Patient-Specific Models for Predicting the Outcomes of Patients with Community Acquired Pneumonia

Shyam Visweswaran, M.D., M.S. and Gregory F. Cooper, M.D., Ph.D.
Center for Biomedical Informatics and the Intelligent Systems Program
University of Pittsburgh, Pittsburgh, Pennsylvania

ABSTRACT

We investigated two *patient-specific* and four *population-wide* machine learning methods for predicting dire outcomes in community acquired pneumonia (CAP) patients. Predicting dire outcomes in CAP patients can significantly influence the decision about whether to admit the patient to the hospital or to treat the patient at home. Population-wide methods induce models that are trained to perform well on average on all future cases. In contrast, patient-specific methods specifically induce a model for a particular patient case. We trained the models on a set of 1601 patient cases and evaluated them on a separate set of 686 cases. One patient-specific method performed better than the population-wide methods when evaluated within a clinically relevant range of the ROC curve. Our study provides support for patient-specific methods being a promising approach for making clinical predictions.

INTRODUCTION

The practice of medicine typically entails predicting outcomes under uncertainty, such as predicting dire outcomes that include mortality and serious complications in patients diagnosed with community acquired pneumonia (CAP). Making better outcome predictions has the potential for improving decisions taken by healthcare providers, leading to better patient care and more efficient utilization of healthcare resources. The management of CAP is an important healthcare problem since it causes significant patient mortality, morbidity, and healthcare resource utilization. In 1994 in the U.S., there were 4.1 million patients diagnosed with CAP of which 1.2 million were hospitalized. In the same year, the total cost of treating CAP was estimated to be about 10 billion dollars.

Accurately predicting the probability of a dire outcome in a patient presenting with CAP is an important factor in deciding whether to admit the patient to the hospital or not. Patients who are very unlikely to have a dire outcome might be safely treated at home with oral antibiotics, while higher risk patients would preferably be treated in the hospital with intravenous antibiotics.

Numerous probabilistic models, based on statistical and machine learning techniques, for making clinical predictions have been described in the literature. Commonly used methods include logistic regression, neural networks, *k*-Nearest neighbor techniques, decision trees, and Bayesian networks [1]. Almost always, a single model is induced from a training set of patient cases, with the intent of applying it to future patient cases. We call such a model a *population-wide* model because it is intended to be applied to an entire population of future cases. A population-wide model is optimized such that it predicts well on average when applied to future patients. In contrast, a *patient-specific* model is specifically constructed for a particular patient. Such a model is optimized to predict especially well for the single patient case for which it is intended. This optimization is achieved by specializing the model structure and parameters, as well as the model search, based on the known features of the patient case at hand.

The discriminative performance of a predictive model is typically evaluated with a Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC). However, in clinical domains, the entire range of the ROC curve may not useful for decision making. Rather, only a narrow region of the ROC curve is useful. Predictive models that perform better in this restricted clinically relevant region are likely to be more helpful to a decision-maker.

In this paper, we investigate the performance of six machine learning methods, including two patient-specific methods, for the clinical problem of predicting dire outcomes in patients diagnosed with CAP. We show that patient-specific methods perform better than population-wide methods when evaluated in a clinically relevant region of the ROC curve.

THE PORT DATASET

The pneumonia database that we used contains several hundred clinical variables on 2287 patients diagnosed with CAP. The data was collected by the Pneumonia Patient Outcomes Research Team (PORT) using a prospective cohort study of hospitalized and ambulatory care patients. The study was conducted from October 1991 to March 1994 at five hospitals in three geographical locations: Pittsburgh, Boston, and Halifax, Nova Scotia. Eligibility criteria were that a patient must (1) be at least 18 years of age, (2) have one or more symptoms suggestive of pneumonia, and (3) have radiographic evidence of pneumonia within 24 hours of presentation [2].

During the study enrollment period, 4002 individuals satisfied the entry criteria for study eligibility, of whom 2287 (57.1%) were enrolled. Based on chart review, hundreds of data items were collected for each of the 2287 patients. Enrolled patients were followed prospectively to assess their vital status and a variety of outcomes at 30 days after the radiographic diagnosis of pneumonia.

One key goal of the PORT project was to develop accurate criteria for prognosis of patients with CAP that could provide guidance on which patients should be hospitalized and which patients might be safely treated at home.

METHODS

Selection of Variables

From the available variables in the PORT dataset, we selected 158 clinical variables that are typically available in the Emergency Department at the time the decision whether to admit or not is made. The variables include demographic information, history and physical examination information, laboratory results, and chest X-ray findings. Of the 158 variables, 120 are discrete and the remaining 38 are continuous. A majority of the discrete variables are binary and the rest have between three to seven values. The 38 continuous variables are derived mainly from laboratory tests and were discretized based on thresholds provided by clinical experts on the PORT project. These 158 discrete variables constituted the input for the machine learning methods.

The binary outcome variable that we selected is called *dire outcome*. A patient was considered to have experienced a dire outcome if any of the following occurred: (1) death within 30 days of presentation, (2) an initial intensive care unit admission for respiratory failure, respiratory or cardiac arrest, or shock, or (3) the presence of one or more specific, severe complications. Based on these criteria, 261 patients (11.4%) had a dire outcome.

Mortality is commonly used as the outcome variable for developing statistical and machine learning predictive models. For example, the pneumonia severity index (PSI) is a model based on logistic regression that predicts patient mortality within 30 days of presentation with CAP. We chose to predict dire outcomes that include mortality as well as severe complications, because it seems likely that the decision about where to treat CAP patients (hospital versus home) is influenced not just by mortality, but also by other possible severe outcomes. A recent paper showed that even small improvements in predicting dire outcomes in CAP patients is projected to lead to significant savings in healthcare costs and improved delivery of healthcare [3].

Training and Test Datasets

The dataset consisting of 2287 patient cases was divided into a training set of 1601 cases (70%) and a test set of 686 cases (30%). This is the same split as described in [3]. The training set was created by randomly sampling from the 2287 cases in the dataset such that both the training and the test datasets had approximately the same proportion of cases with dire outcomes. The training and the test sets contained 182 (11.4%) and 79 (11.5%) cases of dire outcomes respectively. Missing data were filled-in using an iterated *k*-nearest neighbor method. This is a non-parametric EM-style algorithm using Gibbs sampling that is described in detail in [3].

Machine Learning Methods

We evaluated six machine learning classification methods that produce probabilistic outputs: Simple Bayes (SB), logistic regression (LR), artificial neural networks (ANN), *k*-Nearest Neighbor (*k*NN), the Lazy Bayesian Rule (LBR) learner, and the Patient-Specific model-Averaging (PSA) algorithm. The first four methods are among the commonest methods described in the medical literature for constructing predictive models and the last two are patient-specific methods. The LBR algorithm was introduced in the machine learning literature by Zheng and Webb [4]. We developed and applied a modified version of LBR. Both the original LBR and our modification of it are described below. The PSA algorithm is a patient-specific Bayesian model averaging method that we have developed and is described in detail in [5], where it is called the Instance-Specific model-Averaging algorithm. For SB, LR, NN, and *k*NN, we used the implementations in Weka (v3.3.6) for our experiments [6]. We implemented the modified LBR in Matlab (version 7) and the PSA in Java. Models were induced from the training set and were evaluated on the test set. Below, we describe SB, LR, NN, and *k*NN very briefly (see [7] for details) and LBR and PSA in some detail.

1. **Simple Bayes**. Simple Bayes (also known as Naïve Bayes) is a common machine learning method that often has excellent discriminative performance. The Simple Bayes classifier makes the simplifying assumption that features are conditionally independ-

ent given the outcome. For each feature it estimates the conditional probabilities given the outcome from the training set. Given a test case, the classifier chooses the outcome with the maximum posterior probability.

2. **Logistic Regression**. Logistic regression (LR) is commonly used for predictive modeling in the medical literature. Logistic regression derives the log odds of a binary outcome variable in terms of a linear combination of the feature variables. The coefficients of the feature variables are estimated from the training set, typically by using an iterative maximum likelihood method.

3. **Artificial Neural Networks**. Artificial Neural Networks (ANN) generalize logistic regression and are commonly used for learning non-linear relationships. The standard technique for learning ANN uses backpropagation that iteratively revises the parameters of the model based on the errors made by the model on a subset of the training data.

4. ***k*-Nearest Neighbor**. Given a test case to be predicted, the *k*-Nearest Neighbor method selects the *k* most similar training cases according to some similarity measure and averages their outcomes.

5. **Lazy Bayesian Rule**. The Lazy Bayesian Rule (LBR) learner is a classification algorithm that induces a rule from training cases in the neighborhood of the test case that is then used to classify it. A rule generated by LBR consists of (1) a conjunction of a subset of all the feature-value pairs present in the test case as the antecedent, and (2) a local simple Bayes classifier as the consequent. The structure of the local simple Bayes classifier consists of the outcome variable as the parent of all those features that do not appear in the antecedent, and the parameters of the classifier are estimated from the subset of training cases that satisfy the antecedent. Figure 1 shows an example of a LBR rule constructed using six variables from our dataset including the dire outcome variable. The rule has two features in the antecedent and a simple Bayes classifier with three features in the consequent. A greedy step-forward search selects the optimal LBR rule for a test case to be classified. In particular, features in the consequent of the current rule are temporarily moved one at a time from the consequent to the antecedent and evaluated for whether it reduces the overall error rate on the training set. The feature that most reduces the overall error rate is permanently added to the antecedent and removed from the consequent, and the search continues; if no feature decreases the current error rate, then the search halts and the current rule is applied to predict the outcome for the test case. When previously evaluated on 29 datasets from the UCI Machine Learning Repository, LBR had the lowest average error rate when compared to several machine learning methods [4].

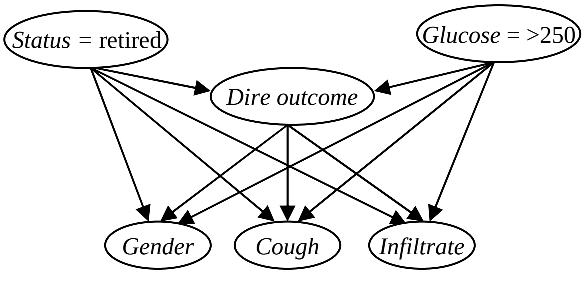

Figure 1. An example of a LBR model (or rule). The two nodes at the top represent features in the antecedent of the LBR rule that have been instantiated to their respective values in the test case. The node in the center (the outcome variable being predicted) and the three nodes at the bottom constitute the local simple Bayes classifier present in the consequent of the LBR rule.

LBR is an example of a patient-specific method that uses the features in the test patient case to direct the search for a suitable model in the model space. The original LBR method evaluates a candidate rule based on the error rate of the local Simple Bayes classifier applied to the relevant subset of the training dataset, using leave-one-out cross-validation. We modified the original LBR (henceforth called modified LBR) by replacing the error rate with the Brier score. For a binary variable like dire outcome that represents an event that either occurs or does not occur, the Brier score *B* for a set of *n* test cases is computed as:

$$B = \frac{1}{n}\sum_{i=1}^{n}(p_i - a_i)^2 \ ,$$

where p_i is the predicted probability of a dire outcome occurring in the i^{th} test case and a_i is the actual outcome for that case. The variable a_i is set to 1 if a dire outcome occurs and is set to 0 if no dire outcome occurs in the i^{th} test case. The Brier score ranges from 0 when all predictions are perfect to 1 when the outcome of every case is predicted incorrectly with perfect confidence. The score is sensitive to both calibration and discrimination, is easily computed, and the Brier scores of two models on the same cases can be statistically compared with the Williams-Kloot statistic [8]. We used greedy step-forward search to select the optimal LBR rule; a current rule was replaced by a candidate rule if the candidate rule had a significantly lower Brier score using a one-tailed test at the

0.001 significance level. For simplicity of implementation, we did not control for multiple testing. Thus, modified LBR may potentially select sub-optimal models and controlling for multiple testing might further improve its performance.

6. **Patient-Specific Algorithm**. The methods described so far select a single model from some model space, ignoring the uncertainty in model selection. Bayesian model averaging is a coherent approach to dealing with the uncertainty in model selection. However, since the number of models is typically enormous, exact model averaging over the entire model space is usually not feasible. The PSA algorithm approximates Bayesian model averaging in a patient-sensitive manner. The current implementation of PSA searches over LBR models, using the features of the test case to direct the search. The prediction for the outcome of the test case is obtained by combining the predictions of the selected models weighted by their posterior probabilities. When evaluated on 29 UCI datasets, PSA had significantly fewer errors on seven datasets and significantly more errors on two datasets when compared to the original LBR algorithm, providing support that patient-specific model averaging can improve on patient-specific model selection [5].

RESULTS AND DISCUSSION

Area under the ROC curve. The AUCs (with 95% confidence intervals) for the six methods are shown in the second column in Table 1. Modified LBR achieves the highest AUC at 0.861 with 95% confidence interval [0.826, 0.896]. This is statistically significantly higher (at the 0.05 level) than LR and *k*-NN but not significantly different from the AUCs of ANN, SB and PSA.

Analysis of the clinically relevant region of the ROC curve. Based on the AUC the patient-specific methods (i.e., modified LBR and PSA) did not perform significantly better than population-wide methods like SB and ANN. However, the ROC curves of different methods may have significantly different performance characteristics in a particular region, even if they have similar AUCs. In clinical decision-making, the entire range of the ROC curve is usually not of interest. In the case of CAP, one of the PORT clinical investigators, who is a CAP expert, assessed that an acceptable *error rate* corresponds to no more than 1 to 2 percent of CAP patients treated at home experiencing a dire outcome. The error rate can be used to determine a point on the ROC curve that is clinically relevant in influencing decisions about where to treat CAP patients.

We identified a point on the ROC curve for each method that is between 1 to 2 percent error and as close as possible to 1%. We describe in detail the performance characteristics of such a point for the modified LBR method. The identified point on the ROC curve closest to 1% is shown in Figure 2. The precise error rate at this point is 1.2% corresponding to 5 patients who were treated at home and had a dire outcome. Operating at this point on the ROC curve, modified LBR recommends treating 400 patients at home out of the 686 in the test set. Based on the actual care given, 280 patients were treated at home of which 5 patients (1.8%) had a dire outcome. Thus, modified LBR recommends treating 17.5% more patients at home than were actually treated at home, which is highly statistically significant (P < 0.0001). Since, modified LBR's error rate is not statistically significantly different from that of the actual care (1.3% versus 1.8%); its recommendations are not likely to reduce the quality of healthcare. We next estimate the potential cost savings from such a reduction in admissions.

The estimated cost for treating a CAP patient in the hospital in the U.S. in 1994 was $7517 compared to $421 for treating the patient at home [9]. If the

Method	AUC with 95% confidence interval	Errors and percent error with 95% confidence interval	Number and percent (with 95% confidence interval) of patients treated at home
Logistic regression (LR)	0.741 [0.681, 0.802]	3 (1.8% [0.36%, 5.07%])	170 (24.8% [21.59%, 28.19%])
Neural Network (ANN)	0.828 [0.783, 0.873]	3 (1.7% [0.36%, 5.01%])	172 (25.1% [21.87%, 28.49%])
k-Nearest Neighbor (*k*-NN)	0.787 [0.738, 0.837]	3 (1.5% [0.31%, 4.30%])	201 (29.3% [25.92%, 32.86%])
Simple Bayes (SB)	0.850 [0.817, 0.883]	5 (1.3% [0.43%, 3.33%])	368 (53.6% [49.83%, 57.42%])
Modified LBR	0.861 [0.826, 0.896]	5 (1.2% [0.41%, 2.89%])	400 (58.3% [54.52%, 62.03%])
PSA	0.853 [0.818, 0.876]	5 (1.3% [0.42%, 2.97%])	390 (56.8% [53.05%, 60.59%])
Actual care	-	5 (1.8% [0.58%, 4.12%])	280 (40.8% [37.11%, 44.60%])

Table 1. Performance of the various methods on the test set of 686 cases. The method for computing the errors in column 3 is described in detail in the text. The last column gives the number of patients that would be treated at home out of 686 when operating at the point on the ROC curve that corresponds to the error rate in the third column. The last row gives the actual error rate and the actual number of patients treated at home.

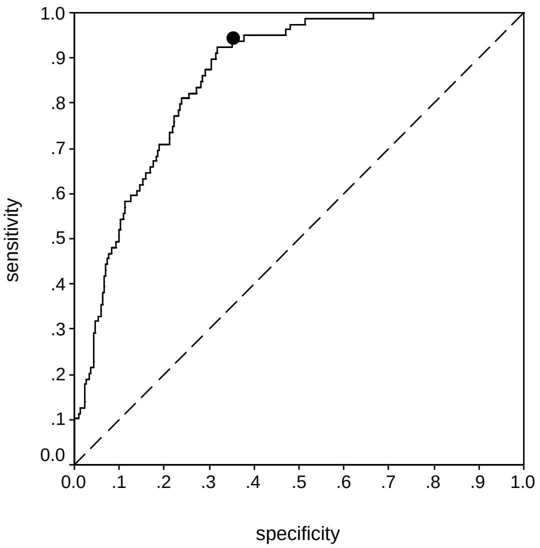

Figure 2. ROC curve on the test set for the modified LBR method. The dot indicates the operating point on the curve that is discussed in the text.

recommendations of the modified LBR model were followed for all the CAP patients in the U.S. in 1994, the 1.2 million CAP hospital admissions in that year would be reduced by 17.5%. This would lead to savings of more than 1.5 billion dollars.

The number and percent of patients recommended for treatment at home for all the methods are given in the last column of Table 1. Modified LBR that was discussed in detail in the previous paragraph, performs significantly better (at the 0.05 significance level) than the other methods except for PSA.

CONCLUSION

We evaluated the performance of six machine learning methods on the clinical problem of predicting dire outcomes in CAP patients. The two patient-specific methods, modified LBR and PSA, and two of the population wide methods, SB and ANN, had similar AUCs. We also analyzed the performance of the methods in a clinically relevant region of the ROC curve that corresponded to an error rate between 1 to 2 percent. Based on the proportion of patients treated at home, modified LBR performed significantly better than all other methods except for PSA.

Although four methods had similar AUCs, they did not all perform similarly in a focused region of the ROC curve. This suggests that for clinical problems like predicting dire outcomes in CAP patients, the AUC can be too broad a measure that is unable to differentiate among various predictive models. A more focused assessment of the clinically relevant region of the ROC curve can be more informative. We briefly discussed the potential cost savings that can accrue from operating at a clinically relevant region of the ROC curve. A more detailed analysis is described in [3].

There are limitations to our study. The patient-specific methods, especially modified LBR, performed well in a small region of the ROC curve that was deemed significant by an expert. However, different physicians may have dissimilar preferences and we did not analyze other possibly relevant regions of the ROC curve. Patient preferences are also important in making decisions of hospital admissions; we did not assess them in this study.

In the future, we plan to develop and apply methods that can directly optimize a specified region of the ROC curve. Such methods have the potential for customizing predictive models to the preferences of the healthcare decision maker.

Acknowledgements

This research was supported by several grants from the National Library of Medicine (R01-HS06468, R01-LM008374 and T15-LM/DE07059-17 to the University of Pittsburgh's Biomedical Informatics Training Program).

REFERENCES

1. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: A methodology review. J Biomed Inform 2002;35(5-6):352-9.
2. Fine MJ, Stone RA, Singer DE, Coley CM, Marrie TJ, Lave JR, et al. Processes and outcomes of care for patients with community-acquired pneumonia: Results from the Pneumonia Patient Outcomes Research Team (PORT) cohort study. Arch Intern Med 1999;159(9):970-80.
3. Cooper GF, Abraham V, Aliferis CF, Aronis J, Buchanan BG, Caruana R, et al. Predicting dire outcomes of patients with community acquired pneumonia. Journal of Biomedical Informatics 2005.
4. Zheng ZJ, Webb GI. Lazy learning of Bayesian rules. Machine Learning 2000;41(1):53-84.
5. Visweswaran S, Cooper GF. Instance-specific Bayesian model averaging for classification. In: Advances in Neural Information Processing Systems (NIPS); 2004 December 13-16; Vancouver, Canada; 2004.
6. Witten IH, Frank E. Data Mining: Practical machine learning tools with Java implementations. San Francisco: Morgan Kaufmann; 2000.
7. Mitchell TM. Machine Learning. 1st ed. New York: McGraw Hill; 1997.
8. Redelmeier DA, Bloch DA, Hickam DH. Assessing predictive accuracy: How to compare Brier scores. J Clin Epidemiol 1991;44(11):1141-6.
9. Lave JR, Lin CC, Fine MJ. The cost of treating patients with community-acquired pneumonia. Semin Respir Crit Care Med 1999;20:189-198.