Using markers of inflammation to predict mortality in hospitalized patients with community-acquired pneumonia with ensemble machine learning methods

Robert Kelley, Timothy Wiemken, William Mattingly, Brian Guinn, Laura Binford, Kim Buckner, Brick Green, Josh Britton, Paula Peyrani, Julio Ramirez

ABSTRACT

Introduction: Community-acquired pneumonia (CAP) is one of the leading causes of morbidity and mortality in the US. Investigators and practitioners continually seek methods to determine severity of disease for hospitalized patients with CAP to guide their management plans. We propose using ensemble machine-learning models to classify whether a patient with CAP will live or die based on markers of inflammation.

Materials and Methods: This study was a secondary analysis of the Community-Acquired Pneumonia Organization's (CAPO) database. Using markers of inflammation such as glucose, hemoglobin, and blood pressure, we conducted an analysis of eight machine-learning models. These models included bagging, four types of boosting, support vector machine, naïve Bayes and simple decision tree to determine their suitability for classifying whether a patient will die or survive based on data collected at hospital admission. For each model we calculated several performance measures. We further developed a prototype of a computer application to implement the best performing model.

Results: Decision Trees with Bagging produced the highest area under the curve (0.95) and performed over 0.90 for several metrics including sensitivity/recall, precision, and F1 score. We implemented our computer application prototype using this model.

Conclusions: We conclude that markers of inflammation collected at hospital admission can be used to classify patients on their likelihood of death from community-acquired pneumonia. The decision tree algorithm with bagging shows promise for implementation into a software application for investigators and practitioners to use at the bedside or embedded into an electronic medical record system.

Introduction

In the United States, community-acquired pneumonia (CAP) is one of the leading causes of morbidity and mortality [1]. A significant number of patients with severe pneumonia will progress to clinical failure and death even with appropriate therapy. It has been suggested that lung and systemic inflammatory responses affect clinical

outcomes even after the pathogen has been eradicated. Studies have documented an exaggerated cytokine response in the systemic circulation in patients with severe pneumonia [2]. Since this "cytokine storm" has been associated with poor patient outcomes, investigators have evaluated the use of immunomodulatory agents such as corticosteroids in patients with severe pneumonia. However, these attempts to control the cytokine storm and improve mortality have offered conflicting results [3, 4]. A possible explanation for these conflicting results is that patients are enrolled into steroid trials based on having severe pneumonia, as defined by pneumonia severity scores. In fact, a general consensus is emerging indicating that our current approach to stratify patients for clinical studies using pneumonia severity scores is not powerful enough to identify patients that may benefit from immunomodulatory therapies [5]. To identify these patients, we need new models that consider the degree of inflammation to predict clinical outcomes in hospitalized patients with pneumonia.

Machine learning models (MLMs) are a family of modeling techniques that extend traditional statistical methods for analyzing data. In general, MLMs are divided into two categories: supervised and unsupervised [6]. For supervised MLMs, the algorithms are "trained" using existing data with the outcome (e.g. pneumonia patients with a record of in-hospital mortality). Once trained, these algorithms are used to predict outcomes for new unseen data. For unsupervised MLMs, algorithms are presented with data without the outcome of interest; the algorithm then "clusters" instances that indicate some type of grouping. Commonly used MLMs

include decision trees, support vector machines, naïve bayes, k nearest neighbor, neural networks, hierarchical clustering, and principal-component analysis. In recent years, the technique of ensemble learning has become more prevalent. Ensemble learning employs multiple learning algorithms and aggregates the results to produce more robust predictions. In particular, bootstrap aggregation (bagging) and boosting are often used for ensemble learning. With bagging multiple learners are employed on different subsets of the data voting is used to generate a predictive model. Boosting is similar to bagging except in the voting step, some learners are weighted differently to more heavily penalize misclassified instances.

The goal of this study was to investigate potential MLMs to predict in-hospital mortality for hospitalized patients with pneumonia using markers of systemic inflammation. In this study, lab values (e.g. white blood cell count, hemoglobin, etc.) recorded on the first day of hospital admission were used surrogates to indicate systemic inflammation because cytokine levels are not typically collected as standard of care. A further objective was to compare the developed inflammation models with models that use the PSI and CURB 65 severity scores and their variables for predicting mortality.

Methods

This study was a secondary analysis of the Community-Acquired Pneumonia Organization's database. To be included in this study, subject records were required to have data in at least 50% of the variables that contribute to the Pneumonia Severity Index [7], the CURB65 [8], and lab values that may indicate systemic inflammation. The variables that were selected for this study are shown in Table 1.

3,420 cases were included in the final dataset before oversampling and 3,918 were included after oversampling.

Table 1: Candidate variables for Machine Learning Models for Predicting In-Hospital Mortality for Patients with Community-Acquired Pneumonia

PSI	**CURB 65**
Age Gender Nursing Home Resident (Y/N) Neoplastic Disease (Y/N) Liver Disease (Y/N) Congestive Heart Failure (Y/N) Cerebrovascular Accident (Y/N) Renal Disease (Y/N) Altered Mental Status (Y/N) Heart Rate >= 125 Beats per minute (Y/N) Respiratory Rate>30 breaths per minute (Y/N) Systolic Blood Pressure<90mm Hg (Y/N) Temperature <35 C or >40C (Y/N) Arterial PH <7.35 (Y/N) Blood Urea Nitrogen >=30 mg/dL (Y/N) Sodium < 130 mmol/L (Y/N) Glucose >=250 mg/dL (Y/N) Hematocrit <30% (Y/N) Pao2 < 60 mm Hg (Y/N) Pleural Effusion (Y/N)	Confusion (Y/N) Blood Urea Nitrogen > 19 mg/dL (Y/N) Respiratory Rate >= 30 (Y/N) Systolic Blood Pressure < 90 mmHg or Diastolic Blood Pressure <=60 mmHg (Y/N) Age >=65
Inflammation Variables	**Other Variables**
Altered Mental Status (Y/N) Heart Rate Respiratory Rate Systolic Blood Pressure Diastolic Blood Pressure Temperature Blood Urea Nitrogen Sodium Glucose Hemoglobin Hematocrit ABG Arterial PH ABG Paco2 ABG Pao2	Age Gender Time to Clinical Stability Length of Stay

Missing data were imputed with additive regression, bootstrapping and predicting mean matching using the *aregImpute* function available in the Hmisc package for R. Only 10% of the cases in the dataset had the outcome of mortality, leading to an imbalanced dataset. Imbalanced datasets often cause difficulty for classification

algorithms [9, 10], therefore the minority class (mortality) was oversampled using the synthetic minority oversampling technique (SMOTE) [11] . SMOTE uses the k Nearest Neighbor method to generate "synthetic" cases with minority outcomes so classification algorithms have more data from which to build predictive models.

The data was further divided into two sets: train, which contained 3135 cases (80%) and test, which contained 783 cases (20%). The models were trained using the train set and tested with the test set.

Seven experiments were conducted; each experiment was performed with a different set of predictors as shown in Table 2.

Table 2: Experiments and variables included in each.

Experiment Number	Variables
1: Inflammation Variables	Age, Sex, Altered Mental Status, Heart Rate, Respiratory Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Temperature, Blood Urea Nitrogen, Sodium, Glucose, Hemoglobin, Hematocrit, ABG Arterial PH, ABG Paco2,ABG Pao2
2: PSI Risk Score Only	PSI Risk (1-5)
3: PSI Variables Only	Age, Gender, Nursing Home Resident, Neoplastic Disease, Liver Disease, Congestive Heart Failure, Cerebrovascular Accident, Renal Disease, Altered Mental Status, Heart Rate, Respiratory Rate, Systolic Blood Pressure, Temperature, ABG Arterial PH, Blood Urea Nitrogen, Sodium, Glucose, Hematocrit, ABG Pao2, Pleural Effusion
4: CURB 65 Score Only	CURB 65 (1-5)
5: CURB 65 Variables Only	Confusion (Altered Mental Status), Blood Urea Nitrogen, Respiratory Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Age
6: Severe vs. Non-Severe CAP	Age, Sex, PSI Category (Severe/Non-Severe) Severe = PSI Score 4 or 5, Non-Severe=PSI Score 1, 2, or 3.
7. Early/Mid/Late Time to Clinical Stability	Age, Sex, Time to Clinical Stability Category (Early = <3 days, Mid=4-7 days, Late=>7 days

For each experiment, the same set of classification algorithms were executed with its respective set of variables. These algorithms were divided into two classes: ensembles which use multiple learners and voting schemes to predict the mortality outcome, and single, in which a single learner was used. The list of algorithms is shown in Table 3. In general, most of the ensemble methods used a variation of the boosting technique in which the weights assigned to the various learners are changed based on misclassification rates calculated in the training dataset. The method for changing these weights is different for each boosting scheme. The details for each scheme can be found at:

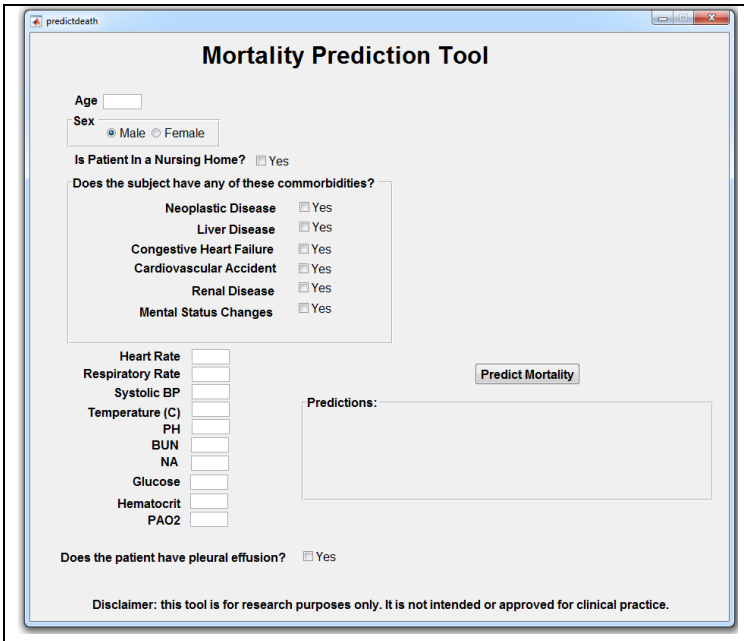
http://www.mathworks.com/help/stats/ensemble-methods.html#bsw8akh.

Table 3: Classification algorithms used in each experiment and its algorithm type

Classification Algorithm	Algorithm Type
Bagging	Ensemble
AdaBoost	Ensemble
LogitBoost	Ensemble
GentleBoost	Ensemble
RUSBoost	Ensemble
Support Vector Machine	Single
Naïve Bayes	Single
Decision Tree	Single

To evaluate the performance of each experiment/algorithm combination, receiver-operating characteristics (ROC) curves and confusion matrices were generated using the results from predictions on the test data set. In addition, a final ROC curve was generated to compare the performance of each experiment with only the bagging ensemble algorithm.

The Spyder Python Development Environment 2.3.5.2, R 3.2.0, and Matlab 2015a were used for all analysis.



Results

Each experiment generated an ROC curve to compare the performance of each algorithm. Figure 1 shows curves for two of the experiments which represent the two main result patterns seen across the experiments. Experiments with a larger number of variables demonstrated better performance than experiments with a limited number of variables.

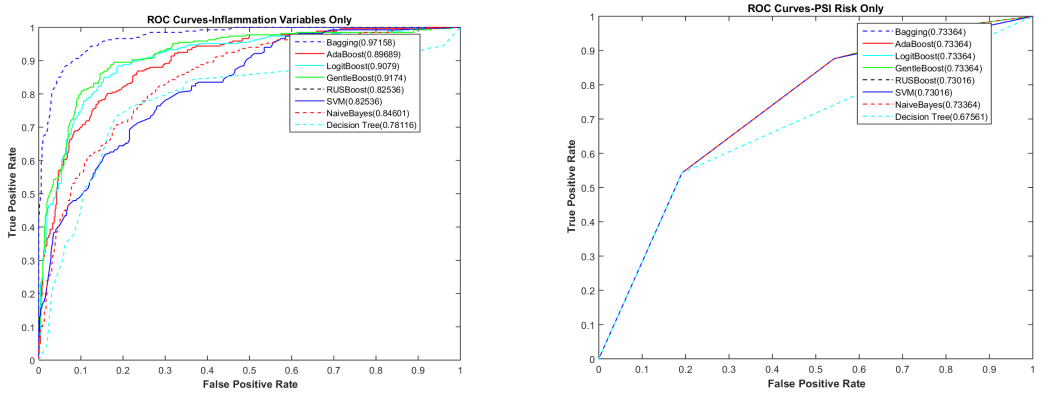


Figure 1: ROC Curves for Experiment 1 and 3

The highest AUC recorded across all experiments was .97 for Experiment 1 which used only inflammation variables and the bagging algorithm. The lowest AUC was .62 for Experiment 4 which used only CURB65 score and the bagging algorithm. The full set of results for each experiment is shown in Table 4

Table 4: Comparison of highest/lowest AUC and algorithms for each experiment

Experiment	Highest AUC (Algorithm)	Lowest AUC (Algorithm)
Experiment 1	**.97 (Bagging)**	.78 (Single Decision Tree)
Experiment 2	.73 (Bagging, AdaBoost, LogitBoost, GentleBoost)	.67 (Single Decision Tree)
Experiment 3	.96 (Bagging)	.80 (Single Decision Tree)
Experiment 4	.66 (all algorithms except Single Decision Tree)	.62 (Single Decision Tree)
Experiment 5	.89 (Bagging)	.81 (RUSBoost, SVM)
Experiment 6	.70 (Bagging)	.65 (Naïve Bayes
Experiment 7	.77 (Bagging)	.73 (RUSBoost, SVM)

For all experiments, the bagging algorithm either demonstrated the highest AUC or tied with other algorithms for highest AUC. Figure 2 shows the ROC curves/AUC values for only the bagging algorithm for each experiment.

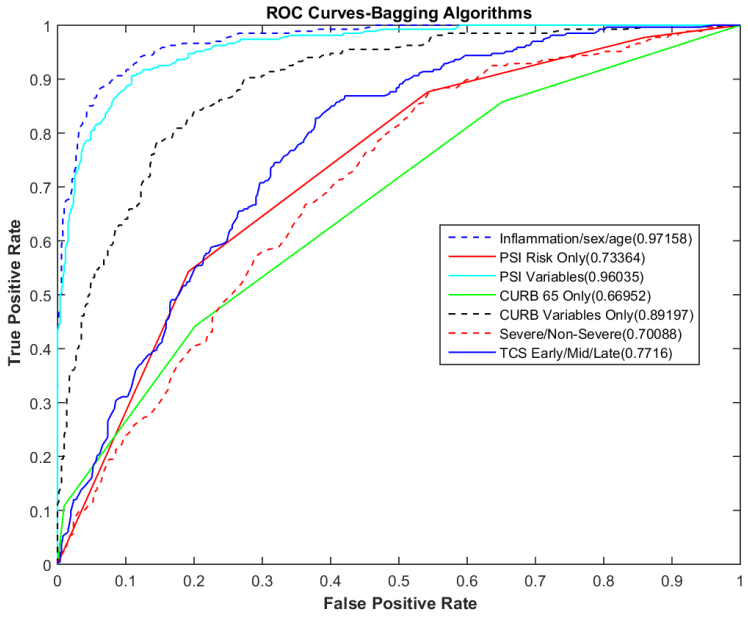


Figure 2: ROC curves for Bagging Algorithm for all experiments

Conclusions

It is clear that MLMs can provide adequate models for predicting mortality based on the markers of systemic inflammation recorded at hospital admission. In particular, bootstrap aggregation (bagging), which was consistently the highest performing algorithm, was extremely effective. Experiment 1 in which only markers of inflammation were used, had the highest AUC. However Experiment 3, which used all variables in the PSI was .96. For all practical purposes the performance was the same in this regard.

It is further clear that single variable scores like the PSI and CURB-65 are not satisfactory predictors for mortality using MLMs. This is likely due to the fact that too much data is compressed into a single data point and does not adequately

consider the issue of variability in continuous values. Using the raw data in the PSI without the cutoffs used to assign points for the score, yield much better results.

This study had some limitations. Most notably because this was not a prospective study, missing data in the many of the variables could not controlled. Consequently, imputation was employed to fill-in missing data. This introduces bias to the results. In similar fashion, oversampling the minority class also introduces bias.

Despite the limitations, it is clear that using laboratory values as surrogates for systemic inflammation can be a useful technique to predict mortality, which in turn, can be used to assign a patient to a particular arm in a randomized control trial. Further work is required to control variable collection and employ the same techniques using only cytokine data recorded from blood samples drawn on hospital admission.

References

[1] "National Center for Health Statistics. Health, United States, 2014," U. D. o. H. a. H. S. CfDCaP, Ed., ed. Hyattsville, MD, 2015.

[2] J. A. Kellum, L. Kong, M. P. Fink, L. A. Weissfeld, D. M. Yealy, M. R. Pinsky*, et al.*, "Understanding the inflammatory cytokine response in pneumonia and sepsis: results of the Genetic and Inflammatory Markers of Sepsis (GenIMS) Study," *Arch Intern Med,* vol. 167, pp. 1655-63, Aug 13-27 2007.

[3] V. F. Corrales-Medina and D. M. Musher, "Immunomodulatory agents in the treatment of community-acquired pneumonia: a systematic review," *J Infect,* vol. 63, pp. 187-99, Sep 2011.

[4] W. Nie, Y. Zhang, J. Cheng, and Q. Xiu, "Corticosteroids in the treatment of community-acquired pneumonia in adults: a meta-analysis," *PLoS One,* vol. 7, p. e47926, 2012.

[5] R. G. Wunderink, "Corticosteroids for severe community-acquired pneumonia: not for everyone," *JAMA,* vol. 313, pp. 673-4, Feb 17 2015.
[6] S. Marsland, *Machine Learning: an Algorithmic Approach*. London: CRC Press, 2015.
[7] M. J. Fine, T. E. Auble, D. M. Yealy, B. H. Hanusa, L. A. Weissfeld, D. E. Singer*, et al.*, "A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia," *New England Journal of Medicine,* vol. 336, pp. 243-250, 1997.
[8] W. S. Lim, M. M. van der Eerden, R. Laing, W. G. Boersma, N. Karalus, G. I. Town*, et al.*, "Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study," *Thorax,* vol. 58, pp. 377-382, May 1, 2003 2003.
[9] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data," *Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans,* vol. 41, pp. 552-568, May 2011.
[10] V. Lopez, A. Fernandez, S. Garcia, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information Sciences,* vol. 250, pp. 113-141, Nov 20 2013.
[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.