A Tree-based Decision Model to Support Prediction of the Severity of Asthma Exacerbations in Children

Ken Farion
Departments of Pediatrics and Emergency Medicine, University of Ottawa
Ottawa, Canada

Wojtek Michalowski, Szymon Wilk[1], Dympna O'Sullivan
Telfer School of Management, University of Ottawa
Ottawa, Canada

Stan Matwin
School of Information Technology and Engineering, University of Ottawa
Ottawa, Canada
Institute of Computer Science, Polish Academy of Sciences
Warsaw, Poland

Abstract

This paper describes the development of a tree-based decision model to predict the severity of pediatric asthma exacerbations in the emergency department (ED) at two hours following triage. The model was constructed from retrospective patient data abstracted from the ED charts. The original data was preprocessed to eliminate questionable patient records and to normalize values of age-dependent clinical attributes. The model uses attributes routinely collected in the ED and provides predictions even for incomplete observations. Its performance was verified on independent validating data (split-sample validation) where it demonstrated AUC (area under ROC curve) of 0.83, sensitivity of 84%, specificity of 71% and the Brier score of 0.18. The model is intended to supplement an asthma clinical practice guideline, however, it can be also used as a stand-alone decision tool.

Keywords:

decision making; asthma; child; retrospective studies; decision trees

[1] Corresponding author:
Szymon Wilk
Telfer School of Management, University of Ottawa
55 Laurier Ave East, Ottawa, ON K1N 6N5
tel.: +1 613-562-5800 x. 4196, fax: +1 613-562-5164, e-mail: wilk@telfer.uottawa.ca

1. Introduction

Asthma exacerbations are one of the most common medical reasons for children to be brought to the emergency department (ED). These visits, and subsequent hospitalizations required by many of these patients, account for nearly 65% of all direct costs of asthma care. Despite such prevalence, several studies demonstrate extensive variation in care provided to asthmatic patients in the ED [1, 2]. In an attempt to standardize care and improve patient outcomes [3], asthma clinical practice guidelines (CPGs) have been developed by national bodies [4, 5], however their clinical use is limited. This may be attributed to several problems common to CPGs developed for other clinical conditions, including their availability in a paper format that requires translation into a computer readable format [6] for better integration with a clinical workflow, and the need for customization to site-specific characteristics – this task alone constitutes up to 90% of the total implementation effort [7].

Our research is concerned with customizing a pediatric asthma CPG to a local setting. More specifically, we aim at simplifying the use of the CPG by using available clinical attributes (signs, symptoms and tests) and by allowing incomplete information about patient's state. Figure 1 presents a general schema of the pediatric asthma CPG published by the Canadian Association of Emergency Physicians (CAEP) [5]. Although the schema delineates the specific CPG, it can be easily generalized for other pediatric asthma guidelines. It begins with a decision step where the severity of the exacerbation is evaluated. Then, for each possible outcome, a corresponding action step specifies how to manage the patient. Asthma CPGs usually have four levels of severity – the first three (mild, moderate and severe) correspond to situations managed in the ED and the last (near death) requires immediate hospitalization in the intensive care unit. Because of the clinical specificity and rarity of patients in this last level, we focus on mild, moderate, and severe conditions only.

A gap between the published CPG and local clinical practices or contexts intervenes mostly during the step of severity evaluation. We encountered this problem when trying to reconcile the decision criteria from the CAEP CPG with patient information collected at the test site (the ED at the Children's Hospital of Eastern Ontario), where only two clinical attributes referenced in the CPG were among those routinely collected. Thus, augmenting the severity evaluation by a decision model suited to the local setting seems to be crucial for successful utilization of the CPG. Action steps may also require customization by matching suggested treatments and management options with local resources and practices [7], however this is beyond the scope of our research.

There has been extensive research on developing decision models (mostly in form of clinical scores or indexes) for pediatric asthma to help predict the severity of an exacerbation. Literature reviews identify more than sixteen such models [8, 9]. Examples include the Asthma Severity Scale (ASS) [10], Clinical Asthma Score (CAS) [11], Pediatric Asthma Severity Score (PASS) [12], Pediatric Respiratory Assessment Measure (PRAM) [13], and Pulmonary Score (PS) [14]. Unfortunately, they suffer from the same shortcomings as any asthma CPG – they rely on clinical attributes that are not routinely used or that cannot be collected for all patients (e.g., results of pulmonary function tests that are applicable to older children [15]) and require complete information characterizing the patient's state.

Due to these limitations we decided to develop a new decision model for predicting the severity that relies on clinical data collected at the local setting and is more flexible in terms of input requirements (it works with information limited to what physicians have deemed necessary in given circumstances). We followed a discovery-driven approach to model development [16] and created it from retrospective chart data. We preprocessed the data to take into account the existence of missing values, inconsistent categorizations, and contextual dependencies between values of some of the attributes. Then we applied machine learning to construct a tree-based decision model from the preprocessed data. The format of a decision tree is common for representing clinical algorithms, and thus familiar for physicians and relatively easy to interpret [17, 18].

The proposed decision model is intended to be used around two hours after starting the ED management process, which is consistent with other asthma studies [19-21] reporting that the results of post-treatment assessments have better predictive capabilities. Following discussions with ED physicians we collapsed the original severity levels into binary classes of mild and moderate/severe because of the clinical importance of early differentiation between the relatively benign condition (mild) and the others (moderate/severe), which require more aggressive intervention. According to clinical practice, both moderate and severe patients should receive systemic steroids, while mild patients should not. Evidence demonstrates that early administration of steroids results in earlier discharge of the patient [22], hence, early and accurate identification of moderate/severe patients in the ED should improve patient outcomes and operational efficiencies. Other differences in management (e.g., anticholinergics) between moderate and severe patients have less impact on the clinical outcome of the patient.

The paper is organized as follows. In Section 2 we describe the retrospective chart study and the process of developing and validating our decision model with focus on data preprocessing. In Section 3 we give a description of the analysis including the characteristic of the collected data, the structure of the developed model and the results of its validation. Finally, we conclude with a discussion in Section 4.

2. Materials and Methods

2.1. Study Setting and Population

This study was conducted at the Children’s Hospital of Eastern Ontario (CHEO) (Ottawa, Ontario, Canada). CHEO is a tertiary-care pediatric teaching hospital affiliated with the University of Ottawa serving patients up to 18 years of age. The ED has 53,000 annual patient visits (approximately 2,800 of them are visits for asthma – 2005/06 data) and is staffed 24-hours per day by specialty-trained Pediatric Emergency Medicine physicians along with fellows, residents and medical students.

Asthma management at CHEO includes a critical pathway [23] that outlines the standardized assessments and treatments patients should undergo to achieve sufficient reduction in respiratory distress prior to discharge home. The pathway is used as the primary nursing documentation tool and becomes part of the patient record. Medical directives are in place allowing the triage nurse to initiate bronchodilator treatments prior to physician assessment, and preprinted order sheets facilitate further treatment and investigation orders by physicians conforming to best evidence.

Records from children 1-17 years of age presenting to the ED between November 1, 2000 and July 30, 2004 for an asthma exacerbation were initially identified from ICD-10 coded discharge diagnoses. The study was approved by the CHEO Research Ethics Board.

2.2. Data Abstraction

A trained data abstractor reviewed each identified asthma visit using standardized inclusion and exclusion criteria (Table 1). When a patient had multiple asthma visits, index visits were identified as the first visit for an exacerbation, with the requirement that this exacerbation be distinctly unique using a two-month washout period from prior exacerbations documented.

For each index visit, the abstractor collected values of clinical attributes describing past history, history of current exacerbation, triage assessment, repeated assessments and final disposition. The complete list of collected attributes is given in Table 2.

Finally, the data abstractor, in consultation with one of the investigators (KF), assigned each visit to one of three groups (mild, moderate and severe) using the duration of the visit and presence of relapse visits as a proxy, according to strict, pre-defined criteria. This allowed us to identify those cases where the initial discharge decision was premature (e.g., the patient was incorrectly discharged and required another ED visit within a few days) and to change the severity assignment accordingly. Then the moderate and severe categories were collapsed together into the moderate/severe one, and we used this final severity assignment (mild or moderate/severe) as the gold standard (a correct decision that was or should have been made) when evaluating performance of the decision models.

2.3. Model Development

2.3.1. Experimental Design

The design of the experiment to develop a tree-based decision model is given in Figure 2. It follows a design presented in [24], where the authors described the process of developing and validating a tree-based decision model to identify high-risk elderly intensive care unit patients. According to this design, a data set is partitioned into developmental and validation sets. The developmental set is used to identify the best decision model in 10-fold cross validation process and the validation set is applied to validate the performance of a selected model.

Our experiment started with attribute filtering where from the list given in Table 2 we excluded attributes with a significant number of missing values – because of the retrospective nature of data abstraction process we were not able to conclude about the reasons for missing values and to impute them (e.g., with "normal" or "typical" values).

In step 2 the entire data set was partitioned into the developmental and validation sets according to the date of visit – records corresponding to visits before a selected date were included in the developmental set, and the remaining records were assigned to the validation set. This allowed us to mimic clinical practice, where a model would be constructed from data collected up to a certain point, and verified afterwards. Such an approach, being a special case of split-sample validation, has resulted in realistic validation of a decision model.

After partitioning the data we proceeded to find the best tree-based decision model (steps 3 – 6). As stated in [25] identification of the best decision model "is conceptually a search process: the algorithms used for their construction are searching a model space for the model that is most appropriate". In our research we used the C4.5 algorithm [26] implemented in the WEKA system [27] to construct decision models. The search space was defined by considering different approaches to preprocessing of the data and various complexity levels of potential decision models.

The C4.5 algorithm follows the divide-and-conquer approach to decision tree induction. It recursively partitions the data into subsets according to splits defined by attributes and their values. For nominal attributes splits correspond to all their possible values and for numeric attributes only two-way (or binary) splits based on a certain threshold value are considered. To select the best split, the algorithm uses two measures based on entropy – the information gain and the gain ratio. A split is selected if it maximizes the gain ratio, providing its information gain is above the average for all considered splits to compensate for highly branching splits that may have been favored by the gain ratio. The partitioning process stops if the data cannot be split any further.

In order to deal with missing values C4.5 does not introduce any additional specialized splits (like surrogate splits in CART [28]). Instead records where the value of the splitting attribute is missing are fragmented into so-called *fractional* records [29]. Fractional records are proportionally distributed among outcomes of a split (i.e., outgoing branches) using a weighting schema that is based on the number of learning records with known values that followed each branch. During construction of a tree, these fractional records are included in computing information gain and gain ratio when selecting subsequent splits. During classification of an unseen record with missing values, fractional records are used to calculate a class probability distribution, and then the most probable class is selected as the predicted one.

A tree constructed by recursive partitioning is very likely to overfit the learning data, and thus to perform poorly on unseen records. To address this problem C4.5 uses postpruning. After growing a full tree it checks specific splitting nodes from the bottom up and decides whether they should be postpruned (replaced by a leaf node or raised up) or not. The decision is made on the basis of estimated error rate – a node is postpruned if it leads to a lower estimate. The extent to which a tree is pruned is controlled by the confidence factor – a parameter that translates into confidence limits used to estimate the error rate. The default value for the confidence factor is 25% and decreasing it results in more aggressive pruning and a smaller size (in terms of the number of nodes) of a tree.

Possible decision models were constructed and evaluated in steps 3 – 6. As it is normally accepted, we used the 10-fold cross validation (10-fcv) schema. The developmental set was randomly split into 10 mutually exclusive subsets. Nine subsets were then combined into a learning set used to construct a decision model, and the 10th subset was used to evaluate the performance of the model. This was repeated 10 times, so each of the subsets was used once for testing and 9 times for learning. For more reliable estimates of performance we repeated 10-fcv 10 times [27] using a different random seed to split the developmental set in each run of cross validation, and we averaged the obtained 100 evaluations to get the final estimates.

In step 3 we preprocessed learning sets to address their undesirable characteristics that may have negatively impacted the quality of constructed models. We dealt with incompleteness, inconsistent categorizations and contextual dependencies between attributes (all these problems are common in medical data [30]). Specifically, we filtered questionable patient records and contextually normalized age-dependent attributes – the applied techniques are described in details in Section 2.3.2 and 2.3.3, respectively. In order to assess the effect of each of these two techniques on the performance of constructed models, we followed the concept of a factorial experiment, with the two binary factors corresponding to the use of two preprocessing techniques.

In step 4 we used the C4.5 algorithm to construct four possible tree-based decision models. They were built for a reduced number of attributes (as we indicated in the beginning, attributes with significant number of missing values were excluded from the analysis) and using contextually normalized values of selected attributes, where applicable. The models, labeled M1 to M4, corresponded to alternative preprocessing options applied in step 3 and are briefly characterized in Table 3. We controlled the complexity of these models by changing values of the confidence factor from 25% to 1% - such range followed the suggestion from [31] and it resulted in models of decreasing sizes.

In step 5 possible decision models were evaluated on the testing sets by comparing their predictions to the gold standard. During evaluation we considered two performance measures – area under Receiver Operating Characteristics (ROC) curve (AUC) [32] and sensitivity (for the cut-off of 0.5). AUC represents the probability that a decision model will rank a randomly chosen record from the critical (positive) category higher than a randomly chosen negative instance [33]. When computing these measures we considered moderate/severe to be a positive category (critical class), and mild to be a negative one.

In step 6 we computed the final estimates of performance and selected the best model that maximized AUC (primary criterion) and sensitivity (secondary criterion). We focused on the measures characterizing the discriminative abilities of a model as it was in line with the overall goal of our research – early differentiation between mild and moderate/severe patients.

Finally, in step 7 the best model selected in the previous step was recreated using the entire developmental set and it was validated using the validation set. To better characterize its predictive performance we expanded the set of performance measures to include overall accuracy, specificity, positive predictive value (PPV), negative predictive value (NPV), all computed for the cut-off of 0.5, and the Brier score, computed as the mean of the squared errors of the probability predictions [24]. The latter is a measure of calibration and thus it complements AUC, which is a measure of discrimination.

2.3.2. Record Filtering

Record filtering was inspired by research showing the positive impact of removing records with missing values on the prediction performance [34, 35]. Instead of using a simple record deletion technique [36] we employed expert knowledge to find “questionable” records. We used the PRAM score [13] as a proxy for such knowledge. PRAM is an evaluative score that uses 5 clinical attributes (suprasternal retractions, scalene contraction, air entry, wheezing and oxygen saturation) to derive an evaluation on a 12-point scale. Although it could not be applied directly to our data, with help of an emergency physician (KF) we developed a set of mapping rules to compute scores for 4 out of 5 attributes considered in PRAM. These rules were then applied to calculate so-called “modified” PRAM (M-PRAM) scores. Since triage (pre-treatment) assessments are reported to be not correlated with patient outcome [21], we applied M-PRAM to repeated (post-treatment) assessments.

The rules for calculating M-PRAM are listed in Table 4. One of the rules corresponding to wheezing indicates invalid combination of values – in such case M-PRAM cannot be computed. Moreover, two rules corresponding to air entry rely on the severity category limiting the applicability of M-PRAM to retrospective data only.

We labeled a record to be “questionable” if it was impossible to compute M-PRAM (because of missing values or their invalid combinations) or if a record associated with moderate/severe exacerbation received M-PRAM score equal to 0 (while it should be 4 or more [13]). The latter allowed us to exclude those records, where the final outcome was clearly inconsistent with recorded findings. We decided not to filter records associated with mild exacerbations even if they had high M-PRAM scores because the misclassification from mild to moderate/severe is a less serious mistake than misclassification in the opposite direction.

2.3.3. Contextual Normalization

The retrospective data included four context-sensitive attributes – heart rate and respiratory rate checked during triage and repeated assessments (TRI_HEART_RATE, TRI_RESP_RATE, REP_HEART_RATE and REP_RESP_RATE, respectively). Each of these attributes had to be considered in the context of the patient's age (REG_AGE), e.g., triage respiratory rate of 28 was normal for a 2-year-old, but abnormal for a 7-year-old. Usually values of such attributes are normalized according to approved medical norms for specific age groups. Such an approach was used in ASS [10], however reported results were not satisfactory.

In our analysis we used a data-driven normalization [37], where values of attributes were normalized using baseline values observed in the same context in a data set. In order to do so we took the mild category as the baseline and for each normalized record we identified records of mild patients with the most similar age (in other words, the nearest "mild neighbors" according to age). In our experiment we considered the size of the baseline neighborhood ranging from 5 to 9, which was inspired by results from [37]. For the nearest baseline neighbors we calculated mean values and standard deviations (SD) for the four context-dependent attributes and used them in the following formula (where the raw value denotes a value before normalization) to calculate their normalized values:

$$\text{normalized value} = (\text{raw value} - \text{mean})/\text{SD}.$$

Thus, a normalized value measures a relative difference between a raw value for a normalized record and a mean value for mild records in the same age group.

Finally, we removed the age attribute from normalized records because it became redundant. Moreover, age alone is not a good predictor of the severity of asthma exacerbation [38].

3. Results

3.1. Collected Data

During the retrospective chart study we extracted information from 775 index visits of 341 patients with asthma exacerbations. They were used to develop the initial data set composed of 362 records. The number of records was smaller than the number of index visits because records were created only for those visits that had a documented repeated assessment at 2 hours ± 20 minutes after triage (i.e., between 100 and 140 minutes). The basic characteristics of the initial data set, including descriptions of the developmental and validation sets are presented in Table 5.

As expected, the majority of records were incomplete. Although all attributes transcribed from paper charts appear on the emergency triage assessment record and the critical pathway used in the ED, many of them were not consistently recorded. Information about missing values is given in Table 6. Ten attributes that had more than 50% of missing values were excluded from the analysis, thus it was conducted on data described with 32 attributes. Although the usefulness of attributes with more than 15% of missing values is reported to be questionable [36], setting a lower threshold would have resulted in removing too much information from the data.

We used the date of October 1, 2003 to partition the data set into the developmental (prior) and validation (after) parts. Selecting this date allowed us to include in the validation set records of visits from fall and winter when usually the number of asthma exacerbations increases, especially for the youngest age group (1-4 years) [39].

3.2. Developed Model

To develop a tree-based decision model we followed the process described in Section 2.3.1. We successfully completed steps 1 –5, however, in step 6 we were not able to confidently select the best model. Table 7 lists evaluation results for the best decision models indentified in step 6 and corresponding to four data preprocessing options. The table also includes values of parameters controlling the contextual normalization (where applicable) and the complexity for resulting models. The estimates of AUC for all identified models were very close – the paired *t*-test with confidence level of 1% conducted on results of 10-fcv runs revealed no statistical differences among them. Therefore, we used estimates of sensitivity as the secondary criterion for selection. The highest sensitivity was observed for M3 and M4 models, and the paired *t*-test confirmed that the sensitivity estimates for M3 and M4 were statistically different from the estimates for M1 and M2, and that there were no statistical differences between the sensitivity estimates for M3 and M4. Thus, we selected these two models for further validation.

In step 7 we recreated the M3 and M4 models with controlling parameters identified in step 6 using the entire developmental set, and then we validated these models on the validation set. The results of validation are reported in Table 8. For all presented performance measures we constructed 95% confidence intervals (95% CI). CI for the Brier score was calculated using a *t*-value assuming normal distribution of the score [40], CI for AUC was calculated using the bootstrap percentile method [41] and for sensitivity, specificity, PPV, NPV and accuracy we used the Wilson's method [24]).

The M4 model turned out to be superior to M3 and its AUC surpassed the desired level of 0.8 [42] thus it was selected as the model for predicting severity of pediatric asthma exacerbations. The model is presented graphically in Figure 3. Its structure in terms of most discriminatory conditions is supported by [8, 19], where the relevance of wheezing and retractions for predicting asthma severity was emphasized. The model uses normalized heart rate and respiratory rate recorded during the repeated assessment (REP_HEART_RATE and REP_RESP_RATE), which further amplifies a good correlation between post-treatment assessments and the predicted clinical outcome [20, 21]. It also utilizes information about prior assessment in the chest clinic (CHEST_CLINIC). The numbers in the leaf nodes indicate how many records from the developmental set were captured in these nodes. The first number shows the total number of records captured by a node, while the second corresponds to the number of misclassified records (i.e., records for which the category was different than the one indicated in the node). Fractional numbers result from the way the C4.5 algorithm handles missing values (i.e., from introducing fractional records).

The M4 model includes attributes from the asthma clinical pathway that are routinely collected in the ED. Considering physicians' familiarity with these attributes this model should be easy to understand and interpret. Moreover, as physicians need to manage asthmatic patients according to the clinical pathway, the use of the M4 model does not force physicians to collect additional patient data.

To further evaluate the reported results we also constructed a logistic regression model using the developmental set and validated it on the validation set. Its performance, reported in Table 8, was worse than the performance of the M4 model on all measures (Figure 4 presents ROC curves for both models).

4. Discussion

The goal of our research was to customize asthma CPG by including site-specific information. We achieved this by proposing a new decision model for predicting the severity of exacerbation that relies on locally available information and provides predictions in the absence of some values. The model is intended to be used at 2 hours after triage and distinguishes between mild and moderate/severe exacerbations that correspond to two major treatment options.

Considering that learning about decision trees is a decision-making component of medical curriculum, a decision tree model generated from data is familiar construct for the physicians [17, 18]. Moreover, positive experience with tree-based models for diagnosing asthma and predicting hospital admission for asthmatic patients was reported by others [18, 43], who evaluated different decision models discovered from data (neural networks, linear discriminant functions, logistic regression).

Bishop et al. [10] and Chey et al. [38] also created models to predict severity of asthma exacerbations and reported predictive performance of their models. We were not able to compare our model with their results because of data incompatibilities and different ways of measuring the model's predictive performance. We used an objectively verified gold standard while calculating sensitivity and specificity (84% and 71% respectively), while ASS developed by Bishop et al. [10] (when tested on the developmental data) had sensitivity of 97% and specificity of 50% in comparison to physicians' predictions (such predictions should not be considered to be a gold standard). The logistic regression model proposed by Chey et al. [38] was tested on a validating hold-out sample but its sensitivity and specificity (88% and 89% respectively) were also calculated in comparison to physicians' predictions.

To address undesirable characteristics of the retrospective data we used two preprocessing techniques – filtering questionable records identified with help of M-PRAM and contextual normalization of age-sensitive attributes. All possible combinations in factorial design were evaluated using multiple runs of 10-fold cross validation. The model developed from the data preprocessed with a help of both techniques demonstrated the best performance when validated on the independent validation set. It is worth noting that good performance was observed despite some of the inputs being incomplete (some attributes used by our model, e.g., REP_HEART_RATE or REP_RESP_RATE, had more than 30% missing values in the validation set).

Our research has some limitations. First, we were able to conduct only a retrospective evaluation. A more complete evaluation should also include comparing the performance of the model with prediction performance of ED physicians on the same validation set (the quasi-Touring test [44]). Unfortunately, we were not able to extract required physicians' information (severity prediction at 2 hours) from charts and it was unrealistic to ask busy ED physicians to analyze more than 100 validation records. Another limitation results from the fact that our decision model relies on subjective attributes, i.e., wheezing or retractions (they are often referred to as "soft" clinical data [45]). Thus, in order to assess their variability we should have analyzed inter-observer agreement as suggested in [8]. However, such analysis was not possible in a retrospective chart study. Finally, the analyzed data set has a limited size (362 records) what may have limited the generality of the constructed decision model and the results of its validation.

Despite these limitations, the results of our research can be generalized. We demonstrated that it is possible to develop a good customized predictive decision model from messy clinical data, provided that it has been preprocessed. Since the model relies on locally collected and available clinical information and is flexible in terms of input requirements, it should facilitate the routine use of a CPG. If necessary, our model can be also used as a standalone decision tool.

5. Acknowledgements

The authors would like to thank William Klement and Morvarid Sehatkar from the School of Information Technology and Engineering, University of Ottawa for help with data analysis and Chris Drummond and Peter Turney from the National Research Council of Canada for advice about contextual normalization. The authors also would like to thank anonymous reviewers for helpful comments and suggestions.

Research reported here was supported by NSERC-CIHR grant from the CHRP program.

6. References

1. McDermott, M. F., Grant, E. N., Turner-Roan, K., Li, T., and Weiss, K. B., Asthma care practices in Chicago-area emergency departments. Chest 116(4 Suppl 1): 167-173, 1999.
2. Barnett, P. J., and Oberklaid, F., Acute asthma in children: evaluation of management in a hospital emergency department. Med. J. Aust. 154(11): 729-733, 1991.
3. Scribano, P. V., Lerer, T., Kennedy, D., and Cloutier, M. M., Provider adherence to a clinical practice guideline for acute asthma in a pediatric emergency department. Acad. Emerg. Med. 8(12): 1147-1152, 2001.
4. National Heart, Lung and Blood Institute, Guidelines for the diagnosis and management of asthma. Available from http://www.nhlbi.nih.gov/guidelines/asthma/index.htm. Last accessed on September 15, 2008.
5. Canadian Association of Emergency Physicians, Guidelines for emergency management of paediatric asthma. Available from http://www.caep.ca/template.asp?id=3272F9E47A064ED4820C829B15BB1BCD. Last accessed on September 15, 2008.
6. Shiffman, R. N., Michel, G., Essaihi, A., and Thornquist, E., Bridging the guideline implementation gap: a systematic, document-centered approach to guideline implementation. J. Am. Med. Inform. Assoc. 11(5): 418-426, 2004.
7. Waitman, L. R., and Miller, R. A., Pragmatics of implementing guidelines on the front lines. J. Am. Med. Inform. Assoc. 11(5): 436-438, 2004.
8. van der Windt, D. A., Nagelkerke, A. F., Bouter, L. M., Dankert-Roelse, J. E., and Veerman, A. J., Clinical scores for acute asthma in pre-school children. A review of the literature. J. Clin. Epidemiol. 47(6): 635-646, 1994.
9. van der Windt, D. A., Promises and pitfalls in the evaluation of pediatric asthma scores. J. Pediatr. 137(6): 744-746, 2000.
10. Bishop, J., Carlin, J., and Nolan, T., Evaluation of the properties and reliability of a clinical severity scale for acute asthma in children. J. Clin. Epidemiol. 45(1): 71-76, 1992.

11. Parkin, P. C., Macarthur, C., Saunders, N. R., Diamond, S. A., and Winders, P. M., Development of a clinical asthma score for use in hospitalized children between 1 and 5 years of age. J. Clin. Epidemiol. 49(8): 821-825, 1996.
12. Gorelick, M. H., Stevens, M. W., Schultz, T. R., and Scribano, P. V., Performance of a novel clinical score, the Pediatric Asthma Severity Score (PASS), in the evaluation of acute asthma. Acad. Emerg. Med. 11(1): 10-18, 2004.
13. Ducharme, F. M., Chalut, D., Plotnick, L., Savdie, C., Kudirka, D., Zhang, X., Meng, L., and McGillivray, D., The Pediatric Respiratory Assessment Measure: a valid clinical score for assessing acute asthma severity from toddlers to teenagers. J. Pediatr. 152(4): 476-480, 2008.
14. Smith, S. R., Baty, J. D., and Hodge, D., Validation of the pulmonary score: an asthma severity score for children. Acad. Emerg. Med. 9(2): 99-104, 2002.
15. Gorelick, M. H., Stevens, M. W., Schultz, T. R., and Scribano, P. V., Difficulty in obtaining peak expiratory flow measurements in children with acute asthma. Pediatr. Emerg. Care 20(1): 22-26, 2004.
16. Hanson, C. W., and Marshall, B. E., Artificial intelligence applications in the intensive care unit. Crit. Care Med. 29(2): 427-435, 2001.
17. Podgorelec, V., Kokol, P., Stiglic, B., and Rozman, I., Decision trees: an overview and their use in medicine. J. Med. Syst. 26(5): 445-463, 2002.
18. Grassi, M., Villani, S., and Marinoni, A., Classification methods for the identification of 'case' in epidemiological diagnosis of asthma. Eur. J. Epidemiol. 17(1): 19-29, 2001.
19. Becker, A. B., Nelson, N. A., and Simons, F. E., The pulmonary index: assessment of a clinical score for asthma. Am. J. Dis. Child. 138(6): 574-576, 1984.
20. Kelly, A. M., Kerr, D., and Powell, C., Is severity assessment after one hour of treatment better for predicting the need for admission in acute asthma? Respir. Med. 98(8): 777-781, 2004.
21. Schuh, S., Johnson, D., Stephens, D., Callahan, S., and Canny, G., Hospitalization patterns in severe acute asthma in children. Pediatr. Pulmonol. 23(3): 184-192, 1997.
22. Smith, M., Iqbal, S. M. S. I., Rowe, B. H., and N'Diaye, T., Corticosteroids for hospitalised children with acute asthma. Cochrane Database Syst. Rev. 1: CD002886, 2003.
23. Pearson, S. D., Goulart-Fisher, D., and Lee, T. H., Critical pathways as a strategy for improving care: problems and potential. Ann. Intern. Med. 123(12): 941-948, 1995.
24. de Rooij, S. E., Abu-Hanna, A., Levi, M., and de Jonge, E., Identification of high-risk subgroups in very elderly intensive care unit patients. Crit. Care 11(2): R33, 2007.
25. Abu-Hanna, A., and Lucas, P. J. F., Prognostic models in medicine. AI and statistical approaches. Methods Inf. Med. 40: 1-5, 2001.
26. Quinlan, R., C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, CA, 302 pages, 1993.
27. Witten, I. H., and Frank, E., Data mining: practical machine learning tools and techniques. 2nd ed. Morgan Kaufmann, San Francisco, CA, 525 pages, 2005.
28. Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A., Classification and Regression Trees. Wadsworth & Brooks, 1984.

29. Kohavi, R., and Quinlan, R., Decision-tree discovery. In: Klosgen W, and Zytkow JM, editors, Handbook of Data Mining and Knowledge Discovery. Oxford University Press, New York, p. 267-276, 2002.

30. Cios, K. J., and Moore, G. W., Uniqueness of medical data mining. Artif. Intell. Med. 26(1-2): 1-24, 2002.

31. Beck, J. R., Garcia, M. E., Zhong, M., Georgiopoulos, M., and Anagnostopoulos, G. C., A backward adjusting strategy and optimization of the C4.5 parameters to improve C4.5's performance. In: Wilson D, and Lane HC, editors, Proceedings of the 21st International Florida Artificial Intelligence Research Society (FLAIRS) Conference (FLAIRS 2008), Coconut Grove, Florida, USA, May 15-17, 2008, p. 35-40.

32. Faraggi, D., and Reiser, B., Estimation of the area under the ROC curve. Stat. Med. 21(20): 3093-3106, 2002.

33. Fawcett, T., An introduction to ROC analysis. Pattern Recogn. Lett. 27(8): 861-874, 2006.

34. Grzymala-Busse, J. W., and Hu, M., A comparison of several approaches to missing attribute values in data mining. In: Ziarko W, and Yao YY, editors, Rough sets and current trends in computing. Second International Conference, RSCTC 2000 Banff, Canada, October 16–19, 2000. Revised papers. Springer, Berlin, Heidelberg, p. 378-385, 2000.

35. O'Sullivan, D., Elazmeh, W., Wilk, S., Farion, K., Matwin, S., Michalowski, W., and Sehatkar, M., Using secondary knowledge to support decision tree classification of retrospective clinical data. In: Ras Z, Zighed D, and Tsumoto S, editors, Mining complex data. ECML/PKDD 2007 Third International Workshop, MCD 2007, Warsaw, Poland, September 17-21, 2007. Revised selected papers. Springer, Berlin, Heidelberg, p. 238-251, 2008.

36. Acuna, E., and Rodriguez, C., The treatment of missing values and its effect in the classifier accuracy. In: Banks D, House L, McMorris FR, Arabie P, and Gaul W, editors, Classification, clustering and data mining applications. Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15-18 July 2004. Springer, Berlin, Heidelberg, p. 639-648, 2004.

37. Turney, P., and Halasz, M., Contextual normalization applied to aircraft gas turbine engine diagnosis. J. Appl. Intell. 3: 109-129, 1993.

38. Chey, T., Jalaludin, B., Hanson, R., and Leeder, S., Validation of a predictive model for asthma admission in children: how accurate is it for predicting admissions? J. Clin. Epidemiol. 52(12): 1157-1163, 1999.

39. Crighton, E. J., Mamdani, M. M., and Upshur, R. E. G., A population based time series analysis of asthma hospitalisations in Ontario, Canada: 1988 to 2000. BMC Health Serv. Res. 1(1): 7-7, 2001.

40. Peek, N., Arts, D. G. T., Bosman, R. J., van der Voort, P. H. J., and de Keizer, N. F., External validation of prognostic models for critically ill patients required substantial sample sizes. J. Clin. Epidemiol. 60: 491-501, 2007.

41. Obuchowski, N. A., and Lieber, M. L., Confidence intervals for the receiver operating characteristic area in studies with small samples. Acad. Radiol. 5(8): 561-571, 1998.

42. Zimmerman, J. E., Kramer, A. A., McNair, D. S., and Malila, F. M., Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today’s critically ill patients. Crit. Care Med. 34(5): 1297–1310, 2006.

43. Li, D., German, D., Lulla, S., Thomas, R. G., and Wilson, S. R., Prospective study of hospitalization for asthma. A preliminary risk factor model. Am. J. Respir. Crit. Care Med. 151(3 Pt 1): 647-655, 1995.

44. Holt, G., Letter to the Editor. Clinical benchmarking for the validation of AI medical diagnostic classifiers. Artif. Intell. Med. 35(3): 259-260, 2005.

45. Kerem, E., Tibshirani, R., Canny, G., Bentur, L., Reisman, J., Schuh, S., Stein, R., and Levison, H., Predicting the need for hospitalization in children with acute asthma. Chest 98(6): 1355-1361, 1990.

Figure 1. General schema of the CAEP CPG

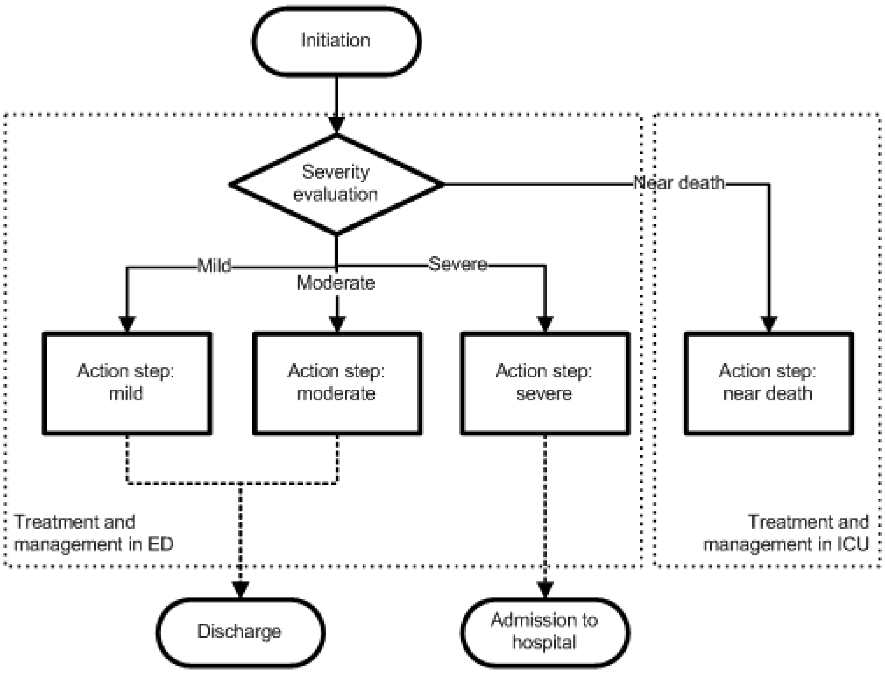

Figure 2. General experimental design

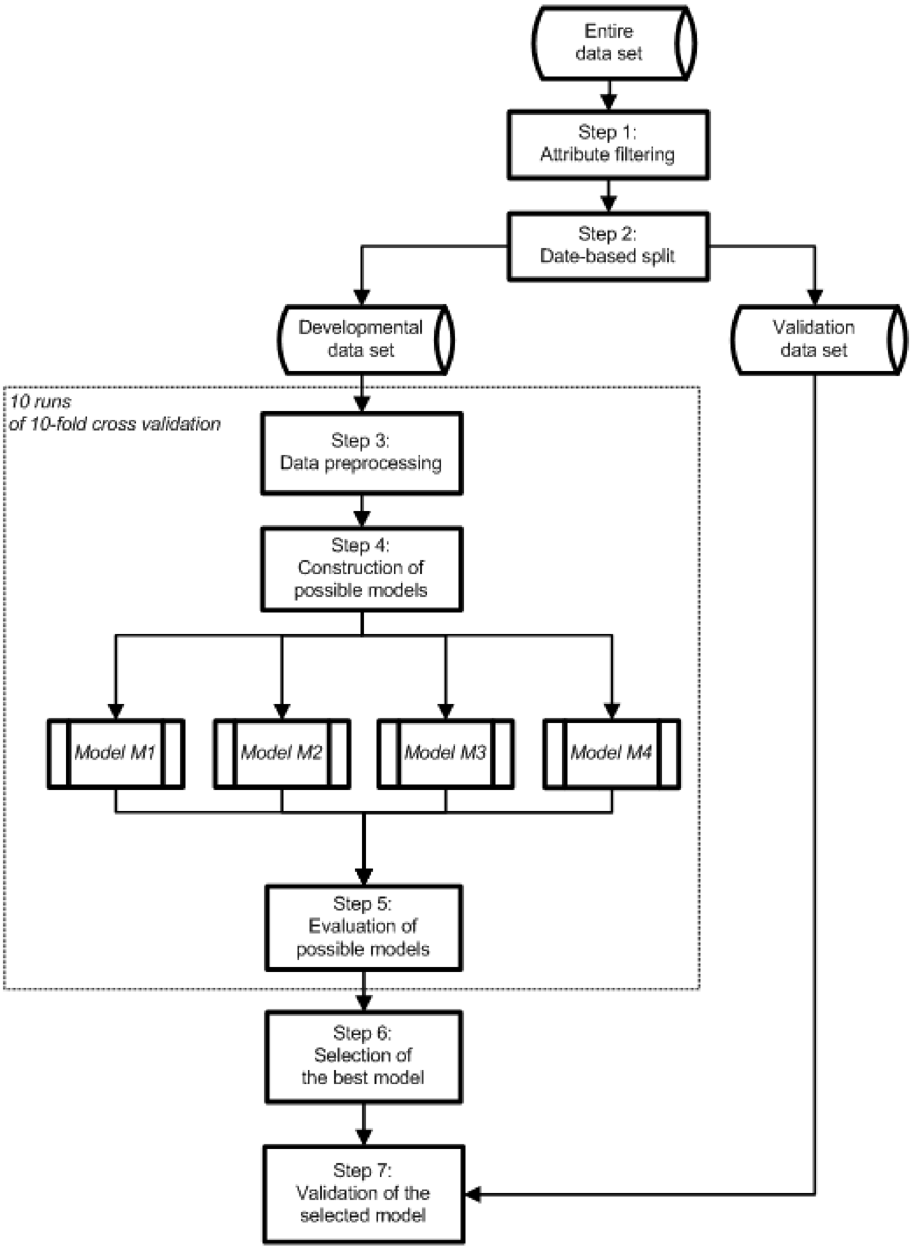
Entire data set
Step 1: Attribute filtering
Step 2: Date-based split
Developmental data set
Validation data set
10 runs of 10-fold cross validation
Step 3: Data preprocessing
Step 4: Construction of possible models
Model M1
Model M2
Model M3
Model M4
Step 5: Evaluation of possible models
Step 6: Selection of the best model
Step 7: Validation of the selected model

Figure 3. The structure of the M4 model

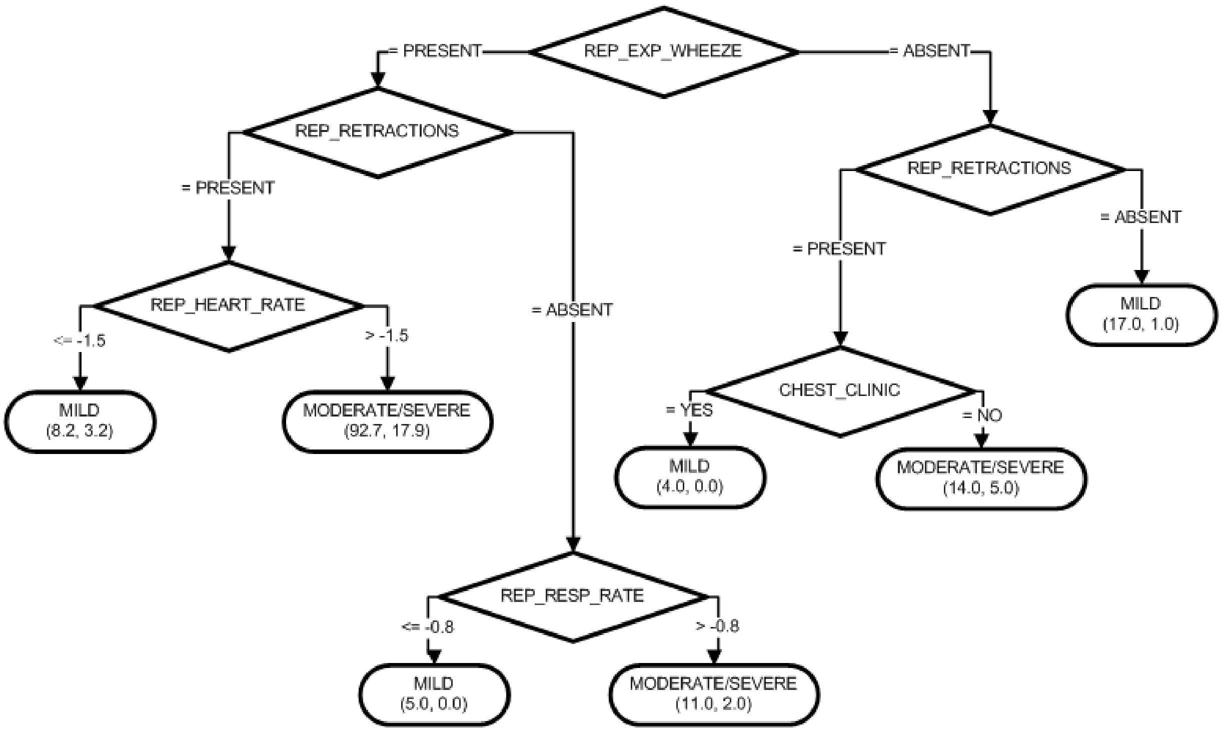

Figure 4. ROC curves for the M4 model and the logistic regression model

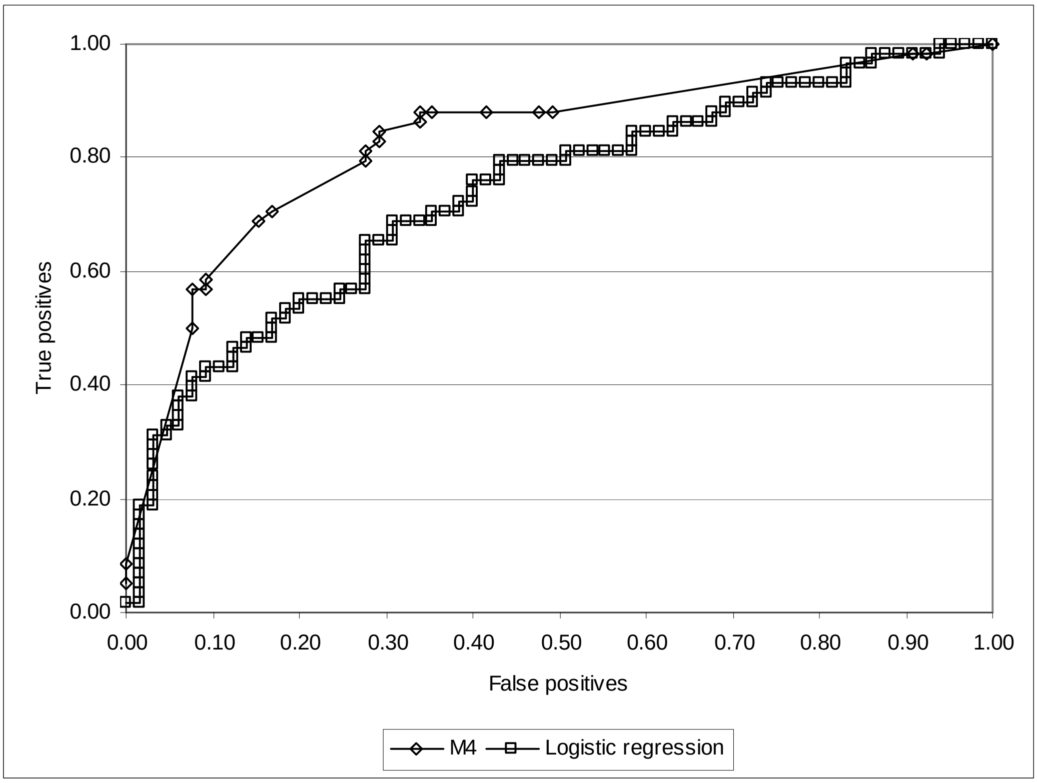

Table 1. Inclusion and exclusion criteria for retrospective chart study

Inclusion criteria

1. Patient age between 1 and 17 years.
2. Pre-existing diagnosis of asthma or reactive airways disease, as reported to the triage nurse or physician. Patients must have been previously prescribed inhaled bronchodilator therapy for at least one previous episode of wheeze, cough, or shortness-of-breath.
3. Presenting complaint of wheeze, cough, shortness-of-breath, or difficulty breathing requiring bronchodilator therapy in the ED with an ED discharge diagnosis or inpatient admission diagnosis of asthma or reactive airways disease.

Exclusion criteria

1. Patients receiving oral steroids chronically for asthma or any other illness
2. Patients receiving oral steroids for more than 48 hours prior to their ED visit for an acute exacerbation.
3. Patients with co-existing pulmonary conditions, cardiac illnesses, gastroesophageal reflux, chronic aspiration, or neuromuscular disease.
4. Patients presenting for medication refills or other non-urgent reasons related to asthma, and not requiring ED treatment.
5. Patients diagnosed with bronchiolitis

Table 2. Attributes transcribed from charts

#	Attribute	Code	Possible values
	Physician evaluation		
1	Age at registration	REG_AGE	numerical (years)
2	Primary care	PRIM_CARE	family doctor, pediatrician, other, none
3	Previous assessment in chest clinic	CHEST_CLINIC	yes, no
4	Current inhaled steroids	CURR_INH_STEROID	< 1 week, 1 – 4 weeks, ≥ 4 weeks, as necessary, none
5	Age at first symptoms	AGE_FIRST_SX	numerical (years)
6	Previous oral steroids	PREV_ORAL_STEROID	< 1 month, 1 – 3 months, 3 – 12 months, ≥ 12 months, none
7	Previous ED visits last year	PREV_ED_LAST_YEAR	1 visit, 2 visits, 3 visits, ≥ 4 visits, none
8	Previous admission	PREV_ADM	floor, ICU, none
9	Smokers in environment	ENV_SMOKE	yes, no
10	Dander in environment	ENV_DANDER	yes, no
11	Carpets in environment	ENV_CARPETS	yes, no
12	Allergies to environment	ALLG_ENV	yes, no
13	Allergies to pets	ALLG_PETS	yes, no
14	Allergies to food	ALLG_FOOD	yes, no
15	History of atopy	PTHX_ATOPY	yes, no
16	Family history of asthma	FMHX_ASTHMA	yes, no
17	Allergy exposure	ALLG_EXP	yes, no
18	URTI symptoms	URTI_SX	yes, no
19	Fever	FEVER	yes, no
20	Duration of symptoms	DUR_ASTHMA_SX	numerical (hours)
21	Bronchodilators in the last 24h	VENT_LAST_24H	numerical
22	Arrival to the ED	ARRV_ED	ambulance, parents
	Triage assessment		
23	Temperature	TRI_TEMP	numerical (Celsius degrees)
24	Respiratory rate	TRI_RESP_RATE	numerical (breaths per minute)
25	Heart rate	TRI_HEART_RATE	numerical (bits per minute)
26	Oxygen saturation	TRI_SAO2	numerical (%)
27	Air entry	TRI_AIR_ENTRY	good, reduced
28	Distress	TRI_DISTRESS	none, mild, moderate, severe
29	Skin color	TRI_COLOR	pink, pale, dusky
30	Expiratory wheeze	TRI_EXP_WHEEZE	present, absent
31	Inspiratory wheeze	TRI_INSP_WHEEZE	present, absent
32	Retractions	TRI_RETRACTIONS	present, absent
	Repeated assessment		
33	Temperature	REP_TEMP	numerical (Celsius degrees)
34	Respiratory rate	REP_RESP_RATE	numerical (breaths per minute)
35	Heart rate	REP_HEART_RATE	numerical (bits per minute)
36	Oxygen saturation	REP_SAO2	numerical (%)
37	Air entry	REP_AIR_ENTRY	good, reduced
38	Distress	REP_DISTRESS	none, mild, moderate, severe
39	Skin color	REP_COLOR	pink, pale, dusky
40	Expiratory wheeze	REP_EXP_WHEEZE	present, absent
41	Inspiratory wheeze	REP_INSP_WHEEZE	present, absent
42	Retractions	REP_RETRACTIONS	present, absent

Table 3. Decision models

		Record filtering	
		No	Yes
Contextual normalization	No	M1	M3
	Yes	M2	M4

Table 4. Expert rules for calculating M-PRAM

PRAM attribute	Conditions	Score
Suprasternal retractions	REP_RETRACTIONS = absent, REP_AIR_ENTRY = good	0
	REP_RETRACTIONS = absent, REP_AIR_ENTRY <> good	1
	REP_RETRACTIONS = present	2
Air entry	REP_AIR_ENTRY = good	0
	REP_AIR_ENTRY = reduced, category = mild	1
	REP_AIR_ENTRY = reduced, category = moderate/severe	3
Wheezing	REP_EXP_WHEEZE = absent, REP_INSP_WHEEZE = absent	0
	REP_EXP_WHEEZE = absent, REP_INSP_WHEEZE = present	invalid
	REP_EXP_WHEEZE = present, REP_INSP_WHEEZE = absent	1
	REP_EXP_WHEEZE = present, REP_INSP_WHEEZE = present	2
Oxygen saturation	SaO2 >= 95%	0
	92% <= SaO2 <= 94%	1
	SaO2 < 92%	2

Table 5. Characteristics of the data sets

	Entire set		Developmental set		Validation set	
Category	Records [n]	Records [%]	Records [n]	Records [%]	Records [n]	Records [%]
Mild	163	45.0	98	41.0	65	52.8
Moderate/severe	199	55.0	141	59.0	58	47.2
Total	362	100.0	239	100.0	123	100.0

Table 6. Missing values in the data set

	Missing values [%]		
Code	Entire set	Developmental set	Validation set
REG_AGE	0.0	0.0	0.0
PRIM_CARE	0.3	0.0	0.8
CHEST_CLINIC	0.0	0.0	0.0
CURR_INH_STEROID	41.4	37.7	48.8
AGE_FIRST_SX	9.9	7.9	13.8
PREV_ORAL_STEROID	13.8	15.5	10.6
PREV_ED_LAST_YEAR	1.1	1.3	0.8
PREV_ADM	1.4	1.7	0.8
ENV_SMOKE	69.9	69.0	71.5
ENV_DANDER	70.7	69.0	74.0
ENV_CARPETS	83.7	83.3	84.6
ALLG_ENV	1.7	1.7	1.6
ALLG_PETS	1.4	1.3	1.6
ALLG_FOOD	1.1	1.3	0.8
PTHX_ATOPY	27.9	22.6	38.2
FMHX_ASTHMA	23.2	20.9	27.6
ALLG_EXP	74.0	74.5	73.2
URTI_SX	3.3	2.9	4.1
FEVER	9.7	9.6	9.8
DUR_ASTHMA_SX	5.8	5.4	6.5
VENT_LAST_24H	22.4	21.8	23.6
ARRV_ED	1.4	1.3	1.6
TRI_TEMP	20.7	22.6	17.1
TRI_RESP_RATE	7.5	9.2	4.1
TRI_HEART_RATE	1.7	2.1	0.8
TRI_SAO2	1.7	2.1	0.8
TRI_AIR_ENTRY	6.6	6.7	6.5
TRI_DISTRESS	59.7	57.3	64.2
TRI_COLOR	2.8	3.3	1.6
TRI_EXP_WHEEZE	65.2	75.3	45.5
TRI_INSP_WHEEZE	73.2	81.2	57.7
TRI_RETRACTIONS	61.9	72.0	42.3
REP_TEMP	85.6	87.9	81.3
REP_RESP_RATE	22.7	17.2	33.3
REP_HEART_RATE	25.4	19.2	37.4
REP_SAO2	20.2	15.9	28.5
REP_AIR_ENTRY	11.0	10.0	13.0
REP_DISTRESS	90.3	91.6	87.8
REP_COLOR	26.2	23.8	30.9
REP_EXP_WHEEZE	14.6	13.0	17.9
REP_INSP_WHEEZE	16.0	15.5	17.1
REP_RETRACTIONS	16.9	15.9	18.7

Table 7. Evaluation results on the developmental set

Measure	M1	M2	M3	M4
AUC ± SD	0.6464 ± 0.1209	0.6346 ± 0.1087	0.6477 ± 0.1301	0.6390 ± 0.1153
Sensitivity ± SD [%]	73.76 ± 11.30	74.54 ± 12.11	78.00 ± 11.82	79.64 ± 11.71
Tree size ± SD	11.6 ± 4.6	20.1 ± 8.0	17.7 ± 4.9	15.5 ± 5.3
Confidence factor	5%	15%	15%	15%
Number of baseline neighbors	-	9	-	8

Table 8. Validation results for the M3 and M4 models and the logistic regression on the validation set (95% CI)

Measure	M3	M4	Logistic regression
Brier score	0.2199 (0.1608; 0.2790)	0.1752 (0.1263; 0.2241)	0.2247 (0.1692; 0.2802)
AUC	0.7391 (0.6426; 0.8259)	0.8275 (0.7461; 0.8991)	0.7379 (0.6464; 0.8243)
Sensitivity [%]	75.86 (63.47; 85.04)	84.48 (73.07; 91.62)	68.97 (56.20; 79.38)
Specificity [%]	64.62 (52.48; 75.12)	70.77 (58.80; 80.42)	67.69 (55.61; 77.80)
PPV [%]	65.67 (53.73; 75.91)	72.06 (60.44; 81.32)	65.57 (53.05; 76.25)
NPV[%]	75.00 (62.31; 84.48)	83.64 (71.74; 91.14)	70.97 (58.71; 80.78)
Accuracy [%]	69.92 (61.31; 77.32)	77.24 (69.07; 83.75)	68.29 (59.62; 75.86)