Journal of Biomedical Informatics 42 (2009) 356–364

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery

L.M. Taft [a,*], R.S. Evans [a,e], C.R. Shyu [a,c], M.J. Egger [a], N. Chawla [d], J.A. Mitchell [a], S.N. Thornton [a,e], B. Bray [a], M. Varner [b]

[a] Department of Biomedical Informatics, University of Utah Health Sciences Center, School of Medicine, 30 North 1900 East, Salt Lake City, Utah 84132, USA
[b] Department of Obstetrics and Gynecology, University of Utah, School of Medicine, USA
[c] Informatics Institute, University of Missouri Columbia, USA
[d] Department of Computer Science & Engg., University of Notre Dame, USA
[e] Department of Medical Informatics, Intermountain Healthcare, USA

ARTICLE INFO

Article history:
Received 2 June 2008
Available online 14 September 2008

Keywords:
Adverse drug events
Pregnancy
Labor and delivery
Oversampling
Data-mining

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.

1. 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

* Corresponding author. Fax: +801 581 4297.
E-mail address: Laritza.Taft@hsc.utah.edu (L.M. Taft).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.09.001

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 inter-changeably 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.

1.1. 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].

1.2. 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].

2. Methods

2.1. 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 g. 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.

2.2. 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 et al. [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 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.

2.3. Data validation

Despite shortcomings, numerous clinical and informatics researchers have proven the usefulness of ICD9 coding systems for clinical research [23]. Table 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 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.

2.4. Statistical procedures

2.4.1. Attribute selection or dimensionality reduction

The original dataset consisted of eighty four variables including maternal comorbidities, demographic information, fetal outcomes and surgical procedures. Principal components analysis (PC) and χ^2 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 χ^2 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].

2.4.2. 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. 10-fold cross-validation divides the data into 10 mutually exclusive subsets, and then combines nine 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. 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 h for one instance of 10-fold and 4.5 h for the decision tree. The total time to run the experiments reported was 136.5 h. 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].

2.4.3. 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.

2.4.4. 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

Table 1
Clinical classification and ICD codes

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, E9408, E9409, E941, E9410, E9411, E9412, E9413, E9419, E942, E9420, E9394, E9395, E9396, E9397, E9398, E9399, E940, E9400, E9401, E9421, E9422, E9423, E9424, E9425, E9426, E9427, E9428, E9429, E943, E9430, E9431, E9432, E9433, E9434, E9435, E9436, E9438,
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
Complicated labor	65983
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
Prolonged labor	66201, 66211, 66221, 66223
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 29383, 29384, 29389, 29534, 29570,
Mental alteration	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
Pregnancy induced	64231,64232, 64233, 64234, 64241, 64242, 64243, 64244, 64251, 64252, 64253
Severe pregnancy induced hype	64261, 64262, 64263
Placenta previa	64100, 64101, 64103, 66351, 7620
Polyhydrramnios	65701,65703
Postdates	64511, 64521, 64523
Precipitaded labor	66131, 66133
Preterm pregnancy	64400, 64403, 64413, 64420, 64421
Previous cesarean	65421, 65423
Premature rupture of membrane	65810, 65811, 65813
Prolonged PROM	65821, 65823, 65831
Shoulder dystocia	66041
Streptococcal infection	380, 4100, 4104, V0251
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
Tobacco use	3051, V1582

(*continued on next page*)

Table 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

Selected comorbidities.

Table 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-ADEMean l0 $p < .001$ for number of drugs used in both groups.
Use of Kappa statistic for agreement between the classification based on 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[a]
Manual revision of free text notes from the electronic medical record and the ICD9 codes classification	Kappa statistic: .55–.75 for agreement between free text in the medical record and ICD9 codes for ADE, trauma, hypertension

[a] The disease incidence found by manual revision was higher when reported by ICD9 codes than in the electronic record. The disagreement is attributed to the fact that coding billing and reporting is done based both on electronic and paper records.

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 non-parametric 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 non-parametric 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.

2.4.5. 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.

2.4.6. IRB approval

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

3. 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.

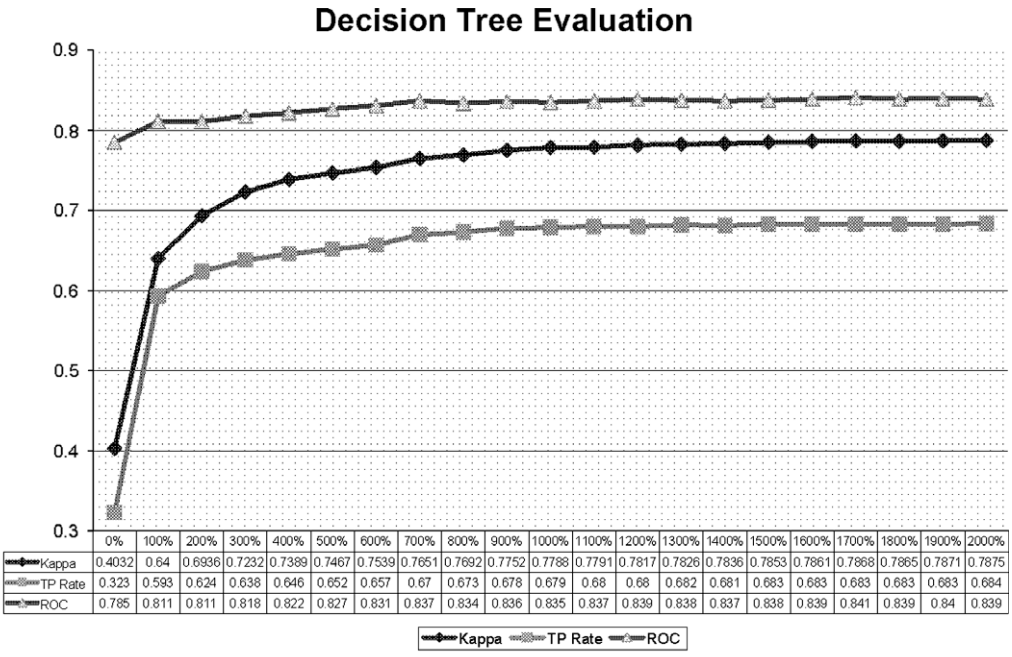
3.1. Performance measures

Figs. 1 and 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. Fig. 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. Fig. 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.

3.2. 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 with synthetic cases remained the same. There was a non-significant 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.

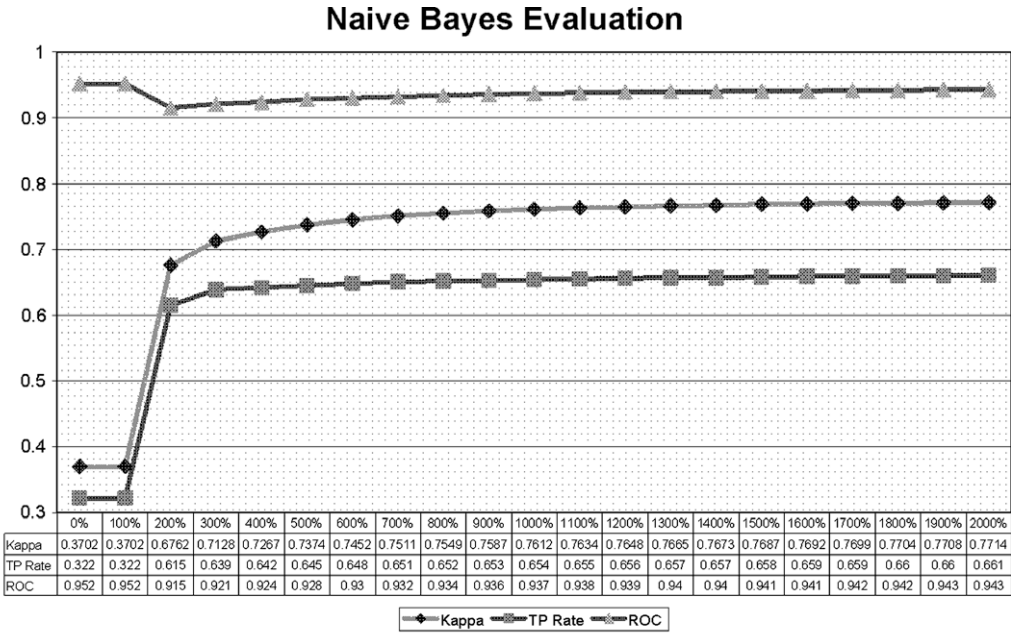
The decision trees in all the models were similar in structure. The first split in the decision tree occurred in patients with external trauma followed by anomalies of the cervix, genito-urinary



	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 Operation Curve (ROC). The Kappa statistic shows the level of agreement for the 10-Fold Cross validation method.

Fig. 1. Performance of the evaluation matrics in the decision tree.



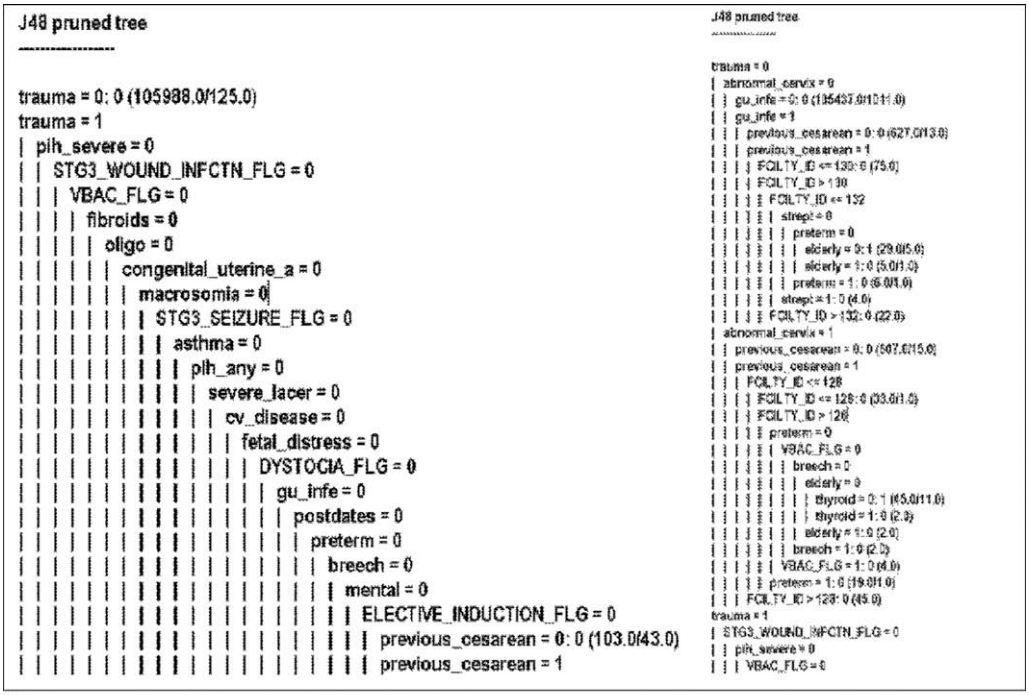
	0%	100%	200%	300%	400%	500%	600%	700%	800%	900%	1000%	1100%	1200%	1300%	1400%	1500%	1600%	1700%	1800%	1900%	2000%
Kappa	0.3702	0.3702	0.6762	0.7128	0.7267	0.7374	0.7452	0.7511	0.7549	0.7587	0.7612	0.7634	0.7648	0.7665	0.7673	0.7687	0.7692	0.7699	0.7704	0.7708	0.7714
TP Rate	0.322	0.322	0.615	0.639	0.642	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 Operation Curve (ROC). The Kappa statistic shows the level of agreement for the 10-Fold Cross validation method.

Fig. 2. Performance of the evaluation matrics in the Naive Bayes classification algorithm.

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. Fig. 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 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



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 equality associated with ADE.

Fig. 3. Comparision of the Structure of the decision trees before and after the SMOTE process.

Table 3
Statistical comparison of classifiers for Kappa statistic and AUC for Naive Bayes and decision trees

Test	Statistic	Value	*p* value
Student's *t*	*t*	29.66	<0001
Sign	*M*	10.5	<0001
Signed rank	*S*	115.5	<0001
Student's *t*	*t*	−2.3	<0.0321
Sign	*M*	−10.5	<0001
Signed rank	*S*	−115.5	<0001

The results show the value for parametric and non-parametric 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 AUC. Non-parametric tests are statistically safer in samples of 30 or less when the assumption of normal distribution is violated.

significant difference for the Kappa statistics both with parametric and non-parametric 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 non-parametric tests.

4. Discussion

The importance of developing automatic detection tools for ADE has 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 inespecific 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, under-reporting 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 an 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].

4.1. Study limitations

In the present study, we found important discordance between the coded data and the text reports in the electronic medical record (Table 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 et al. [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 can not 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 non-preventable 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 over-training. Oversampling leads to overfitting, while random under-sampling 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 non-parametric statistics in the evaluation metrics of both classifiers. The differences of the structure of the decision trees do 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 can not 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 (Fig. 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 hand 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.

4.2. 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 et al. [19].

5. 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.

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. Crit Care Med 2005;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;172(6):390–3.
[5] Tsai PS, Chen CP, Tsai MS. Perioperative vasovagal syncope with focus on obstetric anesthesia. Taiwan J Obstet Gynecol 2006;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;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;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;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;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. Int J Med Inform 2001;63(1–2):109–21.
[12] Weiss GM. Mining with rarity: a unifying framework. SIGKDD Explor Newsl. 2004;6(1):7–19.
[13] Chawla N. V. LA, Hall L.O., 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, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2003 2002:341–78.
[16] Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost: improving prediction of minority class in boosting. Croatia: Dubrovnik; 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, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowledge Data Eng 2003;15(3):1.
[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;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. Utah: Salt Lake City; 2006.
[23] Afifi AA, Clark V. Multivariate analysis, canonical correlation analysis. Computer-aided multivariate analysis. 2nd ed. New York: Van Nostrand Reinhold; 2004. p. 252–70.
[24] Evans RS, Pestotnik SL, Classen DC, Burke JP. Evaluation of a computer-assisted antibiotic-dose monitor. Ann Pharmacother 1999;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, Hall Lawrence O, Bowyer Kevin W. SMOTEBoost: improving prediction of the minority class in boosting 2003.
[28] Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 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;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. Am J Med Genet 2005;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;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. J Clin Epidemiol 2004;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. Int J Cancer 2005;117(4):655–61.
[34] Geller SE, Cox SM, Kilpatrick SJ. A descriptive model of preventability in maternal morbidity and mortality. J Perinatol 2006;26(2):79–84.
[35] Geller SE, Rosenberg D, Cox SM, Kilpatrick S. Defining a conceptual framework for near-miss maternal morbidity. J Am Med Womens Assoc 2002;57(3):135–9.
[36] Holden DA, Quin M, Holden DP. Clinical risk management in obstetrics. Curr Opin Obstet Gynecol 2004;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;274(1):29–34.