where X is the antecedent (product 1) and Y the consequence (product 2). The significance of the rules is measured based on the number of times each item appears in the dataset and the number of times the two items appear as an item set. The *support* of the rule is the ratio of transactions that include all items in the antecedent and the consequence to the number of transactions that include all items in the antecedent; in other words, the number of times that the combination appears in all the transactions. The *confidence* (incidence) of the rule is the ratio of transactions that include all items in the antecedent and the consequence to the number of transactions that include all items in the antecedent. The *expected confidence* is the number of the consequent transactions divided by the total number of transactions. The *lift* is the ratio of the confidence to the expected confidence or the strength of the association [21, 22]. Another interesting statistic that can be calculated is the *Z-score*, which indicates if the confidence is significantly greater than the expected confidence.

Association rules with high support and confidence are significant and therefore used to construct marketing strategies. However, high support and high confidence rules can lack practical interest when the consequence occurs every time the antecedent is present. An example of high support, high confidence rules would be *a patient with fever is likely to have an infection.* In clinical practice, these associations are known as pathognomonic signs of a given disease. Moreover, if an infrequent association of the consequence and the antecedent is found, the rule becomes *interesting* and would likely have been missed by conventional methods and not easily detected even with extensive domain knowledge. Thus, for the rules to be *interesting* and useful in practice, we should seek those with low support, high confidence, and a lift value of 1 or higher [16, 23].

One of the difficulties with association rules is the number of items to consider. A large number of categories require lengthy computing time and memory often exceeding server capacity. In addition, the process generates large number of rules that will in turn be difficult to analyze and might not be relevant (*interesting*). To avoid this complication, it is necessary to combine the items into broad categories that will require less computing resources and yield fewer and more meaningful rules. The literature reports several approaches; the most common one is a combination of mapping item sets to a classification system and domain expertise [21].

Sequence Discovery

Sometimes it is of interest to know the order in which items are used, or the antecedent and the consequence. Such is the case of a pharmacological application where the purpose is to identify the use of antidote drugs to treat an ADE. In this context, the drug that caused the ADE is the antecedent, and the antidote is the consequence [9, 21, 24-26]. Sequence discovery requires a *time variable* computed based on the order in which the events happen.

Methods

Subjects

Records for the present study were extracted from the Enterprise Data Warehouse (EDW) of Intermountain Healthcare in Salt Lake City, Utah. The EDW contains clinical care and coded data for billing and reporting. Data from 106,480 individual patients admitted for L&D during years 2002-2005 were extracted.

Included in the study were patient records with gestational age between 20 and 44 weeks and birth weight between 500 and 4800 grams. Two patient records with maternal age above 55 were excluded as they were determined to be data entry errors. In patients with multifetal pregnancies, the first maternal record was selected for inclusion.

Three hundred seventy-one patient records were identified with an ADE based on ICD9 codes. A control set of 371 patients with no ADE was randomly selected from the remaining pool of records. The random selection function of the Structured Query Language Open Source version (MySQL V.5.1) was used for this purpose.

Pharmacy data including the medication name, delivery route, and the date/time of administration were extracted for each patient in both the ADE and non-ADE groups.

Drug Category Reduction

To reduce the number of categories, the drugs were grouped by main component. Separate categories were created for those containing multiple components. For identification purposes, names were assigned to the categories based on the active ingredients and, for those with multiple ingredients, the name of the main ingredient was used. A low incidence of use of antihistamines and antidepressants in diverse pharmacological forms was found. To include these drugs without increasing the number of categories, a broad class was created for both of these groups and labeled by the pharmacological action. The drug category reduction process included both identifying similar ingredients with different spellings or abbreviations and similar ingredients with different brand names.

The classification was verified with the Unified Medical Language System (UMLS) by one of the authors (LMT). A second author (MV) validated the drug classification. Each of these clinicians individually has over 15 years of clinical obstetric experience.

The drug categories were then divided into two broad groups: those to be used in the analysis as antecedent and those to be considered as consequence. The consequence or antidote group included antihistamines, ephedrine, external use nasal decongestants, and topical steroids. Medications that could be considered as antidotes in other clinical specialties were discarded from the analysis for reasons explained elsewhere in this paper.

Parenteral and oral steroids are used as antidotes in presence of ADE in other specialties; however, in pregnant patients, parenteral steroids are used as inductors of fetal lung maturity, and, therefore, they were not included in the antidote list. Also excluded from analysis were items that, although found in the pharmacology list, 1) do not have pharmacological action, 2) are used unavoidably in the laboring patient, or 3) are not known to cause ADE (for example, enemas, laxatives, and tucks). The antidote drugs have been previously used as triggers in ADE systems, and their action was validated by the two obstetricians involved in the study.

For each drug, the date and time variables were transformed into discrete time variables. The time variable was used to identify the order in which each drug was administered to the individual patient and thus required by the sequence analysis process.

Association Rules and Sequence Analysis

Separate analyses for ADE and non-ADE sets were conducted. For each set, four different association and sequence analysis experiments were conducted. Each experiment utilized a different antidote drug as the consequence variable. This approach allowed the identification of a specific antidote drug in the item set found by the sequence analysis process.

A frequent set support threshold of 5% and a rule support threshold of 10% were used. The support threshold of 5% means that an item set (antecedent-consequence) must be present at least in 5% of the associations in order to be considered a frequent set. The rule support threshold of 10% means that the item set must be greater than or equal to 10% of the rules to be considered a high-support association rule. Likewise, the maximum number of items in an item set was limited to 2.

Software Packages

Software packages used were Statistical Analysis Software (SAS) Release 9.1 and SAS Enterprise Miner Release 4.3.Structured Query Language (MySQL V5.0) Open Source database management system.

Institutional Review Board

Institutional Review Board approval was obtained from both the Intermountain Healthcare, and the University of Utah.

Results

The study included 1400 cases. Of those, 371 ADE cases were identified based on the clinical classification previously described [27]. The overall ratio of ADE to non-ADE patients was 0.36:100. After applying the pharmacological exclusions described above, the ADE group had 6740 individual drug records and the non-ADE group had 4427 records. The total number of drug categories for the ADE group was 121; the total for the non-ADE group was 93 ($p < 0.001$). The drugs per patient were 14 for the ADE group and 10 for the non–ADE group ($p < 0.001$).

Table 4.1 shows the drug categories and the differences found in each group. In the ADE group, the different drugs used and not found in the non-ADE group were anti-hypertensives, bronchodilators, cardiovascular medications, antibiotics of uncommon use in laboring patients, and drugs used for general anesthesia.

The association rule analyses and the sequence discovery analyses showed significant differences between the groups. In the non-ADE group, we found no significant associations for ephedrine, nasal decongestants, and topical steroids. A significant Z-score was found for antihistamines in which the antecedent component of the item set was cephalosporins and opioid analgesics. The total number of significant associations for antihistamines as antidote was 22. Likewise, in the ADE group, no significant associations were found for item sets with nasal decongestants and topical steroids. Significant associations were found for antihistamines (80 significant associations) and ephedrine (16 significant associations). In the analysis for antihistamines, the most significant associations were with nonsteroidal anti-inflammatory drugs, opioid analgesics, oxytocin, and cephalosporins. Significant associations in the ephedrine as antidote group were with antiemetic, cephalosporins, opioid analgesics, and nonsteroidal anti-inflammatory drugs. Table 4.2 shows the associations found for ephedrine in the ADE group. Table 4.3 shows the association rules found for antihistamines in the ADE group.

Discussion

The main purpose of this study was to identify significant associations between therapeutic drugs used in patients during L&D hospitalization and drugs known as antidote drugs and to create logic utilizing data mining methods in patient records with reported ADE. The resulting associations indicated statistical significance as well as clinical relevance. As expected, we identified no associations between the drugs of frequent use in labor and post-partum and antidote drugs.

From our findings, it is clear that the ADE group includes drugs used in patients in L&D with complicated diagnosis that require the use of a larger number of drugs and ingredients with pharmacological actions outside of labor conduction and puerperal care. This observation is in accordance with those by other authors who have found a larger number of drugs used in patients with subsequent ADE [5, 28].

Our work explores how data mining techniques used in fields outside of medicine apply to clinical healthcare data. As noted above, ADE are underreported; there is pressing need to develop electronic methods for detection and reporting. The validation of data mining methodologies with clinical healthcare data opens new possibilities for future research. These methodologies applied to clinical data are likely to improve automatic detection of ADE.

Limitations

The main challenge of the present study is the clear identification of patients with reported ADE. The use of ICD9 codes for this purpose does not provide a clear identification of the type of adverse drug event as well as the timing of its presentation during the hospitalization. We have no reliable way of determining when the ADE happened, if present on admission, occurring during labor and delivery, or happening in the post-partum period. Nonetheless, by utilizing the pharmacy patient records and the time and date of the drug administration, we were able to create a time variable that allowed not only an association rule analysis but also a sequence analysis in the study groups.

Identification of patients with ADE via ICD9 codes might have missed a number of patients in which the AE was not clearly documented in the medical record and thus not included in the discharge coding. Another limitation of the study is the manner in which the drug categories were assigned: some by drug components, and others by drug action. Those categorized by drug action are of infrequent use and were grouped together in an effort to include them in the analysis without loss of information. However, the approach allowed statistical verification of the different drugs used between patients with and without ADE. It also allowed the association and sequence analysis algorithms to generate significant rules and to identify the most common antidotes used in patients admitted for L&D. The approach also opened new research venues in which clinical variables as well as the inclusion of antidote measures (intravenous fluids, changes in maternal posture, etc.) can be used to identify patients with ADE.

Conclusion

Episode mining techniques can be used with healthcare data to identify associations between drugs and antidotes in patients with ADE. The associations can be used to create computer logic that can aid in the electronic prediction of rare events such as ADE.

Acknowledgments

The present study was conducted with data from Intermountain Healthcare. It was supported in part by the grant No. LM007124 from the National Library of Medicine.

Table 4.1 Differences in the drug categories found in ADE and non-ADE groups. The numbers indicate the different drugs included in the category. Blank spaces in the non-ADE groups indicate drugs that were not found in one group and found in the other.

Drug Categories in No ADE Group	**Number of Different Drug Names in Same Category**	**Drug Categories in ADE Group**	**Number if Different Drug Names in Same Category**
ACETAMINOFEN_MIX	40	ACETAMINOFEN_MIX	67
ACETAMINOPHEN	46	ACETAMINOPHEN	80
AMOXICILLIN	2	AMOXICILLIN	4
AMPICILLIN	76	AMPICILLIN	74
ANTIACID	30	ANTIACID	123
ANTIDEPRESANT	20	ANTIDEPRESANT	10
ANTIHISTAMINE	66	ANTIHISTAMINE	188
		ASPIRIN	1
ATROPINES	1	ATROPINES	5
BETABLOCKER	4	BETABLOCKER	41
		BRONCODILATADOR	9
		CALCIUMBLOCKER	2
		CALCIUMGLUCONATE	1
CELESTONE	6	CELESTONE	18
CEPHALOSPORIN	151	CEPHALOSPORIN	244
CLINDAMYCIN	3	CLINDAMYCIN	33
		CLONIDINE	1
CODEINE_MIX	41	CODEINE_MIX	68
		DESFLURANE	1
DEXAMETHASONE	7	DEXAMETHASONE	8
		DICLOXACILLIN	9
DINOPROSTONE	8	DINOPROSTONE	16
		DOXYCYCLINE	2
		ENOXAPARIN	24
EPHEDRINE	26	EPHEDRINE	52
EPIDURAL PCA	36	EPIDURAL PCA	31
EPINEPHRINE	12	EPINEPHRINE	13
ERYTHROMYCIN	7	ERYTHROMYCIN	13
ESTROGENS	1	ESTROGENS	1
FENTANYL	255	FENTANYL	281
FLEET ENEMA	1	FLEET ENEMA	3
FLUCONAZOLE	2	FLUCONAZOLE	6
		FOLICACID	2
		FOLTX	3

Table 4.1 Continued

Drug Categories in No ADE Grosup	Number of Different Drug Names in Same Category	Drug Categories in ADE Group	Number if Different Drug Names in Same Category
FUROSEMIDE	5	FUROSEMIDE	22
GENTAMICIN	8	GENTAMICIN	27
GLUCOPHAGE	5		
GUAIFENESIN	9	GUAIFENESIN	12
HEPARIN	8	HEPARIN	21
HEPATITISB_VACC	1	HEPATITISB_VACC	3
		HESPAN	1
HYDRALAZINE	1	HYDRALAZINE	24
HYDROCODONE	217	HYDROCODONE	265
HYDROXYCHLOROQUINE	1		
		HYPEROSMOTICIV	13
IBUPROFEN	702	IBUPROFEN	734
		IMIPENEM	20
INDOMETHACIN	2	INDOMETHACIN	5
INSULIN	7	INSULIN	22
IRON	37	IRON	67
IVFLUID	990	IVFLUID	1380
		KETAMINE	3
LANOLIN	68	LANOLIN	76
LAXATIVE	392	LAXATIVE	521
LEVOTHYROXINE	1	LEVOTHYROXINE	8
LOCAL ANESTHETIC	412	LOCAL ANESTHETIC	346
LOCAL ANTIFUNGAL	2	LOCAL ANTIFUNGAL	7
LOCALBACTERIOSTATIC	8	LOCALBACTERIOSTATIC	13
LOCALESTEROID	13	**LOCALESTEROID**	15
		LOPERAMIDE	1
MACROLIDE	1	MACROLIDE	16
MAGNESIUM SULFATE	26	MAGNESIUM SULFATE	84
MEPERIDINE	39	MEPERIDINE	32
		METHYLDOPA	17
METHYLERGONOVINE	26	METHYLERGONOVINE	21
		METHYLPREDNISOLONE	6
METOCLOPRAMIDE	37	METOCLOPRAMIDE	92
		METRONIDAZOLE	6
MIDAZOLAM	10	MIDAZOLAM	29
MISOPROSTOL	11	MISOPROSTOL	18
MORPHINE	109	MORPHINE	243

Table 4.1 Continued

Drug Categories in No ADE Group	**Number of Different Drug Names in Same Category**	**Drug Categories in ADE Group**	**Number if Different Drug Names in Same Category**
		MUSCLERELAXAT	15
NALBUPHINE	37	NALBUPHINE	72
NALOXONE	21	NALOXONE	40
NASALDECONGEST	3	NASALDECONGEST	13
		NEOSTIGMINE	1
NICODERM	2	NICODERM	1
NIFEDIPINE	1	NIFEDIPINE	16
		NITROFURANTOIN	3
		NITROGLYCERIN	3
NSAID	103	NSAID	102
OPIODS_SYNTHETIC	12	OPIODS_SYNTHETIC	16
OXYCODONE	768	OXYCODONE	1152
OXYTOCIN	428	OXYTOCIN	473
PAIN SERVICE CODE	43	PAIN SERVICE CODE	46
PENICILLIN	65	PENICILLIN	81
PHENOBARBITAL	6	PHENOBARBITAL	13
PHYTONADIONE	2	PHYTONADIONE	5
		POTASSIUM IV	6
POTASSIUM TB	1	POTASSIUM TB	4
PRAMOXINE	1	PRAMOXINE	1
PREDNISONE	3	PREDNISONE	3
PRENATAL VIT	14	PRENATAL VIT	34
PROGESTERON	4	PROGESTERON	7
PROMETHAZINE	53	PROMETHAZINE	158
PROPOFOL	3	PROPOFOL	9
		PROTAMINE	1
QUINOLONE	5	QUINOLONE	6
RANITIDINE	15	RANITIDINE	31
RHO(D)	6	RHO(D)	3
ROCEPHIN	1	ROCEPHIN	4
ROFECOXIB	7	ROFECOXIB	31
SCOPOLAMINE	7	SCOPOLAMINE	11
SEDATIVE	37	SEDATIVE	125
		SEVOFLURANE	1
SIMETHICONE	10	SIMETHICONE	47
SUCCINYLCHOLINE	2	SUCCINYLCHOLINE	2
SUFENTANIL	27	SUFENTANIL	26
SYNTHE_PENICILIN	2	SYNTHE_PENICILIN	22
		TACROLIMUS	1
TERBUTALINE	11	TERBUTALINE	16

Table 4.1 Continued

Drug Categories in No ADE Group	**Number of Different Drug Names in Same Category**	**Drug Categories in ADE Group**	**Number of Different Drug Names in Same Category**
THIOPENTAL	2	THIOPENTAL	1
		TRIME/SULFA	1
TROMETHAMINE	77	TROMETHAMINE	217
TUCKS	30	TUCKS	28
UNASYN	5	UNASYN	57
URINARYBACTERIOSTATIC	49	URINARYBACTERIOSTATIC	67
VACCINE	33	VACCINE	34
VALACYCLOVIR	3	VALACYCLOVIR	3
		VANCOMYCIN	11
WARFARIN	5	WARFARIN	17
ZOFRAN	58	ZOFRAN	147

Table 4.2 Examples of significant associations found in the ADE group when the experiments were conducted with "Ephedrine" as the consequence or antidote drug. Lift values > 1 indicate strong association. Z scores > 0 indicate reported confidence is significantly better than the expected confidence.

Associations	**Expected Confidence** [1]	**LIFT**[2]	**Z SCORE**[3]
OXYTOCIN & ZOFRAN ==> EPHEDRINE	8.959537572	2.092741935	1.479392784
OXYCODONE & CEPHALOSPORIN ==> EPHEDRINE	8.959537572	1.842154713	1.148864762
ZOFRAN ==> EPHEDRINE	8.959537572	1.843148861	1.141484828
OXYTOCIN & CEPHALOSPORIN ==> EPHEDRINE	8.959537572	1.800208117	1.067494636
OXYTOCIN & MORPHINE ==> EPHEDRINE	8.959537572	1.621725944	0.848156542
OXYTOCIN & OXYCODONE ==> EPHEDRINE	8.959537572	1.576453435	0.743170182
CEPHALOSPORIN ==> EPHEDRINE	9.826589595	1.453781513	0.694832676
MORPHINE ==> EPHEDRINE	8.959537572	1.482358871	0.658032427
OXYCODONE ==> EPHEDRINE	8.959537572	1.38187404	0.489170438
IBUPROFEN & OXYCODONE ==> EPHEDRINE	8.959537572	1.328725038	0.438526188
OXYTOCIN & IBUPROFEN ==> EPHEDRINE	8.959537572	1.302150538	0.382265023
OXYTOCIN ==> EPHEDRINE	8.959537572	1.295166795	0.360788577
IBUPROFEN ==> EPHEDRINE	8.959537572	1.231316726	0.287507056
OXYTOCIN & LOCAL ANESTHETIC ==> EPHEDRINE	8.959537572	1.193345506	0.26376132
LOCAL ANESTHETIC ==> EPHEDRINE	8.959537572	1.1687215	0.225078067

1. The number of the consequent transactions divided by the total number of transactions.
2. The ratio of the confidence to the expected confidence or the strength of the association.
3. Indicates if the confidence is significantly greater than the expected confidence if the value is ≥ 0

Table 4.3 Examples of significant associations found in the ADE group when the experiments were conducted with "Antihistamines" as the consequence or antidote drug. Lift values > 1 indicate strong association. Z scores > 0 reported confidence is significantly better than the expected confidence.

Rule	**Expected Confidence** [1]	**LIFT**[2]	**Z_SCORE**[3]
OXYCODONE & NALBUPHINE	10.69364162	6.234234234	9.092239132
OXYTOCIN & NALOXONE	10.69364162	6.126747437	8.9709401
CEPHALOSPORIN & HYDROCODONE	17.91907514	3.255376344	8.391711956
IBUPROFEN & NALOXONE	10.69364162	5.731473409	8.279277467
NALOXONE	10.69364162	5.731473409	8.279277467
OXYCODONE & SEDATIVE	10.69364162	5.714714715	8.074596546
MORPHINE & NALOXONE	10.69364162	5.481826654	7.962036239
CEPHALOSPORIN & NALBUPHINE	13.00578035	4.296732026	7.723894444
OXYCODONE & HYDROCODONE	10.69364162	5.30752374	7.429095241
OXYTOCIN & ANTIACID	10.69364162	5.167852063	7.188206449
OXYCODONE & ANTIACID	10.69364162	4.967905405	7.049046977
IBUPROFEN & ANTIACID	10.69364162	4.675675676	6.480116913
ANTIACID	23.12138728	2.1625	6.272427558
OXYTOCIN & NALBUPHINE	10.69364162	4.555786556	6.222024508
MORPHINE & HYDROCODONE	10.69364162	4.415915916	6.068428882
NALBUPHINE	10.69364162	4.463144963	5.972812994
OXYCODONE & TROMETHAMINE	10.69364162	4.675675676	5.847828622
FENTANYL & CEPHALOSPORIN	10.69364162	4.605889472	5.768397041
SEDATIVE	10.69364162	4.411014788	5.725312086
CEPHALOSPORIN & TROMETHAMINE	11.84971098	3.971305595	5.569323353
MORPHINE & ZOFRAN	10.69364162	4.472385429	5.554828564
OXYCODONE & MORPHINE	10.69364162	4.724893314	5.541918916
CEPHALOSPORIN & ZOFRAN	10.69364162	4.483524621	5.512344676
OXYTOCIN & SEDATIVE	10.69364162	4.17826337	5.481483725
LOCAL ANESTHETIC & CEPHALOSPORIN	10.69364162	4.198565913	5.47798347
ZOFRAN & TROMETHAMINE	10.69364162	4.198565913	5.47798347
OXYCODONE & CEPHALOSPORIN	10.69364162	4.721070585	5.447072596
CEPHALOSPORIN & MORPHINE	14.73988439	3.209782838	5.388600831
IBUPROFEN & SEDATIVE	10.69364162	4.091216216	5.331355337

1. The number of the consequent transactions divided by the total number of transactions.
2. The ratio of the confidence to the expected confidence or the strength of the association.
3. Indicates if the confidence is significantly greater than the expected confidence if the value is ≥ 0

Table 4.3 continued

Rule	Expected Confidence [1]	LIFT[2]	Z_SCORE[3]
OXYCODONE & ZOFRAN	10.69364162	4.250614251	5.143786282
OXYTOCIN & TROMETHAMINE	10.69364162	4.143003763	4.973502974
IBUPROFEN & CEPHALOSPORIN	10.69364162	4.375402926	4.960634189
OXYTOCIN & CEPHALOSPORIN	10.69364162	4.298605057	4.738942027
FENTANYL & TROMETHAMINE	10.69364162	3.701576577	4.727301655
LOCAL ANESTHETIC & MORPHINE	10.69364162	3.701576577	4.727301655
IBUPROFEN & TROMETHAMINE	10.69364162	3.886275886	4.643146592
FENTANYL & MORPHINE	10.69364162	3.814366999	4.55339074
IBUPROFEN & MORPHINE	10.69364162	4.020207123	4.532115588
TROMETHAMINE	10.69364162	3.887640449	4.521806156
OXYTOCIN & MORPHINE	10.69364162	4.076230076	4.520957179
MORPHINE	10.69364162	4.091216216	4.457104674
CEPHALOSPORIN & METOCLOPRAMIDE	16.1849711	2.347857143	4.373899694
OXYTOCIN & ACETAMINOPHEN	10.69364162	3.455934195	4.363006106
OXYTOCIN & ZOFRAN	10.69364162	3.798986486	4.338990922
OXYCODONE & URINARYBACTERIOSTATIC	10.69364162	3.435190292	4.293174695
OXYTOCIN & URINARYBACTERIOSTATIC	10.69364162	3.416839917	4.229060704
OXYCODONE & METOCLOPRAMIDE	10.69364162	3.445234708	4.217244671
FENTANYL & ZOFRAN	10.69364162	3.428828829	4.159702985
CEPHALOSPORIN	22.5433526	1.964468864	4.100300793
OXYCODONE & FENTANYL	10.69364162	3.589407589	4.034208595
IBUPROFEN & ZOFRAN	10.69364162	3.493911494	3.967696909
OXYTOCIN & METOCLOPRAMIDE	10.69364162	3.308939709	3.927350211
ZOFRAN	10.69364162	3.517480784	3.864851992
URINARYBACTERIOSTATIC	10.69364162	3.224603914	3.864296066
FENTANYL & HYDROCODONE	10.69364162	3.117117117	3.761092186
OXYTOCIN & OXYCODONE	10.69364162	3.698274546	3.71795073
METOCLOPRAMIDE	10.69364162	3.117117117	3.601072962
IBUPROFEN & METOCLOPRAMIDE	10.69364162	3.005791506	3.536156236
OXYTOCIN & HYDROCODONE	10.69364162	3.029311001	3.4517207
OXYTOCIN & PROMETHAZINE	10.69364162	2.961261261	3.431875182
IBUPROFEN & OXYCODONE	10.69364162	3.395431145	3.394002421

1. The number of the consequent transactions divided by the total number of transactions.
2. The ratio of the confidence to the expected confidence or the strength of the association.
3. Indicates if the confidence is significantly greater than the expected confidence if the value is ≥ 0.

Table 4.3 continued

Rule	**Expected Confidence** [1]	**LIFT**[2]	**Z_SCORE**[3]
IBUPROFEN & PROMETHAZINE	10.69364162	2.838803089	3.266662894
HYDROCODONE	10.69364162	2.970429253	3.265422956
PROMETHAZINE	10.69364162	2.833742834	3.185342421
IBUPROFEN & HYDROCODONE	10.69364162	2.830014225	3.112730369
OXYCODONE & LOCAL ANESTHETIC	10.69364162	2.796665825	2.890292934
ACETAMINOPHEN	25.14450867	1.431724138	2.718090196
OXYTOCIN & FENTANYL	10.69364162	2.757449757	2.637224904
FENTANYL	10.69364162	2.694457169	2.490245565
OXYTOCIN & IBUPROFEN	10.69364162	2.805405405	2.473717672
IBUPROFEN & FENTANYL	10.69364162	2.583926031	2.420184541
OXYTOCIN	10.69364162	2.776681118	2.347538297
IBUPROFEN	10.69364162	2.662306435	2.231214176
LOCAL ANESTHETIC	10.69364162	2.203197962	1.813471929
LOCAL ANESTHETIC & FENTANYL	10.69364162	2.078078078	1.797913837
OXYTOCIN & LOCAL ANESTHETIC	10.69364162	2.117287098	1.758678404
IBUPROFEN & LOCAL ANESTHETIC	10.69364162	2.058473568	1.674933228

1. The number of the consequent transactions divided by the total number of transactions.
2. The ratio of the confidence to the expected confidence or the strength of the association.
3. Indicates if the confidence is significantly greater than the expected confidence if the value is ≥ 0

References

[1] IOM. To Err is Human:Building a Safer Health System; 1999.

[2] Quality Interagency Coordination Task Force to the President. Federal Actions to reduce medical errors and their impact 2000

[3] JCAHO. Joint Commission. [cited 2005; Available from: http://www.jcaho.org/

[4] IOM. Preventing Medication Errors: Institute of Medicine; 2006.

[5] Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc. 2003 Mar-Apr;10(2):115-28.

[6] Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005 Jul-Aug;12(4):448-57.

[7] Tolentino HD, Matters MD, Walop W, Law B, Tong W, Liu F, et al. A UMLS-based spell checker for natural language processing in vaccine safety. BMC medical informatics and decision making. 2007;7:3.

[8] Gardner RM, Evans RS. Using computer technology to detect, measure, and prevent adverse drug events. J Am Med Inform Assoc. 2004 Nov-Dec;11(6):535-6.

[9] Handler SM, Altman RL, Perera S, Hanlon JT, Studenski SA, Bost JE, et al. A systematic review of the performance characteristics of clinical event monitor signals used to detect adverse drug events in the hospital setting. J Am Med Inform Assoc. 2007 Jul-Aug;14(4):451-8.

[10] Classen DC, Pestotnik SL, Evans RS, Burke JP. Computerized surveillance of adverse drug events in hospital patients. 1991. Qual Saf Health Care. 2005 Jun;14(3):221-5; discussion 5-6.

[11] Evans RS, Pestotnik SL, Classen DC, Bass SB, Menlove RL, Gardner RM, et al. Development of a computerized adverse drug event monitor. Proc Annu Symp Comput Appl Med Care. 1991:23-7.

[12] Evans RS, Gardner RM, Bush AR, Burke JP, Jacobson YA, Larsen RA, et al. Development of a computerized infectious disease monitor (CIDM). Comput Biomed Res. 1985 Apr;18(2):103-13.

[13] Briggs GG, Wan SR. Drug therapy during labor and delivery, part 2. Am J Health Syst Pharm. 2006 Jun 15;63(12):1131-9.

[14] Briggs GG, Wan SR. Drug therapy during labor and delivery, part 1. Am J Health Syst Pharm. 2006 Jun 1;63(11):1038-47

[15] Witten IH, Frank E. Data mining : practical machine learning tools and techniques. 2nd ed. Amsterdam ; Boston, MA: Morgan Kaufman 2005.

[16] Tan P-N, Steinbach M, Kumar V. Introduction to data mining. 1st ed. Boston: Pearson Addison Wesley 2006.

[17] Morgan XC, Ni S, Miranker DP, Iyer VR. Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. BMC Bioinformatics. 2007 Nov 15;8(1):445.

[18] Chen Q, Chen YP. Mining frequent patterns for AMP-activated protein kinase regulation on skeletal muscle. BMC Bioinformatics. 2006;7:394.

[19] Creighton C, Hanash S. Mining gene expression databases for association rules. Bioinformatics (Oxford, England). 2003 Jan;19(1):79-86.

[20] Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. J Am Med Inform Assoc. 1998 Jul-Aug;5(4):373-81.

[21] Cerrito PB, SAS Institute. Introduction to data mining using SAS Enterprise Miner. Cary, N.C.: SAS Institute 2006.

[22] Refaat M. Data preparation for data mining using SAS. San Francisco: Morgan Kaufmann Publishers 2007.

[23] McIntosh T, Chawla S. High confidence rule mining for microarray analysis. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM. 2007 Oct-Dec;4(4):611-23.

[24] Handler SM, Wright RM, Ruby CM, Hanlon JT. Epidemiology of medication-related adverse events in nursing homes. The American journal of geriatric pharmacotherapy. 2006 Sep;4(3):264-72.

[25] Szarfman A, Tonning JM, Doraiswamy PM. Pharmacovigilance in the 21st century: new systematic tools for an old problem. Pharmacotherapy. 2004 Sep;24(9):1099-104.

[26] Field TS, Gurwitz JH, Harrold LR, Rothschild JM, Debellis K, Seger AC, et al. Strategies for detecting adverse drug events among older persons in the ambulatory setting. J Am Med Inform Assoc. 2004 Nov-Dec;11(6):492-8.

[27] Taft LM, Evans RS, Shyu CR, Egger MJ, Chawla N, Mitchell JA, et al. Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. Journal of biomedical informatics. 2008 Sep 14.

[28] Classen DC, Pestotnik SL, Evans RS, Burke JP. Computerized surveillance of adverse drug events in hospital patients. Jama. 1991 Nov 27;266(20):2847-51.

CONCLUSIONS

Different KDD techniques were used to extract important medical information from healthcare data. Data transformation, normalization, dimensionality reduction, and descriptive and predictive tasks were all important steps in this study.

The studies conducted in this dissertation allowed the validation of KDD techniques in the healthcare setting with complex obstetrical data. Through the application of data transformation and data preprocessing, it was demonstrated that obstetrical data can be successfully analyzed and useful electronic models can be generated. The overall results of this study indicate that there are still new areas of informatics that remain to be explored using healthcare data. As techniques continue to improve and more coded clinical data becomes available, it will be possible to incorporate automatic processes in the electronic health record that will aid in the early detection of AE.

The techniques utilized in the studies will likely improve in performance if direct clinical data instead of coded billing data is used. Hence, it is important to move forward with research on better ways to store real time coded clinical healthcare data.

There is continuous, ongoing research in disciplines outside of healthcare that develop methodologies to help discover hidden relationships in large, complex datasets. It has been demonstrated that these techniques can also be proven useful in healthcare; continued research is recommended.

CONTRIBUTION TO BIOMEDICAL INFORMATICS

The work presented in this dissertation demonstrated the ability to use specific KDD methods for the analysis of complex datasets in healthcare data for the extraction of useful information. The work presented here validates the use of KDD processes in healthcare data by demonstrating clinically significant conclusions in an area of medicine to which these processes had not been previously applied. These findings are encouraging and lead us in new research directions.

One of the most common complaints with decision support systems is false-positive alerts. The sequence analysis study efficiently shows the application of techniques that involve a time variable. If applied in a real time clinical setting, it could result in a decrease in false-positives alerts, reduce the noise in decision-support systems and possibly increase user acceptance.

Patient data were transformed, manipulated, and normalized with the application of KDD techniques. As noted in the introduction, the KDD process comprises different phases. We exploited an array of methodologies and demonstrated how they can all be applied to healthcare data. Through data manipulation and handling, we were able to extract information that was not otherwise evident from initial analysis of the dataset. The dataset in which the work was performed was complex and the incidence of the sentinel events chosen sparse; various techniques allowed the extraction of meaningful conclusions. The present work stimulates future research to explore other KDD methods to glean medical information from complex and sparse data.

We have demonstrated that new techniques have feasible applications in healthcare data, even if their primary purpose is the application in arenas not related to medicine. Our results encourage innovation and experimentation. New technologies continue to appear daily and are made rapidly available to the end user. Data capturing systems evolve to facilitate user interaction; KDD knowledge also evolves and new algorithms are designed. New research is encouraged as the quality and availability of the data improves; our research demonstrates that complex experimentation is in order.

Our work also demonstrates the importance of domain knowledge and fortifies the position of the field of Biomedical Informatics as an independent discipline. The Biomedical Informatics researcher and practitioner understand the possibility embedded in the use of computational techniques and appreciates the complexity of the clinical domain.