Insights from Predicting Pediatric Asthma Exacerbations from Retrospective Clinical Data

William Elazmeh
Stan Matwin
S.I.T.E.
University of Ottawa, Canada

Dympna O'Sullivan
Wojtek Michalowski
School of Management.
University of Ottawa, Canada

Ken Farion
Emergency Department
Children's Hospital of Eastern Ontario
Ottawa, Canada

Abstract

The paper presents ongoing issues, challenges, and difficulties we face in applying machine learning methods to retrospectively collected clinical data. The objective of our research is to build a reliable prediction model for early assessment of emergency pediatric asthma exacerbations. This predictive model should be able to distinguish between patients with *mild* or *moderate/severe* asthma attacks at a medically acceptable level of performance. Our real-life data set presents us with some difficult challenges which we communicate in this paper. Our approach to overcoming some of these difficulties is to use external expert knowledge to aid with classification by decomposing the classification problem into a two-tier concept, where concepts can be explicitly described in terms of the external knowledge source. Such an approach also has the advantage of significantly reducing the size of the training set required.

Introduction

Applying artificial intelligence techniques in medical domains is a active area of research with many open problems. Machine learning researchers are confronted with demands to solve medical problems such as classification and prediction using clinical data that features difficult domain-specific characteristics and properties (Mullins *et al.* 2006). Some of these challenges are described by Cios & Moore, who describe the fact that medical data are often heterogeneous in source as well as data structures, and that the pervasiveness of missing values for technical and/or social reasons can create problems for automatic methods for classification and prediction. Furthermore, the task of translating physicians' interpretations based on years of clinical experience to mathematical models poses a serious challenge. Despite these difficulties medical data mining can be most rewarding, as the appropriate formulation of medical queries coupled with the retrieval of relevant information or predictions can mean providing comfort or respite to a sick person or in more serious cases saving or extending a person's life (Cios & Moore 2002). Our efforts are focused on the domain of emergency pediatric asthma. The domain is worthy of investigation as according to Lozano *et al.*, emergency medical care provided to asthmatic children accounts for 65% of all direct costs of asthma care. The provision of computer-based decision support to emergency physicians treating asthma patients has been shown to lead to an increase in the overall effectiveness of health care delivered in emergency departments and has already been demonstrated by the use of classification using decision trees (Kerem *et al.* 1990; Lieu *et al.* 1998). However, for asthma-related predictions, the performance of these decision tree models ranged from 79% sensitivity and 75% specificity (Kerem *et al.* 1990) to 32% sensitivity and 94% specificity (Lieu *et al.* 1998). This performance remains inadequate in terms of achieving a balance between high sensitivity and high specificity. Analyzing the trade-off between sensitivity and specificity is common in medical domains and is analogous to Receiver Operating Characteristics (ROC) analysis (Sox *et al.* Boston 1998; Faraggi & Reiser 1998) used in machine learning (Provost & Fawcett 1997).

In this paper, we communicate our experiences as well as the lessons learned from analyzing retrospectively collected asthma data using established machine learning methods. Following several attempts to build a reliable prediction model, we identify a two-tier concept in the data with aid from external medical knowledge. The external knowledge employed is the Preschool Respiratory Assessment Measure (PRAM) for asthma severity determination and the measure is used to partition our data set into typical and non-typical cases for more effective classification. Thus, a different classifier is built for each tier and a crucial task is to identify which tier a target case conforms to. Our work in developing a complete classification model remains ongoing, however we present some initial experimental results that support our findings with regard to two-tiered descriptions of flexible concepts (Bergadano *et al.* 1992). The rest of this paper is organized as follows. We begin by providing an in-depth review of asthma exacerbations assessment in emergency medicine. We continue with a description of some of the issues associated with retrospectively collected clinical data and describe the data set used in this research. Following this we outline our external medical knowledge source (PRAM), and our rationale in applying this information to our classification and prediction task in the form of a two-tiered classifier. We continue with a discussion of a number of selected classifiers applied to this task. We describe the results of an initial evaluation by presents some experimental results and finally we conclude with a discussion.

Assessment of Asthma Exacerbations

For a patient suffering from an asthma exacerbation, the early identification of the severity based on patient's symptoms is a crucial part of the diagnostic procedure. Using available clinical information, a physician must determine if an asthma exacerbation is *mild*, *moderate*, or *severe* and recommend an appropriate treatment. Early identification of the severity of an asthma exacerbation has implications for the child's management in the emergency department. Patients with a *mild* attack are usually discharged following a brief course of treatment (less than 4 hours) and resolution of symptoms, patients with a *moderate* attack receive more aggressive treatment over an extended observation in the emergency department (up to 12 hours), and patients with a *severe* attack receive maximal therapy before ultimately being transferred to an in-patient hospital bed for ongoing treatment (after about 16 hours in the emergency department). In clinical practice, a decision on the severity and subsequent disposition of an attack is ideally made as soon as possible after arrival of the patient to the emergency department to ensure key therapies have been instituted. Underestimation of severity may result in inadequate treatment, premature discharge and a possible return visit, while overestimation of severity may result in an extended emergency department stay and unnecessary utilization of hospital resources. Our goal is to develop a classification algorithm using all available patient information that can predict an early recommendation to determine the severity of the asthma exacerbation.

Clinical Data Collection and Issues

Collections of clinical information exist in two primary formats. Information may have been collected in a retrospective manner, usually transcribed from paper patient charts. In such circumstances there is usually no prior intention that the information may be used for analysis. Alternatively, the information may have be gathered as part of a prospective data collection with the intent purpose that the information will be used for analysis, including classification tasks. Regarding the quality of classification, the analysis of prospectively collected data is highly favorable. Such data tends to be collected in a well-defined systematic and structured manner complete with fine-grained levels of detail describing the problem and problem domain. However, prospective data collection is expensive, particularly in medical fields. This cost is measured not only in monetary terms but also in terms of the extra time required by already busy medical professionals in meticulously completing data entry tasks.

As a result most of the clinical information available for computer-based medical classification has been collected in a retrospective manner. Retrospective data poses difficulties for classifiers in many fields but there are a number of unique challenges posed by clinical information. Firstly, the vast majority of the information is collected using often physically disparate paper-based patient charts and must therefore be transcribed to an electronic format. This process can lead to two distinct problems. Firstly it may result in the addition of noise through handwriting misinterpretation and mistranslation. Secondly, the process is known to lead to a loss of valuable clinical knowledge as information is reduced to a format that can be more readily employed by electronic parsers. Examples of such processes are the translation of certain clinical attributes from a range of values to nominal forms or the elimination of important clinician notes or annotations as this information cannot be directly mapped to predefined clinical attributes required for classification. A further difficulty that is particularly pronounced in retrospective clinical data is that of missing values. Missing values occur for a number of reasons. Busy clinicians may not have the required time to manually complete all patient information. However more interestingly, the issue of missing values in clinical data is directly related to the concept of tacit clinical knowledge. With a wealth of clinical experience physicians can discriminate which clinical information is most important for a particular patient and can prioritize what information they record based on this knowledge. Furthermore, they are aware of dependencies between clinical attributes. For example, they may know that if a particular symptom is not present in a particular patient then one or more other symptoms will also be absent and therefore the recording of this information is pointless. This is not however, transparent to the classification algorithms. However, despite these shortcomings, the high costs of prospective data collection in medical fields means that retrospective collections of data remains the most widespread sources of data available to computer-based classifiers. Classification efforts must therefore focus on addressing these challenges associated with retrospective data by developing robust and scalable solutions.

The clinical data used in this study was collected as part of a retrospective chart study conducted in 2004 at the Children's Hospital of Eastern Ontario (CHEO). The study includes patients who visited the emergency department from 2001 to 2003 for treatment of an asthma exacerbation. The data describes clinical and non-clinical information recorded at triage by a nurse and and during further assessments by a medical physician. Based on these assessments, the physician initiates management of the asthma exacerbation, including repeated bronchodilator treatments (so-called masks) at intervals ranging from a few minutes to a few hours. Asthma management also includes systemic corticosteroids for patients recognized as having *moderate/severe* exacerbations and all prescribed treatments are recorded in patients' charts. Throughout the stay, a patient is reassessed regularly to evaluate their response to the prescribed treatments. Depending on a variety of external factors (patient's condition, clinicians' workload, etc.), these repeated assessments are performed at irregular intervals and are inconsistently recorded in the chart. However, it is important to collect as much information as possible from these repeated assessments as this data shows the progression of the patient state over time, which is an important characteristic for machine learning techniques for automatic prediction. All information characterizing physician evaluation, triage

Table 1: PRAM (Chalut, Ducharme, & Davis 2000)

Signs	0	1	2	3
Suprasternal indrawing	absent		present	
Scalene retractions	absent		present	
Wheezing	absent	expiratory	inspiratory and expiratory	Audible without stetho-scope /absent with no air entry
Air entry	normal	decreased bases	widespread decrease	absent/ minimal
Oxygen saturation	$\geq$95%	92-95%	<92%	

assessment, repeated assessments and prescribed treatments are combined to produce a data record for each visit to the emergency department. To select the most relevant assessment taking place nearest to the 2 hour point from triage, we identify the 12 most complete reassessment values between the 100 and 140 minutes starting from triage. If results of tests (temperature, respiratory rate, heart rate and saturation) are missing in the selected reassessment, we extract them from a repeated assessment in the preceding 20 minutes, if one is available. After the records have been created the values of attributes are discretized and categorized by experts. All attributes appear in the emergency triage assessment record and asthma pathway used in the emergency department at CHEO, thus their values are routinely collected and recorded. Each record was reviewed and assigned to one of the two groups (*mild* or *moderate/severe*) using an asthma exacerbation severity category documented in the patient chart and confirmed later by lack of a subsequent visit within the next 24 hours. This procedure allowed us to control for those cases where the initial visit resulted in a premature discharge and subsequent readmission. In this sense we used the confirmed severity group as a gold standard for creation and evaluation of prediction models. The result is a data set of 362 visits to the emergency department that was split into 239 for training and 123 for testing based on patient's visit date.

External Medical Knowledge: PRAM

The goal of our research is to develop an effective prediction model for automatic classification of pediatric asthma patients. Due to the outlined shortcomings of retrospective clinical data, it is difficult to develop such a model based on only the data itself. As such our approach involves the incorporation of external sources of medical knowledge that can be used to analyze and organize the data in a meaningful manner to accurately describes the clinical problem at hand as well as specific problem instances (patients). Specifically for our purposes we use the Preschool Respiratory Assessment Measure (PRAM) asthma index (see table 1). PRAM is a discriminative and responsive index of acute asthma severity for preschool children. It is based on five clinical attributes commonly recorded for pediatric asthma patients, *suprasternal indrawing, scalene retractions, wheezing, air entry* and *oxygen saturation*. PRAM is based on a 12 point scoring scale calculated using scores of 0, 1, 2, and 3 assigned to the attributes depending on their presence or absence, or increased or decreased values. It is a well known and clinically validated criterion for interpreting the severity of airway obstruction where a *mild* exacerbation is assigned a score of less than 5 and a *moderate/severe* exacerbation is assigned a score of greater than or equal to 5. The objective of the PRAM scoring systems is to uniformly select and weight vital patient signs observed during management of an asthma patient. The method can address problems of physician subjectivity by introducing a more rigid stratification into the diagnostic process by drawing attention to the most discriminatory attributes and allowing physicians to focus on fewer and important pieces of information.

In this research we use the PRAM index as a secondary knowledge source to identify examples to use for training a classification model. In the data set a decision (class) is recorded for each patient along with other clinical information. This class indicates whether the patient has suffered from a *mild* or *moderate/severe* exacerbation. Using the attributes outlined by the PRAM scoring system we can calculate a PRAM score for a given patient by mapping these PRAM attributes to attributes in our data set. This value can then be compared with the decision (class label) to identify those patients whose attribute values comply with the PRAM scoring system and who can be considered as "typical" asthma patients. This set of "typical" patients must have complete values for all PRAM attributes in their patient record and can be used to build a classifier that identifies such "typical" patients and can automatically predict the severity of an asthma exacerbation based on the complete PRAM score. Other patients in the data set with only partial or incomplete PRAM values must be identified using a different classification mechanism. The classification task is the identification of "typical" and other patients based on the external knowledge source and the prediction task is whether a patient is experiencing a *mild* or *moderate/severe* asthma exacerbation based on the classification result.

Two-tiered description of flexible concepts

Our classification of asthma patients as "typical" or 'other draws on related work into concept representation (Bergadano *et al.* 1992). The premise is to distinguish between instances based on the extent of their *typicality* by deriving a two-tier description of, what they call a *flexible concept*. In this representation, the base concept (the first tier) describes the explicit and common meaning of typical instances while the second tier, the inferential concept interpretation, defines permitted modifications and variations from the base concept, thus, describes the remaining instances of data that cannot be perfectly matched by the base concept. This approach is significantly different from classic machine learning methods, empirical or analytical, which assume that concepts are precise, independent of the context, and are representable by a single symbolic description.

Table 2: Evaluation Metrics for Classification of Patient Length of Stay as M= Mild or O= Other. TM= True Mild, TO= True Other, FM= False Mild, and FO= False Other

	Predictions	
Class	M	O
S	TM	FO
O	FM	TO

$$\text{sensitivity} = \frac{\text{correctly classified positives}}{\text{total positives}}$$

$$\text{specificity} = \frac{\text{correctly classified negatives}}{\text{total negatives}}$$

$$\text{sensitivity}(\mathtt{M}) = \frac{TM}{TM+FO}, \text{specificity}(\mathtt{M}) = \frac{TO}{FM+TO}$$

$$\text{sensitivity}(\mathtt{O}) = \frac{TO}{TO+FM}, \text{specificity}(\mathtt{O}) = \frac{TM}{FO+TM}$$

$$\text{predictive accuracy} = \frac{TM+TO}{TM+TO+FM+FO}$$

The benefits of the two-tier approach allows the efficient classification of *typical* instances that match the base concept while the remaining instances can be classified by an alternate classification model. Furthermore, they show that this division of tiers in the classification concept is suitable for learning flexible concepts, i.e., concepts that lack precise definition and are context-dependent.

In this work, we present empirical evidence to suggest that, when using external medical knowledge in the form of PRAM rules while predicting the severity of asthma exacerbations in children at the two-hour time is a two-tier concept. Our experiments show that the added medical knowledge can be effectively used to describe a subset of "typical" patients which can be used to enhance the classification.

Data Analysis

The desired classifier predicts the severity of asthma exacerbations for a patient based on their length of stay in the emergency department. A Mild stay represents *mild* asthma exacerbations while an Other represents *moderate* or *severe* asthma exacerbations. The confusion matrix and evaluation metrics are described in Table 2. The matrix shows the result of classification in terms of *True Mild* (TM), *True Other* (TO), *False Mild* (FM), and *False Other* (FO). The standard methods used in medicine to measure diagnostic testing are *sensitivity* and *specificity* and *predictive accuracy*. In this work we analyze our data set in such terms, however we use the Area Under the ROC curve (AUC) as an alternative to accuracy (Ling, Huang, & Zang 2003), following research from the machine learning community that demonstrated that measuring the accuracy of classification is inadequate (Provost & Fawcett 1997) and, in some cases, inappropriate.

As described by (Cios & Moore 2002), *sensitivity* measures how often the classifier finds a set of positive examples. For example, *sensitivity(M)* measures how often the classifier finds patients with Mild asthma exacerbations. Similarly, *sensitivity(O)* measures how often the classifier finds patients with Other or *moderate/severe* asthma exacerbations. *Specificity* measures how often what the classifier finds is indeed what it was looking for. For example, *specificity(M)* measures how often what the classifier predicts is indeed a patient with a Mild asthma exacerbation. Similarly, *specificity(O)* measures how often what the classifier predicts is indeed a patient with a Other length of stay, or a *moderate/severe* asthma exacerbations. By inspecting the *sensitivity* and *specificity* expressions shown in table 2, *sensitivity(M)* equals *specificity(O)* and vise versa. Therefore, the analysis of *sensitivity(M)* against *specificity(M)* is equivalent to comparing *sensitivity(S)* against *sensitivity(O)*. For our problem it is important to put the interpretation of this analysis into perspective. The sensitivities of Mild and Other show how often the classifier finds patients in these classes respectively. Maximizing these two sensitivities is important for better classifier performance. In addition, their specificities show how effective the classifier is in finding patients in both classes. However, the specificity of Mild is more important than the specificity of Other. Higher specificity of M ensures that fewer patients of class Other are misclassified. Thus, fewer patients with *moderate/severe* asthma exacerbations are being misclassified as *mild* which in real life equates to the fact that fewer patients experiencing a *moderate/severe* asthma exacerbation are sent home too early from the emergency department. The inverse, misclassifying Other, is a less serious error. In this work, we report classifier performance in terms of *sensitivity*, *specificity*, the overall *predictive accuracy*, and the *AUC*.

An important aspect of classifier evaluation in medical domains is the comprehensibility of the classification model. Physicians place strong emphasis on the medical interpretation and explanation of classification models. This is because the explanatory capability of the prediction model is important to medical experts such that they can understand the reasoning behind predictions made by the machine learning system (Perner 2005). As such, this requirement eliminates several candidate methods for classification. In the context of this paper, the issue of model comprehensibility is limited to the selection of learning methods. We address this issue in more details in the next section.

Table 3: Selected Classification Models

Classifier		Model
NB	Naive Bayes	probabilistic
J48	Decision Tree	tree-based
NBT	decision tree with a Naive Bayes at each leave	tree-based
LWL	Locally Weighted Learning using KNN and Naive Bayes to assign weights to instances	instance based

Machine learning methods

The choice of learning models applicable in medical domains is limited to those models that offer systematic explanation of the prediction process. Such models include classifiers that estimate probabilities (probabilistic), classi-

Table 4: Experiment Settings

Setting	Train	Test	Description
A	239	123	train/test is based on patient arrival time in the emergency
B	239	88	train as A and test on patients that do not conform to PRAM
C	50	88	train on patient that comply with PRAM and test as B

fiers that identify training examples similar to the test example to be classified (case-based), classifiers that describe classification decisions based on a selected set of attributes (tree-based), and classifiers that produce rules that can be applied to a given test patient for classification (rule-based). For these reasons we select classifiers listed in table 3. Furthermore and despite the medical comprehensibility requirement, two additional candidate classification models are the Random Forests (RF) and the Bagged Decision Trees. Strong empirical evidence suggests they perform well despite their complexity (Caruana & Niculescu-Mizil 2006). At the least, they can be used to put classifier evaluation results into perspective, however, the relatively small size of data can cause the Bagged Decision Trees to over-fit. Therefore, we only include the Random Forests RF.

Experiments

Initially, the objective of our experiments was to build a reliable and comprehensible classification model to predict the severity of asthma exacerbations based on the data we described earlier. However, in an attempt to improve classifier performance, we use external medical knowledge, in the form of PRAM rules, to identify patients that illustrate the *crisp* classification model. Following the experimental settings shown in table 4, our experiments show that PRAM rules are able to identify *typical* asthmatic patients. We train classifiers listed in table 3 on training data for each of these settings and test on their corresponding test sets. With the assistance of a medical physician, we determined a mapping between signs considered by PRAM (see table 1) to attributes recorded in our data set. This is a tedious task that requires medical interpretations and expertise. Never-the-less, all signs *retractions, inspiratory wheezing, expiratory wheezing, air entry,* and *Oxygen saturation* are mapped to attributes recorded in our data set, with the exception of *Suprasternal indrawing* for which we failed to determine a mapping. In both of the original training and testing sets, only 24 and 35 records contained values for the mapped attributes respectively. Applying PRAM to these 59 patients yields 50 correct classifications, 5 incorrect classifications, and 4 outliers. The process involved computing PRAM scores for each of the 59 patients and comparing PRAM's decision of severity to the class label present in the patient record. Following discussion with an expert physician, the outlier records were disqualified from our study.

In experiment setting A, the 239 training instances and 123 testing instances are split based on patients' arrival time in the emergency department. These settings are used to produce a baseline for classifier performance indicating how well a particular classification model performs on this data set. In setting B, the training set remains the same as that of setting A, however, the testing set contains those patients for whom PRAM scores could not be computed due to missing values in the PRAM selected attributes. This reduces the testing set to only 88 records to avoid testing on instances that are selected for training in subsequent setting (C). For experimental setting C, the training set consists of those selected 50 records (extracted from both original training and testing sets) for which PRAM scores compute asthma exacerbation severity scores and produce classifications that correctly match their class labels (*Mild* and `Other`). The testing set is identical to that of setting B. In this final setting, the idea is to train classifiers on patients that have been identified by PRAM to describe those *typical* patients. Our idea is to compare classifier performance when (A) trained on the original training set and tested on the original testing set, (B) trained on the original training set and tested on the remaining patients (after removing those that conform with PRAM), and (C) trained on those records identified by PRAM and tested on the same patients as B. Our experimental results are presented in table 5 as percentages of *sensitivity*, *specificity*, *AUC* computed for the class `Mild` along with the overall *predictive accuracy*. The bold entries are the highest values among the three experimental settings. Recall, *sensitivity* of `Mild` shows how often patients with *mild* exacerbations are detected, *specificity* of `Mild` shows the proportion of what the classifier captures as patients with *mild* exacerbations is indeed correct, *AUC* is a scalar score in the ROC space, and the overall *predictive accuracy* shows the percentage of correctly classified test instances.

Table 5: Classifier Performance as a Percentage

Model	Sensitivity	Specificity	Accuracy	AUC
NB_A	79	74	76	80
NB_B	79	75	77	82
NB_C	**85**	**78**	**82**	**88**
$J48_A$	40	**90**	63	**74**
$J48_B$	56	83	**68**	73
$J48_C$	**71**	63	67	66
NBT_A	66	78	72	76
NBT_B	69	**83**	**75**	**80**
NBT_C	**77**	68	73	76
LWL_A	54	**83**	68	75
LWL_B	58	**83**	69	77
LWL_C	**88**	78	**83**	**84**
RF_A	65	**86**	**75**	**82**
RF_B	60	70	65	75
RF_C	**83**	55	71	77

For the first four basic models (NB, J48, NBT, and LWL), we observe that training on the 50 instances identified by PRAM produces classifier performance higher or equivalent to that of training on the original 239 records. For setting (C), *sensitivity* is consistently highest, *specificity* is stable but decreases slightly, *predictive accuracy* is higher for most models, with a strong increase in *AUC* for NB model. In fact, the NB model produces significantly higher performance when trained on the selected 50 instances (setting C) than training on the original training set (settings A and B). This observation is important because it suggests that building a classification model on instances selected by PRAM produces classifiers capable of performing the same or better than those built by training on the significantly larger training set. Therefore, PRAM is external medical knowledge that can be considered to describe *typical* asthmatic patients. However, the overall classifier performance remains inadequate for medical acceptance. This calls for further investigation of the two-tier concept present in this data and the development of a suitable classifier for those patient's that do not conform to the PRAM scoring system.

At this point, consider the performance of the Random Forests (RF). It is clear that training this model on fewer training instances causes a consistent deterioration of performance, with the exception of *sensitivity*. A possible explanation for this deterioration lies in its basic principle of building a tree ensemble by training on randomly selected subsets of features. Reducing the size of the training set can easily affect its performance. Furthermore, an ensemble of trees built by training on the original training set may be capable of capturing the underlying concept and its variations. This is can explain the strong performance of the RF model when trained on the original 239 training instances. However, when comparing with the performance of the NB classifier trained on the selected 50 records (setting C), we see that the use of PRAM can capture more of the *typical* patients rather than using a random selection of features to build an ensemble. This suggests the need for further consideration of addressing the two-tier concept that PRAM is clearly capable of identifying. We need to address the issue of how to build a classifier capable of recognizing patients that belong to which tier of such a concept.

Conclusions

In summary, this paper presents work-in-progress towards building an effective classification model to predict the severity of asthma exacerbations in children in the emergency department. The desired medical quality of such predictions requires due care and attention in addressing several issues and challenges particular to medical data. Our experiments employ an external medical source of knowledge, PRAM asthma severity scores, to identify *typical* patients to decompose the classification problem into a two-tier concept. Our results show that building an appropriate model to address such decomposition can significantly improve classifier results while reducing the size of the training set. Our research in this area is continuing in a number of directions. First, we are investigating appropriate methods to determine how patients can be classified according to different tiers of the concept. Second, we are determining which classification model is best suited for which tier for more effective prediction.

Acknowledgments

The authors would like to thanks Dr. Szymon Wilk for help in creating and describing the retrospective data set used in this research. Research reported in this paper was supported by the grants from NSERC/CIMR Collaborative Health Research Initiative.

References

Bergadano, F.; Matwin, S.; Michalski, R. S.; and Zhang, J. 1992. Learning two-tiered descriptions of flexible concepts: the poseidon system. *Machine Learning* 8:5–43.

Caruana, R., and Niculescu-Mizil, A. 2006. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd Intl. Conf. on Machine Learning* 161–168.

Chalut, D. S.; Ducharme, F. M.; and Davis, G. M. 2000. The preschool respiratory assessment measure (pram): A responsive index of acute asthma severity. *Pediatrics* 137(6):762–768.

Cios, K. J., and Moore, G. W. 2002. Uniqueness of medical data mining. *A. I. in medicine* 26(1-2):1–24.

Faraggi, D., and Reiser, B. 1998. Computer-based models to identify high-risk children with asthma. *American Journal of Respiratory and Critical Care Medicine* 157(4):1173–80.

Kerem, E.; Tibshirani, R.; Canny, G.; Bentur, L.; Reisman, J.; Schuh, S.; Stein, R.; and Levison, H. 1990. Predicting the need for hospitalization in children with acute asthma. *Chest* 98:1355–1361.

Lieu, T. A.; Quesenberry, C. P.; Sorel, M. E.; Mendoza, G. R.; and Leong, A. B. 1998. Computer-based models to identify high-risk children with asthma. *American Journal of Respiratory and Critical Care Medicine* 157(4):1173–80.

Ling, C. X.; Huang, J.; and Zang, H. 2003. Auc: a better measure than accuracy in comparing learning algorithms. *Canadian Conference on AI* 329–341.

Lozano, P.; Sullivan, S.; Smith, D.; and Weiss, K. 1999. The economic burden of asthma in us children: estimates from the national medical expenditure survey. *The Journal of allergy and clinical immunology* 104(5):957–63.

Mullins, I. M.; Siadaty, M. S.; Lyman, J.; Scully, K.; Garrett, C. T.; Miller, W. G.; Muller, R.; Robson, B.; Apte, C.; Weiss, S.; Rigoutsos, I.; Platt, D.; Cohen, S.; and Knaus, W. A. 2006. Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Computers Biology and Medicine* 36(12):1351–77.

Perner, P. 2005. Intelligent data analysis in medicine - recent advances. *A. I. in medicine* 37(1):1–5.

Provost, F., and Fawcett, T. 1997. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. *The Third International Conference on Knowledge Discovery and Data Mining* 34–48.

Sox, H. C. J.; Blatt, M. A.; Higgins, M. C.; and Marton, K. I. Boston, 1998. *Medical Decision Making*. Butterworths.