Comparing predictions made by a prediction model, clinical score, and physicians: Pediatric asthma exacerbations in the emergency department

Ken J. Farion[1,2,3], Szymon Wilk[3,4], Wojtek Michalowski[3], Dympna O'Sullivan[3,5], Jelber Sayyad-Shirabad[6]

[1] Division of Emergency Medicine, Children's Hospital of Eastern Ontario
[2] Departments of Pediatrics and Emergency Medicine, University of Ottawa
401 Smyth Rd., Ottawa, Ontario, K1H 8L1 Canada

[3] MET Research Group, Telfer School of Management, University of Ottawa
55 Laurier Ave. E., Ottawa, Ontario, K1N 6N5 Canada.

[4] Institute of Computing Science, Poznan University of Technology
Piotrowo 2, 60-965 Poznan, Poland

[5] Center for Health Informatics, School of Informatics, City University London
Northampton Square, London, EC1V 0HB, United Kingdom

[6] School of Electrical Engineering and Computer Science, University of Ottawa
800 King Edward, Ottawa, Ontario, K1N 6N5, Canada

Keywords: Asthma; Pediatrics; Emergency Medical Services; Decision Support Techniques

Corresponding author:
Szymon Wilk
Institute of Computing Science
Poznan University of Technology
Piotrowo 2, 60-965 Poznan, Poland
E-mail: szymon.wilk@cs.put.poznan.pl
Tel.: +48 61 665 29 30
Fax: +48 61 877 15 25

Abstract

Background: Asthma exacerbations are one of the most common medical reasons for children to be brought to the hospital emergency department (ED). Various prediction models have been proposed to support diagnosis of exacerbations and evaluation of their severity.
Objectives: First, to evaluate prediction models constructed from data using machine learning techniques and to select the best performing model. Second, to compare predictions from the selected model with predictions from the Pediatric Respiratory Assessment Measure (PRAM) score, and predictions made by ED physicians.
Design: A two-phase study conducted in the ED of an academic pediatric hospital. In phase 1 data collected prospectively using paper forms was used to construct and evaluate five prediction models, and the best performing model was selected. In phase 2, data collected prospectively using a mobile system was used to compare the predictions of the selected prediction model with those from PRAM and ED physicians.
Measurements: Area under the receiver operating characteristic curve and accuracy in phase 1; accuracy, sensitivity, specificity, positive and negative predictive values in phase 2.
Results: In phase 1 prediction models were derived from a data set of 240 patients and evaluated using 10-fold cross validation. A naive Bayes (NB) model demonstrated the best performance and it was selected for phase 2. Evaluation in phase 2 was conducted on data from 82 patients. Predictions made by the NB model were less accurate than the PRAM score and physicians (accuracy of 70.7%, 73.2% and 78.0% respectively), however, according to McNemar's test it is not possible to conclude that the differences between predictions are statistically significant.
Conclusion: Both the PRAM score and the NB model were less accurate than physicians. The NB model can handle incomplete patient data and as such may complement the PRAM score. However, it requires further research to improve its accuracy.

1. Introduction

Asthma is the leading chronic disease in childhood and asthma exacerbations are one of the most common medical reasons for children to be brought to the hospital emergency department (ED). These visits, and the subsequent hospitalizations required by a large proportion of patients, account for more than 60% of all costs of asthma care [1].

Clinical practice guidelines (CPGs) for pediatric asthma, for example the EPR-3 CPG developed by the National Heart, Lung and Blood Institute [2], define four categories of asthma control severity, three categories of exacerbation severity (mild, moderate, and severe) and a subset of life-threatening severity. In this study we are not concerned with asthma control severity, as it looks at how symptomatic a chronically ill patient is on a day-to-day basis. Instead, we are concerned solely with the severity of exacerbations, excluding life-threatening situations. In the academic center where our study was conducted – the Children's Hospital of Eastern Ontario (CHEO) – patients with a mild exacerbation are usually discharged home after a brief course of treatment (typically less than 4 hours in the ED), patients with a moderate exacerbation undergo prolonged and more aggressive treatment in the ED or in an observation unit (typically 4-16 hours in total), and patients with a severe exacerbation receive maximal stabilization therapy before being transferred to an in-patient hospital bed for ongoing therapy (or spend typically over 16 hours in the ED). These lengths of stay are generally in line with data reported by others, for example [1,3].

It is important that the severity of a child's asthma exacerbation is determined as soon as possible, so that appropriate therapies are prescribed and provided in a timely fashion [4]. Underestimation of the severity of the exacerbation results in inadequate treatment so that discharge is delayed, or in premature discharge and possible need for a return visit. Conversely, overestimation results in patients receiving unnecessary therapies and unnecessarily occupying ED resources when they could be safely managed at home.

According to a review by Sanders and Aronsky [5], the development of decision models for detecting and evaluating asthma and its exacerbations is a very active research field. These models are usually developed from data using statistical or machine learning methods. There are studies [6-8] that compare the performance of these two groups of methods in a variety of settings. While some of the reported results are more conclusive than the others, a prevailing view is that it is difficult to establish general guidelines for selecting a specific method [9].

Building on our earlier research [10] in this paper we apply machine learning to develop models for predicting severity of asthma exacerbation for patients already diagnosed with asthma and presenting to the ED. Thus, we contribute to the research on computer-based support for asthma management as advocated by Sanders and Aronsky [5] and expand the line of research presented by Dexheimer et al. [11] and Sanders and Aronsky [12] by focusing on exacerbation severity assessment during the course of asthma management in the ED. Specifically, the purpose of our study is to answer the following research questions:

- What is the best performing prediction model (evaluation measures are presented in details in Section 3.2.1) that can be used to evaluate the severity of pediatric asthma exacerbations early in the ED visit?
- How does the prediction model's performance compare with predictions derived by the Pediatric Respiratory Assessment Measure (PRAM) score [13] and those made by physicians?

We consider the PRAM score as a comparator because it was in use in the ED of CHEO where the research took place.

2. Literature review

Sanders and Aronsky define four domains in biomedical informatics related to asthma management: detection and diagnosis, monitoring and prevention, patient education, and therapy of acute or chronic asthma [5]. In this review we are concerned with the first domain, as it is directly related to our research. Specifically, the biomedical informatics research on detection and diagnosis of asthma is concerned with two important and complementary lines of research: identifying asthma and its exacerbations, and evaluating their severity [5].

The first line of research is well represented by work conducted at the Vanderbilt University Medical Canter. Sanders and Aronsky [14] presented a Bayesian network model to identify patients eligible for an asthma CPG. The model was constructed using expert knowledge and its parameters were learned from data describing 3017 patient visits. The model was tested on an independent test set with 1006 visits, where it demonstrated the value for an area under the receiver operating characteristic (ROC) curve (AUC) of 0.96 (notions of the ROC and AUC are described in detail in Section 3.2.1). The Bayesian network was applied in the asthma detection system [1]. Among 1100 patients identified by the system as having asthma exacerbation, the diagnosis was confirmed for 704 patients, thus yielding an accuracy of 64%. Also Sanders et al. [15] described a rule-based algorithm for identifying asthma exacerbations in the ED. The rules were defined by the experts and used patient data

available in electronic form during triage. The algorithm was tested on a data set with 1835 ED visits and achieved a sensitivity of 44.8% and specificity of 91.6% (AUC was not reported in the paper).

Dexheimer et al. [11] expanded an earlier study by Sanders et al. [14] by applying several machine learning techniques to previously collected data. Specifically, they considered a neural network model, a support vector machine model, a Gaussian process (a probabilistic extension to the neural network) and an automatically constructed Bayesian network. When verified on the test set, these models demonstrated AUC ranging from of 0.94 (the neural network model) to 0.96 (the Gaussian process and the Bayesian network model). The AUC for the support vector machine was not reported.

Most research on evaluation of asthma exacerbation severity is concerned with asthma CPGs and asthma severity scores. For example, Powell et al. [16] described a study where they checked adherence of ED physicians classifying asthma exacerbation severity to the CPG developed by the National Asthma Council Australia. They reported that the adherence was only moderate (kappa = 0.48) with physicians underestimating the severity (the authors did not report the accuracy of severity predictions). Limited compliance of physicians with CPGs is reported also by Eccles et al. [17]. Another example of a CPG supporting exacerbation severity evaluation is the EPR-3 CPG [2]. It was implemented in a system for assessment asthma severity and management in a clinic setting [18]. The system was tested in a cohort study, where it provided accurate predictions of asthma severity for 63 out of 100 patients.

Selected asthma exacerbation severity scores have been reviewed by Birken et al. [19]. The authors pointed at the lack of formal development processes for scoring algorithms. Moreover, the review did not produce conclusive results with regards to the clinical value of asthma scores. While most scores are focused on responsiveness (changes in the score values correspond to the changes in the state of an asthmatic patient), there are some that have discriminative capabilities and can be used as severity predictors. Examples of such scores include PRAM [13] (discussed in Section 3.3) and Pediatric Asthma Severity Score (PASS) [20]. Prediction capabilities of these two scores were evaluated by Gouin et al. [21], who used the length of stay as a proxy of exacerbation severity and distinguished between patients who stayed in the ED < 6 hours, and those who stayed ≥ 6 hours and/or were admitted to the ward. Both PRAM and PASS were applied to 283 patients and the AUC for PRAM was 0.69 and for PASS it was 0.70.

A different approach to evaluating exacerbation severity is presented by Zolnoori et al. [22], who describe a system using fuzzy rules. There are two types of rules – evaluating rules that describe relationships between symptoms and laboratory results and the exacerbation severity, and meta-rules that control the inference process. Both types of rules rely on expert knowledge – they were derived by experts or from a medical literature. The system was tested on 25 asthmatic patients and its outcomes were compared to the outcomes of the physicians. While both matched perfectly (kappa = 1) it is important to note that no gold standard measure was used in this comparison.

A system for evaluating asthma severity was presented by Sefion et al. [23]. Their ADEMA system used an instance-based model that included 190 cases and in a leave-one out test it demonstrated a prediction accuracy of 61.0% for asthma severity and 71.2% for therapy respectively (no AUC values were reported in the paper).

Machine learning techniques were also applied in a study by Farion et al. [10], where a decision tree model was constructed and verified using data transcribed retrospectively from paper charts. The learning data set of 239 patient visits was preprocessed by removing questionable cases (a simplified version of the PRAM score was used for this purpose) and by contextual normalization of age-dependent numerical attributes. The decision model was tested on an independent test set comprised of 123 patient visits, where it demonstrated AUC of 0.83. A logistic regression model constructed and evaluated using the same data sets was less accurate and produced AUC of 0.74.

The study described in this paper adds to the body of knowledge on applying machine learning methods to build prediction models for asthma exacerbations. Moreover, to the best of our knowledge, it is the first study that directly compares machine learning prediction models with asthma exacerbation severity scores, thus providing a better insight into the strengths and weaknesses of both types of predictors.

3. Methods

3.1. Study setting

The study described in this paper was conducted at CHEO (Ottawa, Ontario, Canada). Our study site is a tertiary, academic pediatric hospital with about 60,000 annual ED visits. It has the only pediatric ED serving an estimated population of over 450,000 children in Eastern Ontario and Western Quebec, staffed with emergency physicians (staff EPs), fellows, and residents.

Information management in the ED is handled by an admission-discharge-transfer system (Epic) and an ED information system (Allscripts ED) that exchange information using HL7 messages. There is a secure wireless network that provides mobile access to these systems as well as to vital signs monitors.

3.2. Study design and population

The CHEO Research Ethics Board approved our study; the study population included pediatric asthma patients who satisfied the inclusion and exclusion criteria listed in Table I when they presented to the ED. The patients were categorized by exacerbation severity that distinguished between mild and moderate/severe exacerbations. Such categorization corresponds to different therapies required by patients with mild and moderate/severe exacerbations – the latter includes patients who are given systemic corticosteroids and anticholinergics as early as possible to reduce the length of stay and requirement for hospital admission [4]. Thus, therapeutically it is reasonable to combine moderate and severe exacerbations into a single category.

The study was conducted in two phases. In phase 1 we prospectively collected patient data using structured paper forms, applied machine learning techniques to the collected data in order to construct and evaluate several prediction models, and selected the best performing model. In phase 2 we collected a new set of patient data using a mobile system, applied the prediction model selected in phase 1 to this data set, and compared the predictive performance of the model to the performance of the PRAM score and physicians. Both phases are described in detail in Section 3.2.1 and 3.2.2 respectively.

In phase 1 the inclusion/exclusion criteria (see Table I) were applied by a triage nurse, and in phase 2 they were applied automatically to HL7 admission messages. Potentially eligible patients were flagged (color stickers in phase 1, electronically in phase 2) and enrolled at the discretion of attending physicians (staff EPs, fellows and residents). This enrollment was subject to many factors including current ED workload, physician's willingness to follow the study protocol, and external circumstances (such as the H1N1 regime in place). Approached patients (parents) were asked to provide informed verbal consent to participate in the study, and once consent was obtained, a patient was again re-checked by the physician for eligibility.

When collecting data (paper forms or a mobile system) attending physicians were instructed to collect information related to patient history, the current asthma exacerbation episode, and conducted assessments – primary (triage) or secondary (repeated). Clinical

attributes considered during data collection are presented in Table II. We need to note that values for the last group of attributes (#23-32) may have been recorded multiple times for all secondary assessments.

3.2.1. Phase 1

In phase 1 (conducted between November 2006 and May 2007), prospective data was collected using paper forms that were attached to charts of eligible patients. Primary assessments were recorded by the triage nurse first, assessing the patient on arrival; secondary assessments were collected at various times during the visit by the treating physician or nurse. Patient history was collected once by the physician. When feasible, paired observations were conducted by two observers to assess interobserver reliability. From the data collected or recorded in the patient`s chart, the PRAM score at triage was determined. Treating physicians were asked to rate the patient`s exacerbation severity and predicted length of stay, as mild (less than 4 hours), moderate (4-16 hours), or severe (greater than 16 hours or requiring admission).

A research assistant (RA) routinely audited the collected data to identify and address potential problems (i.e., lack of relevant data). At 10-14 days after the visit, each enrolled patient (or parent) was contacted by phone to inquire about the patient's condition. The complete patient chart was also independently reviewed. All this information was used to determine the final exacerbation severity category for the visit.

The collected data set was processed to minimize the number of missing values associated with the assessment attributes and to limit the number of secondary assessments associated with each patient. If a recorded assessment had missing data, we used data from the corresponding paired assessment to fill in these values. To identify the secondary assessment that would be used for analysis, we determined the average time from primary to first secondary assessment and calculated the standard deviation. For each patient`s data, the assessment closest to this average time was used, providing that it occurred within the time window of one standard deviation before or after the average [average time ± standard deviation]. Patient records without a secondary assessment within this window were discarded. Finally, from such data we created a data set described by values of 42 attributes (values of attributes #23-32 from Table II were given twice – for the primary and the secondary assessment).

This data set was used to construct prediction models and to evaluate their performance, where final severity of exacerbation verified through follow-up and

independent chart review was used as the gold standard. We considered the following prediction models constructed using machine learning techniques (these are models often considered in clinical problems [24]) and implemented by the WEKA system (version 3.6) [25]:

- A naive Bayes model (denoted as NB),
- A decision tree model (denoted as DT)
- An ensemble of decision tree models (denoted as E-DT),
- A support vector machine model (denoted as SVM)
- An instance-based model with 1 and 10 nearest neighbours (denoted as IB1 and IB10 respectively).

Decision trees in the DT and E-DT models were constructed using the C4.5 algorithm, their ensemble was built using a bagging strategy, and SVM was developed using the SMO algorithm. Witten et al. [26] provide a comprehensive and detailed description of algorithms and techniques applied in constructing relevant prediction models (the reader is particularly referred to Chapter 4, 6 and 8). For all prediction models we used default values of parameters provided by WEKA (only for the SVM model we used calibrated parameter values to ensure proper probability estimates for AUC). In a preliminary experiment we checked the preprocessing techniques proposed by Farion et al. [10], however, their usefulness when applied to prospective data turned out to be negligible, therefore we finally did not consider them.

All models were evaluated using the 10-fold cross validation scheme [26] – for more reliable results cross validation was repeated 10 times. We used AUC – area under the ROC curve [11] – as the primary evaluation criterion. The ROC curve is obtained by plotting sensitivity versus (1 – specificity) (we assumed moderate/severe category to be the positive class, and mild to be the negative one), and AUC measures how well the prediction model separates the classes and captures the trade-off between sensitivity and specificity. We also considered prediction accuracy as the secondary criterion.

Finally, we selected a prediction model that had the best AUC and prediction accuracy. In order to check if differences in performance between specific models were statistically significant, we used a paired Student's t-test. In this test we considered all possible pairs of prediction models and averaged results obtained by these models in individual folds. Evaluation and comparison of prediction models, including statistical tests were conducted using WEKA (version 3.6).

3.2.2. Phase 2

Phase 2 took place from February 2009 to March 2010, and similarly to phase 1, it was a prospective observational data collection. However instead of paper forms, physicians used a mobile system – MET3-AE – to record and collect patient information. The data collection was combined with another study conducted in the ED which aimed to evaluate the usability of MET3-AE and assess the motivation of physicians to use computer-based support at the point of care. The MET3-AE system and its use in our study are briefly described in Section 3.4.

Physicians collected information on patient history and current exacerbation, and results of primary and secondary assessments. Unlike in phase 1, a single observer approached each patient. This decision was dictated by the fact that it would be very difficult to have the same multiple observers doing patient reassessments. Each collected record was audited and patient follow-up was conducted similarly to phase 1 to establish a final exacerbation severity assessment that was considered as the gold standard for performance evaluation.

The collected data set was preprocessed to obtain the structure consistent with the data set from phase 1. For each patient we retained two assessments– the primary assessment and a secondary assessment closest to the time point calculated in phase 1.

Finally, we evaluated predictive performance of the selected prediction model, the PRAM score, and physicians by comparing their predictions to the gold standard (in the case of the PRAM score we used the threshold described in Section 3.3 to translate the total score into an exacerbation severity level). Due to the requirements of the AUC measure (numerical probability estimates) we were not able to calculate the AUC for physicians, therefore in phase 2 we decided to use sensitivity and specificity instead. These measures were supplemented with the overall prediction accuracy and with positive and negative prediction values (PPV and NPV respectively).

To check if there were statistically significant differences between predictions made by the selected model, the PRAM score, and physicians, we used the McNemar's test [27] implemented in the R system (version 2.14) [28]. In this test we considered all possible pairs of predictors, and for each patient record we compared predictions made by the selected predictors (NB model – PRAM, NB model – physician, PRAM – physician). As we considered individual records instead of their sets (folds) in this phase, we were not able to obtain numerical outcomes for specific records, thus the Student's t-test was not applicable.

3.3. PRAM score

The PRAM score was developed to assess asthma exacerbation severity among preschool-aged patients [29]. A subsequent prospective validation study demonstrated excellent performance for school-aged children as well [13], making it a reliable evaluation tool for the entire pediatric population. The PRAM score relies on a set of five clinical attributes (observations and tests) and it maps values of these attributes into a 4-point scale. Values of all attributes are required to calculate the score.

While the PRAM score is primarily meant to monitor patient state over time (change in the score reflects changes in airway obstruction), it can be used to categorize the severity of asthma exacerbations with the following thresholds: total score between 0-4 indicating mild exacerbation, total score between 5-8 indicating moderate exacerbation, and total score between 9-12 indicating severe exacerbation [7]. Following its evaluation, the authors of the score concluded that PRAM is a responsive but only moderately discriminative score [29]. As a result, revised thresholds have been adopted in the ED of CHEO. These revised thresholds for the total PRAM score are: 0-3 – mild exacerbation, 4-7 – moderate mild exacerbation, and 8-12 – severe exacerbation. In the our study we used these revised thresholds for the PRAM score when predicting asthma exacerbation severity, thus a total score between 0 and 3 was interpreted as mild exacerbation, and a score of 4 or above was translated into the moderate/severe exacerbation.

3.4. MET3-AE system

The MET3-AE system is a clinical decision support system for supporting management of pediatric asthma exacerbations [30]. Wilk et al. [31] described the design of MET3-AE system, and Sayyad Shirabad et al. [32] described its implementation. The system runs on desktop computers and mobile devices including tablets, and interacts with hospital information systems using HL7 messages. Diagnostic capabilities of MET3-AE are provided by a dedicated prediction model. In this paper we describe the research behind development and evaluation of this model (including a number of computational experiments).

MET3-AE was piloted in the ED at CHEO in order to assess its usability and to learn about the motivation of physicians to use computerized support at the point of care [33]. We used this pilot to prospectively collect patient data using MET3-AE. Due to restrictions imposed by the CHEO Research Ethics Board, the diagnostic support function of the system was unavailable to physicians, thus the system acted as a mobile electronic patient chart. The version of MET3-AE used in the pilot was expanded by adding:

- A screen to record and evaluate eligibility criteria (i.e., inclusion and exclusion criteria from Table I),
- A function to automatically compute and store the PRAM score for each recorded assessment,
- A function to record and store predictions of exacerbation severity (mild or moderate/severe) made by a physician during each assessment,
- A helper application to support patient record audit and follow-up in order to establish verified exacerbation severity.

4. Results

4.1. Results of phase 1

4.1.1. Collected data

Physicians approached 472 patients deemed potentially eligible at nursing triage. From this number, after subsequent assessment, they excluded 83 patients. From the remaining 389 eligible patients, 98 patients were excluded because of insufficient data recorded on paper forms (missing history or triage assessment). This produced a data set with 291 records. During preprocessing we established that the average time point for the first secondary assessment was 82 minutes after the primary assessment and the standard deviation was 48 minutes. Thus, for each record we selected the secondary assessment that was located within the time window of 34-130 minutes and was closest to its middle point. Records with no secondary assessment in the time window were excluded. This eliminated 51 records and the final set comprised 240 patient records, each described by 42 attributes corresponding to the patient's history, current asthma exacerbation, primary assessment and a selected secondary assessment. The average age of patients in the final set was 6.0 years (standard deviation of 4.0 years), and 131 (54.6%) of them suffered from moderate/severe exacerbation.

4.1.2. Selected prediction model

Results obtained by cross validation for all evaluated prediction models are given in Table III. In terms of AUC, the best performance was achieved by the NB model (AUC of 0.74). The difference between this model and the E-DT model (AUC of 0.70) was not statistically significant, while the differences between the NB model and the remaining models were statistically significant. The NB model also had the best predictive accuracy (68.0%). Differences between the NB model and the E-DT and IB10 models (prediction

accuracy of 64.1% and 63.6% respectively) were not statistically significant, and for the remaining models the differences were statistically significant.

According to these results the two best performing models are NB and E-DT (contrary to our expectations, the single DT didn't perform as well as in our earlier retrospective study [10]). We decided to select the NB model due to its simpler structure. This decision was further supported by Sajda [34] who discussed the advantages of methods based on Bayesian reasoning in the context of biomedical decision making, including the ability to deal with noisy and incomplete data.

4.2. Results of phase 2

4.2.1. Collected data

Physicians approached 129 patients deemed potentially eligible by automatic parsing of HL7 admission messages. From this number, after subsequent assessment they excluded 16 patients. From the remaining 113 eligible patients, 11 patients were excluded during audit and follow-up (inability to verify disposition decision or questionable data) and an additional 20 patients were eliminated because of a lack of some data required for calculating a PRAM score. This produced a final data set comprised of 82 patients and this set was used to conduct comparative evaluations of the NB model, PRAM score, and physicians' predictions. The average age of patients in the final set was 4.9 years (standard deviation of 3.5 years), and 56 (68.3%) of them suffered from moderate/severe exacerbation.

4.2.2. Predictive performance

The predictions made by the NB model, the PRAM score, and physicians are presented in Table IV (confusion matrix) and Table V (performance indicators). While all performed almost identically when predicting a mild asthma exacerbation – the only difference being 1 misclassified patient by the NB model, the difference was more pronounced with the moderate/severe category – the NB model and the PRAM score misclassified 5 and 4 more patients than the physicians, respectively.

While the physicians demonstrated better performance than the NB model and the PRAM score, according to the McNemar's tests we were not able to state that the differences between outcomes in pairs of the predictors were statistically significant.

5. Discussion

5.1. General remarks

The NB model predicted the severity of asthma exacerbations (mild vs. moderate/severe) comparably to the PRAM score, but not as well as physicians (accuracy of 70.7% vs. 73.2 vs. 78.0%, respectively). Specifically, the NB model and PRAM score underestimated severity (see Table V), while the physicians were better when identifying moderate/severe asthma patients. A plausible explanation for such a situation might be that predicting moderate/severe asthma needs more than reliance on basic symptoms, signs and tests and involves to a greater extent physicians' tacit knowledge that is not captured in the scores or prediction models. However, we were not able to show that these differences in predictive performance were statistically significant.

Despite differences in the study designs (setting different than ED, prediction of asthma severity instead of exacerbation severity) results produced by the NB model in phase 2 compare favorably with results reported in related studies. The NB model was more accurate than the system for asthma management described by Hoeksema et al. [18] (70.7% vs. 63.0%) and the ADEMA system [23] (70.7% vs. 61.0%). Such performance of the NB model is encouraging.

5.2. Limitations of the study

Any prospective study conducted in the ED is difficult to design as it interrupts the workflow by requiring physicians to perform additional tasks while carrying on with regular patient management. The setting and associated study design ramifications did not allow us to carry out direct observation of physicians doing data collection and making prediction decisions. We also left patient enrollment entirely to the discretion of the attending physicians, and we did not prompt them to enroll even when parsing HL7 messages identified eligible patients.

During phase 2, only one observer saw each patient. We have discussed this decision in Section 3.2.2. Moreover, the experiment did not follow the randomized controlled trial design and was confined to a single pediatric center (there are three pediatric academic centers in the Province of Ontario, and CHEO is one of them). This was because of the nature of the primary study involving MET3-AE and its usability assessment.

Finally, phase 2 of the study was conducted only in an academic center. A true value provided by a prediction model should be assessed in the community or small hospitals,

where physicians may not be as experienced and comfortable in assessing the pediatric population.

5.3. Implications and future research

Using data collected during phase 2, we are continuing to work on improving the predictive performance of the NB model. Such a model should perform better in predicting moderate/severe asthma exacerbations in order to become part of a system that is accepted by physicians. We expect that the model's performance will be improved if the original set of attributes is augmented with those associated with tacit knowledge physicians' use when assessing a patient. We are working on ways of capturing this knowledge for pediatric asthma exacerbations and subsequently codifying it for use in developing a prediction model.

Although our study was conducted in an academic hospital, we feel that the most important setting for this type of evaluation will be in the ED of community or small hospitals. If proven successful in such a setting, the NB model and similar prediction models should have true impact on improving the quality of care delivered to pediatric asthma patients.

6. Conclusions

The NB model was the best performing prediction model among those considered in phase 1. It demonstrated the best performance in terms of AUC and the observed differences were in most cases statistically significant. Therefore, it was selected for comparison with the PRAM score and physicians in phase 2.

In light of the relatively close performance of the NB model and the PRAM score, it is possible to argue about the gains in flexibility when using a prediction model like NB instead of a score. This is because the NB model can be applied to incomplete patient data while a medical score cannot (note that we had to exclude 20 patients from the study because we did not have sufficient data to calculate PRAM score). In summary, the NB model shows promise in predicting the severity of asthma exacerbations but needs further research to improve the accuracy of predictions and to further verify it on a larger data set. In light of the model's ease of use, it can be considered as a tool to use in conjunction with established medical scores, such as the PRAM score.

Clinical relevance statement

Evaluation of predictive abilities of a prediction model, medical score, and human specialist using prospectively collected data is rarely reported in the literature, but results of such evaluations are important for developing clinical decision support systems. While the study reported here showed that physicians working in an academic hospital were better at predicting pediatric asthma exacerbations than the PRAM score or a machine-learning prediction model (the NB model), the performance of the latter two was very close. Considering that the use of the NB model does not require knowing values for all clinical attributes, such a model can be considered as an adjunct to clinical scores routinely used in asthma management.

Protection of human and animal subjects

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and was approved by the CHEO Ethics Review Board.

Acknowledgements

The authors would like to thank the reviewers for their constructive comments.

This research was conducted when Szymon Wilk and Dympna O'Sullivan were post-doctoral fellows at the MET Research Group, University of Ottawa.

The authors acknowledge Tomasz Buchert, Bartosz Kukawka, and Tomasz Maciejewski for their work on customizing the MET3-AE system for phase 2 of the study, and the ED physicians and nurses at CHEO are acknowledged for participating in the study.

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (Collaborative Health Research Program), University of Ottawa Research Chairs Program, Children's Hospital of Eastern Ontario Research Institute, and the Polish National Science Centre.

Conflict of interest statement

No conflicts of interest exist.

References

[1] Dexheimer JW, Abramo TJ, Arnold DH, Johnson KB, Shyr Y, Ye F, Fan KH, Patel N, Aronsky D. An asthma management system in a pediatric emergency department. Int. J. Med. Inform. 2013;82(4):230-8.

[2] National Asthma Education and Prevention Program. Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma-Summary Report 2007. J. Allergy Clin. Immunol. 2007;120(5 Suppl):S94-138.

[3] Chu S, Tan J, Seabrook JA, Rieder MJ. Paediatric asthma severity score and length of stay in patients presenting to a paediatric emergency department. Paediatr. Perinat. Drug. Ther. 2008;8(4):150-3.

[4] Zemek R, Plint A, Osmond MH, Kovesi T, Correll R, Perri N, Barrowman N. Triage nurse initiation of corticosteroids in pediatric asthma is associated with improved emergency department efficiency. Pediatrics. 2012;129(4):671-80.

[5] Sanders DL, Aronsky D. Biomedical informatics applications for asthma care: a systematic review. J. Am. Med. Inform. Assoc. 2006;13(4):418-27.

[6] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, PA: ACM; 2006, p. 161-8.

[7] Lim T-S, Loh W-Y, Shih Y-S. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Mach. Learn. 2000;40(3):203-28.

[8] Perlich C, Provost F, Simonoff JS. Tree induction vs. logistic regression: A learning-curve analysis. J. Mach. Learn. Res. 2003;4:211-55.

[9] Xue J-H, Titterington DM. Comment on "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes". Neural Process. Lett. 2008;28(3):169-87.

[10] Farion K, Michalowski W, Wilk S, O'Sullivan D, Matwin S. A tree-based decision model to support prediction of the severity of asthma exacerbations in children. J. Med. Syst. 2010;34(4):551-62.

[11] Dexheimer JW, Brown LE, Leegon J, Aronsky D. Comparing decision support methodologies for identifying asthma exacerbations. Stud. Health Technol. Inform. 2007;129(Pt 2):880-4.

[12] Sanders DL, Aronsky D. Prospective evaluation of a Bayesian Network for detecting asthma exacerbations in a Pediatric Emergency Department. AMIA Annu. Symp. Proc. 2006:1085.

[13] Ducharme FM, Chalut D, Plotnick L, Savdie C, Kudirka D, Zhang X, Meng L, McGillivray D. The Pediatric Respiratory Assessment Measure: a valid clinical score for assessing acute asthma severity from toddlers to teenagers. J. Pediatr. 2008;152(4):476-80.

[14] Sanders DL, Aronsky D. Detecting asthma exacerbations in a pediatric emergency department using a Bayesian network. AMIA Annu. Symp. Proc. 2006:684-8.

[15] Sanders DL, Gregg W, Aronsky D. Identifying asthma exacerbations in a pediatric emergency department: a feasibility study. Int. J. Med. Inform. 2007;76(7):557-64.

[16] Powell CV, Kelly AM, Kerr D. Lack of agreement in classification of the severity of acute asthma between emergency physician assessment and classification using the National Asthma Council Australia guidelines (1998). Emerg. Med. 2003;15(1):49-53.

[17] Eccles M, McColl E, Steen N, Rousseau N, Grimshaw J, Parkin D, Purves I. Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. BMJ. 2002;325(7370):941.

[18] Hoeksema LJ, Bazzy-Asaad A, Lomotan EA, Edmonds DE, Ramirez-Garnica G, Shiffman RN, Horwitz LI. Accuracy of a computerized clinical decision-support system for asthma assessment and management. J. Am. Med. Inform. Assoc. 2011;18(3):243-50.

[19] Birken CS, Parkin PC, Macarthur C. Asthma severity scores for preschoolers displayed weaknesses in reliability, validity, and responsiveness. J. Clin. Epidemiol. 2004;57(11):1177-81.

[20] Gorelick MH, Stevens MW, Schultz TR, Scribano PV. Performance of a novel clinical score, the Pediatric Asthma Severity Score (PASS), in the evaluation of acute asthma. Acad. Emerg. Med. 2004;11(1):10-8.

[21] Gouin S, Robidas I, Gravel J, Guimont C, Chalut D, Amre D. Prospective evaluation of two clinical scores for acute asthma in children 18 months to 7 years of age. Acad. Emerg. Med. 2010;17(6):598-603.

[22] Zolnoori M, Zarandi MH, Moin M. Application of intelligent systems in asthma disease: designing a fuzzy rule-based system for evaluating level of asthma exacerbation. J. Med. Syst. 2012;36(4):2071-83.

[23] Sefion I, Ennaji A, Gailhardou M, Canu S. ADEMA: a decision support system for asthma health care. Stud. Health Technol. Inform. 2003;95:623-8.

[24] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int. J. Med. Inform. 2008;77(2):81-97.

[25] Weka 3: Data Mining Software in Java [cited 2013 Jun 23]. Available from: http://www.cs.waikato.ac.nz/ml/weka/.

[26] Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 3rd ed. Morgan Kaufmann; 2011.

[27] Dietterich TG. Approximate statistical test for comparing supervised classification learning algorithms. Neural Comput. 1998;10(7):1895-923.

[28] The R Project for Statistical Computing [cited 2013 Jun 23]. Available from: http://www.r-project.org/.

[29] Chalut DS, Ducharme FM, Davis GM. The Preschool Respiratory Assessment Measure (PRAM): a responsive index of acute asthma severity. J. Pediatr. 2000;137(6):762-8.

[30] Wilk S, Michalowski W, Farion K, Sayyad Shirabad J. MET3-AE system to support management of pediatric asthma exacerbation in the emergency department. Stud. Health Technol. Inform. 2010;160(Pt 2):841-5.

[31] Wilk S, Michalowski W, O'Sullivan D, Farion K, Sayyad-Shirabad J, Kuziemsky C, Kukawka B. A task-based support architecture for developing point-of-care clinical decision support systems for the emergency department. Methods Inf. Med. 2013;52(1):18-32.

[32] Sayyad Shirabad J, Wilk S, Michalowski W, Farion K. Implementing an integrative multi-agent clinical decision support system with open source software. J. Med. Syst. 2012;36(1):123-37.

[33] O'Sullivan D, Doyle JS, Michalowski W, Wilk S, Farion K, Kuziemsky C. Assessing the motivation of MDs to use computer-based support at the point-of-care in the emergency department. AMIA Annu. Symp. Proc. 2011:1045-54.

[34] Sajda P. Machine learning for detection and diagnosis of disease. Annu. Rev. Biomed. Eng. 2006;8:537-65.

Table I. Inclusion and exclusion criteria used in the study

Inclusion criteria (all of)
1. Patient age between 1 and 17 years. 2. Pre-existing diagnosis of asthma or reactive airways disease. Patients must have been previously prescribed inhaled bronchodilator therapy for 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.
Exclusion criteria (any one of)
1. Patients receiving oral steroids chronically for asthma or any other illness, or for more than 48 hours prior to their ED visit for an acute exacerbation. 2. Patients presenting for medication refills or other non-urgent reasons related to asthma, and not requiring ED treatment.

Table II. Clinical attributes characterizing a pediatric asthma exacerbation

#	Attribute	Values
Medical history		
1	Age at registration	numerical (years)
2	Primary care	family doctor, pediatrician, walk-in/other, none
3	Previous chest clinic assessment	yes, no
4	Age of first asthma symptoms	< 1 year, 1-3 years, > 3 years
5	Current inhaled steroids	< 1 week, 1-4 weeks, > weeks, none
6	Last exacerbation	< 1 month ago, 1-3 months ago, 3-12 months ago, > 12 months ago
7	Last oral steroids	< 1 month ago, 1-3 months ago, 3-12 months ago, > 12 months ago, never
8	Previous ED visits for asthma	1, 2, 3, 4 or more, none
9	Previous admission for asthma	floor, ICU, none
10	Smokers at home/daycare	yes, no
11	Carpets in bedroom	yes, no
12	Pet allergies	yes, no
13	Food allergies	yes, no
14	Environmental allergies	yes, no
15	Atopy/eczema	yes, no
16	Family history of asthma	yes, no
Current exacerbation		
17	Duration of symptoms	< 12 hours, 12-48 hours, >48 hours
18	URTI symptoms	yes, no
19	Fever	yes, no
20	Exposure to allergen	yes, no

#	Attribute	Values
21	Bronchodilators in the last 24 hours	1-3, 4-6, > 6, none
22	Transport to ED	self/parents, ambulance
Assessment (primary or secondary)		
23	SaO2 (oxygen saturation)	numerical (%)
24	Temperature	numerical (°C)
25	Heart rate	numerical (bpm)
26	Respiratory rate	numerical (bpm)
27	Skin color	pink/normal, dusky, pale
28	Distress	none, mild, moderate, severe
29	Suprasternal retractions	absent, present
30	Scalene retractions	absent, present
31	Air entry	normal, diminished at bases, diminished at widespread, minimal
32	Wheezing	absent, expiration, inspiration/expiration, audible with a stethoscope

Table III. Performance indicators for evaluated prediction models (95% CI); NB = naive Bayes model, DT = decision tree model, EDT = ensemble of decision trees models, SVM = support vector machine model, IB1, IB10 = instance-based model with 1 and 10 nearest neighbors respectively

	Model					
	NB	DT	EDT	SVM	IB1	IB10
AUC	0.74 (0.73, 0.76)	0.59 (0.57, 0.62)	0.70 (0.68, 0.72)	0.63 (0.61, 0.65)	0.56 (0.54, 0.58)	0.68 (0.66, 0.70)
Accuracy [%]	68.0 (66.4, 69.6)	60.9 (59.1, 62.7)	64.1 (62.4, 65.8)	59.2 (57.3, 61.1)	56.3 (54.2, 58.3)	63.6 (61.7, 65.4)

Table IV. Confusion matrix: predictions of the NB model, PRAM score, and physicians

	Predicted severity					
	NB model		PRAM score		Physicians	
Confirmed severity	Mild	Moderate/severe	Mild	Moderate/severe	Mild	Moderate/severe
Mild	19	7	20	6	20	6
Moderate/severe	17	39	16	40	12	44

Table V. Performance indicators for the NB model, PRAM score, and physicians' predictions (95% CI); PPV = positive predictive value, NPV = negative predictive value

	NB model		PRAM score		Physicians	
Accuracy [%]	70.7	(60.9, 80.6)	73.2	(63.6, 82.8)	78.0	(69.1, 87.0)
Sensitivity	0.70	(0.58, 0.82)	0.71	(0.60, 0.83)	0.79	(0.68, 0.89)
Specificity	0.73	(0.56, 0.90)	0.77	(0.61, 0.93)	0.77	(0.61, 0.93)
PPV	0.85	(0.74, 0.95)	0.87	(0.77, 0.97)	0.88	(0.79, 0.97)
NPV	0.53	(0.36, 0.69)	0.56	(0.39, 0.72)	0.63	(0.46, 0.79)