Model comparison in Emergency Severity Index level prediction

Seifu J. Chonde a,*, Omar M. Ashour b, David A. Nemhhard a, Gül E. Okudan Kremer a,b

a The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA
b School of Engineering Design, The Pennsylvania State University, University Park, PA 16802, USA

A R T I C L E  I N F O
Keywords:
Neural networks
Naive Bayesian networks
Ordinal logistic regression
Emergency Severity Index (ESI)

A B S T R A C T
Emergency Department (ED) triage is a process of determining illness severity and accordingly assigning patient priority. The Emergency Severity Index (ESI) is a 5-level acuity categorization system that aids in triage. This paper compared the capabilities of predicting ESI level using ordinal logistic regression (OLR), artificial neural networks (NNs), and naive Bayesian networks (NBNs). Data were obtained from Susquehanna Williamsport Hospital for 947 patients over a one month period in 2008. It contained the assigned ESI level, chief complaint, systolic blood pressure, pulse, respiration rate, temperature, oxygen saturation level (SaO₂), age, gender, and pain level. An OLR model was fit using a subset of these covariates. NBNs and NNs were modeled to relax the inherent assumptions of linearity and covariate independence in logistic regression. These three techniques were compared using incremental training dataset sizes between 50% and 100% of given data. All models were >60% accurate using the entire dataset for training. It was found that NBNs and NNs were robust to data size changes and all models had evaluation speeds of less than 0.5 s. At this time the use of NBNs is recommended considering speed, accuracy, data utilization, model flexibility, and interpretability of the model.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction and motivation

Emergency Departments (EDs) are specialized facilities for the acute care of patients who present themselves without prior appointments. Patients arrive at an ED either via personal means of transport or via ambulance. Patients arrive experiencing an array of complaints. Some of these complaints are life-threatening and require immediate attention. Patients with other complaints are considered non-urgent and can wait until medical resources are available to be treated. Classification of the acuity of complaints is on a patient-by-patient basis and is known as medical triage.

Patient prioritization via triage is a critical task as there is a large resource competition due to crowding in EDs. In the United States, in 2006 there were 119.2 million visits to EDs and this number continues to increase while the number of EDs decreases (Pitts, Niska, Xu, & Burt, 2008). Multiple triage schemes are used throughout the world. The United States commonly uses Emergency Severity Index (ESI). With this triage algorithm, patients are assigned an ESI level 1 (most urgent) to 5 (least urgent) considering the patients’ acuity, pain, and resource needs. In general, nurses perform triage and have extensive training, both practical and academic, in how to use the ESI rules and to best assess patient acuity. Misclassification may occur in the form of undertriage or overtriage. The effects of this misclassification may have severe consequences for waiting patients. For example in 2008, Beatrice Vance was experiencing pain and showed up at the Vista Medical Center Emergency Room in Lake County, Illinois (SoRelle, 2006). She was triaged and sent to the waiting room. When she was called a few hours later, she was found dead of an acute myocardial infarction.

Improving medical triage is challenging yet necessary. This paper compares three techniques in their efficiency of prediction of ESI level on a set of clinical data. Specifically, this paper uses the approach of Zhang and Bivens to compare the three techniques on two performance characteristics along with three other non-performance characteristics, which are provided below respectively (Zhang & Bivens, 2007).

- **Model accuracy given a dataset size**, which demonstrates data sizes necessary for algorithmic stability in prediction of ESI level (performance characteristic).
- **Model computation speed**, which enables the models to be built and processed in a small response time for the data (performance characteristic).
- **Data utilization**, which accounts for the amount of data input into the models and the amount of structural knowledge input into the models (non-performance characteristic).
- **Model interpretability**, which allows the model to be easily disseminated and the results understood with confidence by analysts (non-performance characteristic).

...
We compare and contrast naïve Bayesian networks (NBNs), artificial neural networks (NNs), and ordinal logistic regression (OLR) considering the above characteristics, and we develop several recommendations based on our analysis of data from a regional hospital. This work is part of an ongoing larger effort to develop an expert system to aid nurse decision making in real time. In other words, to transfer the currently used experience-based algorithm to a data-driven and technology assisted process. The paper is organized as follows: Section 1 defines the triage problem, and presents the related background literature. Section 2 explains the methodology of comparing the candidate model types. Section 3 presents the experimental results and discusses their impact. And finally, Section 4 presents conclusions and recommends the directions of future research.

1.1. Problem definition

Patient prioritization during medical triage is performed to assign an appropriate acuity rating or preference to each patient in accordance with the severity of his/her illness or injury; thus, the problem falls into the class of recommendation systems. Specifically, triage is a filtering system performed by triage nurses. Nurses consider each patient's characteristics, and then compare these characteristics to previously triaged patients and the systematic triaging rules of the ESI algorithm to assign an ESI level for the patient of interest. In the current study, we assume that the patients' assigned ESI levels in the clinical data set, collected from Susquehanna Williamsport Hospital in 2008, are the “absolute truth” representing the patients' condition with no misclassifications.

The objective of this paper is to compare and contrast the ability of three analytical techniques in ESI level prediction using the clinical dataset. Different triage systems are fairly reliable in assessing the most urgent triage priority but there is no consistency among these systems for the least urgent levels (Cooper, 2004; Fields, Claudio, Okudan, Smith, & Freivalds, 2009). Thus, it is important to improve the capability of the triage systems to make differentiations among the less urgent waiting patients within the same ESI classification. We envision this research as a step in comparing the efficiency and feasibility of using analytical techniques in ESI prediction to arrive at a triage decision support system. Three techniques that are compared are: (1) ordinal logistic regression (OLR), (2) naïve Bayesian networks (NBNs), and (3) neural networks (NNs).

1.2. Literature review

Conventional First-In-First-Out, processing does not work effectively in EDs because each patient is unique and patient complaints are not equally severe. Correspondingly, the ESI, the most common triage algorithm in the United States, was developed in 2000 by Wuerz et al. (2000). It relies upon nurses' judgment to assign a severity category to each patient. ESI 1 and ESI 2 are for critical patients where ESI 1 patients are considered as those that are severely critical (dying) and ESI 2 are those that are not dying but still in critical condition. ESI 3, 4, and 5 patients are those that are non-critical and can wait. Resource requirements including testing materials and time are used in addition to patient acuity to distinguish between ESI 3, 4, and 5 categories. Since its inception, the ESI algorithm has been updated four times to its current version, which is summarized in Fig. 1.

Triage assessment is a result of complex interactions between the health care provider (i.e., triage nurse) and patients (Cooper, 2004). The skills and personal qualities of the triage nurses are important for an effective triage (Andersson, Omberg, & Svedlund, 2006). These skills and contextual factors are not fully known to date due to the complexity of the triage process (Göransson, Ehnfors, Fonteyn, & Ehrenberg, 2008). Much of the decision making is mainly based on nurse's experience, knowledge, and intuition (Andersson et al., 2006; Patel, Gutnik, Karlin, & Pusic, 2008). Göransson et al. (2008) presented findings revealing that the triage nurse decision making during ED triage varied; some studies showed that there is a difference between experts and the beginner level nurses, while others found that less experienced and more experienced triage nurses' decision making was largely the same.

In addition to ESI, several other triage systems exist, including the Australasian (National) Triage Scale (ATS), the Canadian Triage and Acuity Scale (CTAS), and the Manchester Triage System from the United Kingdom, to name a few. We remark that while each system has its merits, our focus is on ESI as an example of the larger group of systems. Details and discussions of alternate systems are given by (van der Wulp, Hebe, Leenen, & Stel, 2011). Göransson, Ehrenberg, Marklund, and Ehnfors (2006) investigated the quality of triage nurses decision making. In their study, they used the Canadian Triage and Acuity Scale (CTAS) in Swedish EDs, and they studied the relationship between the personal characteristics of the registered nurses and the accuracy in their acuity rating of patient scenarios. They concluded that there is no relationship between personal characteristics of the registered nurses and their ability to triage, and they expected that the decision making might be affected by other factors.

Tanabe et al. (2005) stated that physicians and nurses face a serious limitation with ESI version 3, wherein one cannot determine how acutely ill the level 2 patients in the waiting room are, when they deal with the scenario of having “six level 2 patients in the waiting room”. Moreover, two levels of the ESI level 2 patients have been identified with the clinical experience; those who can safely wait for physician evaluation for at least 10 min without clinical deterioration, and those who cannot wait (Tanabe et al., 2005). It has been shown that computer applications...
designed to assist in triage provide a method of documentation and can assist in decision-making (Sadeghi, Barzi, Sadeghi, & King, 2006). There have been analytical and data mining techniques applied to the problem of ESI prediction. Specifically, regression has been used to determine the most important factors and BNs have been used to determine causal networks that determine posterior probabilities of being in each ESI category given a set of input data.

A series of BN models for triage decision support of patients complaining of non-traumatic abdominal pain was introduced by Sadeghi et al. (2006). Results indicate that the artificial triage system was as efficient as the emergency specialist and may be useful for telephone triage and triage in EDs. Claudio, Fields, Okudan, Smith, and Freivalds (2009) performed a multivariate linear regression considering age, chief complaint, pulse, and respiration rate to predict a continuous ESI value. The objective was to identify the statistically significant predictors and compare the performance of the created linear regression to an actual nurse response. The regression presented a systematic equation for breaking ties among patients triaged similarly by a nurse; however, it is not presented if this tie breaking procedure is accurate in the long-run. Additionally, the only statistically significant predictors were pulse and respiration rate.

A similar study was conducted by Garbez, Carrieri-Kohlman, Stotts, Chan, and Neighbor (2011) to identify specific factors used by triage nurses to differentiate ESI level 2 patients from ESI level 3 patients. The study consisted of 18 triage nurses between 27 and 55 of age with an average of 5 months in using ESI (range 0–7 months). Each time a triage nurse classified a patient as ESI 2 or ESI 3 they were asked to select 3 or 4 factors that allowed them to distinguish between the ESI 2 and ESI 3 barrier. The four statistically significant factors were determined to be chief complaint, vital signs, medical history, and an “other” factor. The “other” factor was patient clinical presentation 71% of the time.

Van der Wulp et al. (2011) published an OLR to predict ESI level using 544 patients seen at the ED of the University Medical Center of Utrecht, Netherlands. The authors used simple regression for each covariate to determine its significance. If a covariate was significant, then it was added to the multivariate model. The results presented indicate that vital signs, age, and chief complaint are significant predictors of ESI level using logistic regression. The authors did not split their dataset into training and validation subsets. Instead the authors used all data and performed testing indicating a Nagelkerke pseudo $R^2$ of 0.297, which was interpreted as the improvement of the full model over the null model in explaining the categorical variation of the ESI levels.

2. Emergency Severity Index prediction models

This section presents the data collection methods including a description of the data and rationale for what data were not used in the model building. This section also contains the background information on the modeling methodologies necessary to understand and implement the three types of models.

2.1. Data description

This is a retrospective study to develop prediction techniques for ESI level assignments during triage by emergency specialists. This study assumes that the assigned ESI level during triage is the true appropriate ESI level and that there is no undertriage or overtriage present in the data. The dataset used was obtained from Susquehanna Williamsport Hospital for 947 patients over a one-month period. Only one patient was observed with an ESI level 1 so this level is not considered in the analysis. The dataset has fewer ESI 1 patients than the 3% presented by Wuerz, Milne, Etel, Travers, and Gilboy (2000). Also, some patients did not have an ESI level recorded and were removed from the study. The resultant data set after cleaning the data included 870 patients. There were 100 (11.5%) ESI level 5 patients; 329 (37.8%) ESI level 4 patients; 363 (41.7%) ESI level 3 patients; and 78 (9.0%) ESI level 2 patients. Of these, 388 (44.7%) were male patients, 460 (52.8%) were female patients, and 22 (2.5%) did not indicate gender.

The chief complaints are summarized in Table 1. There are 19 chief complaint categories that range from 2 to 186 observations. Previous efforts to predict ESI level indicate there is a relationship between the chief complaint and the ESI level assigned, and thus we utilize complaint categories in our approach (Coiffi, 1998; Garbez et al., 2011; Sadeghi et al., 2006).

Vital signs were recorded before the patient saw a physician during the triage process. The vital signs recorded were systolic and diastolic blood pressure, pulse, respiration rate, temperature, and oxygen saturation level. These vital signs were binned according to expert nurse judgment as demonstrated by Ashour and Okudan (2011). The binning process categorizes the continuous vital sign values into bins of "High", "Relatively High", "Medium", "Relatively Low", and "Low" based on the severity of an observed value for the vital sign. Please refer to Ashour and Okudan (2011) for justification and presentation of the bins. The binning process is summarized in Table 2.

Table 2 indicates that over 51% of respiration rate data was missing from the dataset. In a typical data analysis all respiration rate data would be disregarded due to incompleteness; however, respiration rate was retained here because it was useful for the machine learning techniques studied. For all vital sign covariates not statistically significant predictors were pulse and respiration rate.

In the main experimentation, the data were used to build the models of interest. The data was split into training and testing sets for the OLR model building. The training dataset sizes were 50%, 60%, 70%, 80%, 90%, and 100% and the testing dataset sizes were the remaining amount. This was done such that a total amount of data of $n = 870$ was used for both training and testing. Each training dataset size was sampled 10 times and the results were averaged. This process was repeated for NBN model construction. In the NN, the training subset was broken into data for network construction (training) and data for network generalization (validation). 80% of the training data was used for network construction. Each model methodology is briefly presented below.

The above training and testing procedures were repeated starting from a random sample of 50% ($n = 435$) of the cleaned data. This was done to check if different trends in performance were found starting from a much smaller data size. This may be the case if one of the model types is unable to perform well with these smaller datasets or if one of the model types is more robust to data size variations. The results from using the smaller dataset were contrasted to the results from using the full dataset to determine if a different set of trends exists.
2.2. Model building

The following subsections provide details on the OLR, NBN, and NN models.

2.2.1. Ordinal logistic regression

Ordinal logistic regression is a regression model that extends the logistic regression model for dichotomous variables to more than two ordered response categories. Ordering implies that the response value follows some understood sequence. Moreover, ordinal logistic assumes proportional (linear) odds across the response categories such that the coefficients are constant across the ordinal levels and only the intercept changes as the response category changes.

In binary logistic regression, one may model the log-odds of an event using the logit function in only one form, namely: \( \text{logit}(\pi) = \log(\pi/(1 - \pi)) \) where \( \pi \) is the probability of the event of interest occurring. When there are more than two categories in a response variable multiple logits must be considered. If the response variable is ordinal then the proportional-odds cumulative logit model is of interest. Consider a response variable \( Y \) with naturally ordered categories, 1, 2, \ldots, \( N \). The probability of a given category being observed is denoted \( \pi_n \) and a cumulative probability of a response less than or equal to \( n \) is:

\[
P(Y \leq n) = \pi_1 + \cdots + \pi_n
\]

such that a cumulative logit is defined in Eq. (2):

\[
\log \left( \frac{P(Y \leq n)}{P(Y > n)} \right) = \log \left( \frac{\pi_1 + \cdots + \pi_n}{\pi_{n+1} + \cdots + \pi_N} \right)
\]

and successive cumulative logits are defined in Eq. (3):

\[
L_1 = \log \left( \frac{\pi_1}{\pi_2 + \cdots + \pi_N} \right)
\]

\[
L_2 = \log \left( \frac{\pi_1 + \pi_2}{\pi_3 + \cdots + \pi_N} \right)
\]

\[
L_{n-1} = \log \left( \frac{\pi_1 + \pi_2 + \cdots + \pi_{n-1}}{\pi_N} \right)
\]

This model is converted to the proportional odds cumulative logit model with appropriate \( p \) covariates and \( \alpha_n \) intercepts in Eq. (4):

\[
L_n = \alpha_n + \sum_{i=1}^{p} \beta_i X_i \quad n = 1, \ldots, N - 1
\]

In the ESI prediction problem the above equation is used to predict the odds of assigning a category that is at least as severe as \( n \). Note that in rare cases it may be advantageous to ignore inherent response ordering but in general it is understood that ordering should not be ignored. Interaction terms were not considered in this study because of rank deficiency in the data that did not permit accurate estimation of interaction effects.

2.2.2. Bayesian Networks

A naïve Bayesian network (NBN) is a directed acyclic graph (DAG) with \( n \) evidence nodes connected to a target node as shown in Fig. 2 below. The naïve Bayes model assumes that the evidence nodes are conditionally independent. This assumption leads to the joint probability:

\[
P(\text{ESI}, \text{Evidence}) \propto P(\text{ESI}|\text{Evidence}) \prod_{i=1}^{n} P(\text{Evidence}_i) \]

which is known as the naïve Bayes classifier (Russell & Norvig, 2009).

The evidence is the set of all data from the dataset. This model can be written as:

\[
P(\text{ESI}|\text{Evidence}) = k P(\text{ESI} = c) \prod_{i=1}^{n} P(\text{Evidence}_i = f_i|\text{ESI} = c)
\]

where \( k \) is a constant that represents the probability of simultaneously observing each piece of evidence. The data in this ESI recommendation problem is sparse. In particular, there are combinations of evidence that are not present for each ESI level. As an example, this data does not have a patient SaO2 level of “Relatively High” that was assigned an ESI level of 4; however, this combination is possible to occur. The model in Eq. (6) above indicates that the probability of observing this ESI level 2 patient is zero. This warrants correction. A simple, first-order Laplace smoothing is applied to prohibit the model from having zero probability for any unobserved combination of ESI and evidence. This changes the count based probabilities into pseudo-counts in the following way:

\[
P(\text{Evidence}_i = f_i|\text{ESI} = c) = (t + 1)/(N + j)
\]

where \( t \) is the number of data records, and the \( i \)th evidence node has the value \( f_i \) and ESI level \( c \). \( N \) is the cardinality of the dataset, and \( j \) is the number of levels of the \( i \)th evidence node.

The learning process involves calculating maximum likelihoods for the conditional probabilities given the observed data where an empirical prior is used. This ESI recommendation problem is
interested in a classification and not the individual estimates of the probability of being in an ESI level given the observed evidence. Once training is complete, the model can classify the likelihood of new data being in each class. Deterministic classifications are performed using the maximum a posteriori (MAP) decision rule by selecting the class with the highest probability. In the problem of classifying ESI using the five vital signs, the chief complaint, the age, the pain level, and the gender classification can be done as shown in Eq. (8).

\[
\text{classifier} = \text{argmax}_c P(\text{ESI} = c | \text{Evidence}) = \arg \max_c \prod_{i=1}^{n} P(\text{ESI} = c | \text{Evidence}_i)
\]

2.2.3. Artificial neural networks

Applications of NNs include constraint satisfaction problems, forecasting problems, industrial and manufacturing control problems, pattern recognition problems, and diagnostic problems. In medicine NNs have been used in areas such as medical diagnosis, imaging, and prediction of treatment or patient outcome (Drew & Monson, 2000). NNs’ widespread use is in part due to their flexibility in capturing high-order interactions that statistical techniques such as regression cannot easily observe.

Artificial neural networks (NNs) have been shown to be advantageous in recommendation systems and classification problems (Russell & Norvig, 2009). NNs are composed of a network of neurons similar to that of a brain. Neurons function by receiving inputs, performing logical processing based on thresholds, and sending differing output signals based on the comparison of the inputs to the thresholds. Networks of neurons function by weighting the input and output of upstream neurons to produce desired outputs of downstream neurons. The efficacy of NNs change with experience as the network adapts to different inputs. NNs can adapt to varying amounts of input data.

A history of NNs in the medical field is presented by Drew and Monson (2000). The authors indicate that the most common NN type is the multilayer feedforward perceptron (MLP). MLPs consist of an input layer, one or more hidden layers, and an output layer. MLPs are trained in a technique using backpropagation, a learning algorithm that attempts to minimize the least mean square difference between the desired output and the observed output over the entire training set. Simple backpropagation has been extended using various techniques. Scattered conjugate gradient backpropagation is used in this work for its superlinear convergence rate as indicated by the original work of Moller (1993). The scattered conjugate gradient algorithm terminates training when a maximum number of epochs is reached, a maximum time is exceeded, the performance has reached some criterion, the differential gradient of performance is below a minimum threshold for improvement in epochs, or the validation performance has continuously become worse (Demuth, Beale, & Hagan, 2007). In this work, the data is divided into training, validation, and testing datasets; where the training data sets up the network, the validation data refines the network and allows convergence to be reached, and the testing data estimates the applicability of the trained network to external data.

3. Results and discussion

Each modeling approach was setup and run using between 50% \((n = 435)\) and 100% \((n = 870)\) of the data in training with the remainder used in testing. In NNs the allocated fraction of training data is split with 80% of the training data used for model development (training) and 20% of the training data used for model generalization (validation). Average classification values were obtained from confusion matrices using 10 independent sample training datasets. A confusion matrix summarizes the classification predicted with the actual classification of the data.

The confusion matrix in Table 3 shows the various types of error made in an exemplary trained network. In our four-class prediction problem, the matrix is 5 \(\times\) 5 with the left-most 4 \(\times\) 4 referring to class predictions. The columns of this \(4 \times 4\) matrix represent the actual ESI level and show the distribution of predicted ESI level. The other 9 cells refer to prediction accuracy averages. The column averages show how often an actual ESI level is correctly predicted. The row averages show how often a predicted ESI level is correctly assigned to an actual ESI level. The corner average shows the overall accuracy on the dataset.

This procedure was repeated for each modeling methodology starting from 50% of the total data \((n = 435)\) and using between 50% \((n = 217)\) to 100% \((n = 435)\) of the data as training datasets. Results were similar to those presented here and are not displayed.

### Table 3

<table>
<thead>
<tr>
<th>Nurse’s assigned ESI level</th>
<th>ESI 5</th>
<th>ESI 4</th>
<th>ESI 3</th>
<th>ESI 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted ESI level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESI 5</td>
<td>48 (6%)</td>
<td>41 (5%)</td>
<td>20 (2%)</td>
<td>0 (0%)</td>
<td>109 (44%)</td>
</tr>
<tr>
<td>ESI 4</td>
<td>40 (5%)</td>
<td>192 (22%)</td>
<td>51 (6%)</td>
<td>10 (1%)</td>
<td>293 (66%)</td>
</tr>
<tr>
<td>ESI 3</td>
<td>12 (1%)</td>
<td>93 (11%)</td>
<td>273 (31%)</td>
<td>28 (3%)</td>
<td>406 (67%)</td>
</tr>
<tr>
<td>ESI 2</td>
<td>0 (0%)</td>
<td>3 (0%)</td>
<td>19 (2%)</td>
<td>40 (5%)</td>
<td>62 (65%)</td>
</tr>
<tr>
<td>Total</td>
<td>100 (48%)</td>
<td>329 (58%)</td>
<td>363 (75%)</td>
<td>78 (51%)</td>
<td>870 (64%)</td>
</tr>
</tbody>
</table>

#### 3.1. Ordinal logistic regression

Ordinal logistic regression was performed in R 2.14 using the package “ordinal” (Christensen, 2012). Simple, univariate regression was performed on each covariate and the results indicate that chief complaint, age, and saturated oxygen level should be considered in the multivariate case. Table 4 below shows the results of the multivariate model fit with the entire patient dataset. The reference categories are Abdomen Female for chief complaint; and, High for SaO2.

The results in Table 4 provide information on the relative odds for each category in relation to the reference category holding all else constant. For chief complaint this indicates that there is, on average, a larger probability of finding a patient with a Neurological complaint, a Psychiatric Suicide complaint, or a Seizure at an ESI level as severe as or more severe than the average patient with an Abdomen (Female) complaint. All other complaints have a smaller probability of finding a patient at an ESI level as severe as the average ESI level of an Abdomen (Female) patient. Additionally, this model purports that having a Saturated Oxygen level of High increases the average probability of having more severe ESI levels compared to having any other Saturated Oxygen level holding all else constant.

These results are intuitive in that having a high Saturated Oxygen level indicates that a patient is more likely to be in a severe state. Also, Neurological Complaints, Psychiatric Suicides, and Seizures are generally considered life threatening, so it is reasonable that they are classified as more severe than Abdomen (Female)
Table 4
Ordinal logistic regression model for entire dataset.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Est.</th>
<th>Std. error</th>
<th>Odds ratio</th>
<th>95% CI</th>
<th>Sig. value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant (5</td>
<td>4)</td>
<td>−7.87</td>
<td>1.52</td>
<td>0.000 ***</td>
<td></td>
</tr>
<tr>
<td>Constant (4</td>
<td>3)</td>
<td>−4.88</td>
<td>1.51</td>
<td>0.001 ***</td>
<td></td>
</tr>
<tr>
<td>Constant (3</td>
<td>2)</td>
<td>−1.12</td>
<td>1.49</td>
<td>0.449</td>
<td></td>
</tr>
<tr>
<td>Chief complaint Abdomen male</td>
<td>−0.07</td>
<td>0.44</td>
<td>0.933</td>
<td>(0.393, 2.214)</td>
<td>0.875</td>
</tr>
<tr>
<td>Alleged assault</td>
<td>−2.36</td>
<td>0.97</td>
<td>0.095</td>
<td>(0.014, 0.637)</td>
<td>0.015</td>
</tr>
<tr>
<td>Back pain</td>
<td>−2.68</td>
<td>0.39</td>
<td>0.069</td>
<td>(0.032, 0.147)</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>Chest pain</td>
<td>2.68</td>
<td>0.43</td>
<td>14.614</td>
<td>(6.263, 34.08)</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>Extremity</td>
<td>−2.71</td>
<td>0.33</td>
<td>0.066</td>
<td>(0.035, 0.126)</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>Eye ENT &amp; Dent</td>
<td>−4.56</td>
<td>0.36</td>
<td>0.011</td>
<td>(0.005, 0.021)</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>General medicine Headache</td>
<td>−1.28</td>
<td>0.29</td>
<td>0.277</td>
<td>(0.158, 0.487)</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>Head/face trauma</td>
<td>−0.2</td>
<td>0.42</td>
<td>0.820</td>
<td>(0.360, 1.868)</td>
<td>0.636</td>
</tr>
<tr>
<td>Multiple trauma</td>
<td>−1.42</td>
<td>0.51</td>
<td>0.243</td>
<td>(0.089, 0.663)</td>
<td>0.066 **</td>
</tr>
<tr>
<td>MVC Neurological</td>
<td>−1.06</td>
<td>0.59</td>
<td>0.346</td>
<td>(0.108, 1.107)</td>
<td>0.074</td>
</tr>
<tr>
<td>Pediatric</td>
<td>21.24</td>
<td>8829</td>
<td>1.7809</td>
<td>(0.001, 1.7809)</td>
<td>0.998</td>
</tr>
<tr>
<td>Pediatric illness</td>
<td>−2.31</td>
<td>0.36</td>
<td>0.099</td>
<td>(0.049, 0.199)</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>Pediatric injury</td>
<td>−2.2</td>
<td>0.43</td>
<td>0.111</td>
<td>(0.047, 0.259)</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>Psychiatric injury</td>
<td>0.15</td>
<td>0.52</td>
<td>1.16</td>
<td>(0.419, 2.16)</td>
<td>0.775</td>
</tr>
<tr>
<td>Respiratory Seizure</td>
<td>−1.6</td>
<td>0.41</td>
<td>0.201</td>
<td>(0.091, 0.446)</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>Seizure</td>
<td>0.39</td>
<td>0.87</td>
<td>1.476</td>
<td>(0.267, 1.666)</td>
<td>0.656</td>
</tr>
<tr>
<td>Skin rash abscess</td>
<td>−3.81</td>
<td>0.42</td>
<td>0.022</td>
<td>(0.010, 0.050)</td>
<td>0.000 ***</td>
</tr>
<tr>
<td>Saturated O₂</td>
<td>−4.04</td>
<td>1.49</td>
<td>0.018</td>
<td>(0.001, 0.327)</td>
<td>0.007 **</td>
</tr>
<tr>
<td>SaO₂-low</td>
<td>−3.52</td>
<td>1.64</td>
<td>0.030</td>
<td>(0.001, 0.740)</td>
<td>0.032 *</td>
</tr>
<tr>
<td>SaO₂-medium</td>
<td>−4.87</td>
<td>1.52</td>
<td>0.008</td>
<td>(0.000, 0.149)</td>
<td>0.001 **</td>
</tr>
<tr>
<td>SaO₂-missing</td>
<td>−4.76</td>
<td>2.16</td>
<td>0.009</td>
<td>(0.000, 0.590)</td>
<td>0.028 *</td>
</tr>
<tr>
<td>SaO₂-rel. high</td>
<td>−3.99</td>
<td>1.54</td>
<td>0.018</td>
<td>(0.001, 0.378)</td>
<td>0.010 **</td>
</tr>
<tr>
<td>Age</td>
<td>0.02</td>
<td>0.00</td>
<td>1.025</td>
<td>(1.017, 1.033)</td>
<td>0.000 ***</td>
</tr>
</tbody>
</table>

Sig. codes: 0 ** 0.001 *** 0.000 * 0.05 † 0.01 †† 0.05 ††† 1.

Table 5
Example classification matrix for OLR.

<table>
<thead>
<tr>
<th>ESI 5</th>
<th>ESI 4</th>
<th>ESI 3</th>
<th>ESI 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESI 5</td>
<td>33 (4%)</td>
<td>19 (2%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>ESI 4</td>
<td>54 (6%)</td>
<td>220 (26%)</td>
<td>70 (8%)</td>
<td>8 (1%)</td>
</tr>
<tr>
<td>ESI 3</td>
<td>11 (1%)</td>
<td>84 (10%)</td>
<td>280 (32%)</td>
<td>32 (4%)</td>
</tr>
<tr>
<td>ESI 2</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>13 (2%)</td>
<td>38 (4%)</td>
</tr>
<tr>
<td>NA*</td>
<td>2 (0%)</td>
<td>6 (1%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>100 (100%)</td>
<td>329 (67%)</td>
<td>363 (72%)</td>
<td>78 (49%)</td>
</tr>
</tbody>
</table>

* Model unable to make predictions for 8 patients because of missing age data.

3.2. Naïve Bayesian networks

The recommendation system for ESI levels can be constructed as a NBN as shown in Fig. 3. The NBN classification was done in R 2.14 using the package “e1071” (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2012). The probability of being in each ESI level was output and the most probable class was selected as the deterministic prediction. These values were compared to the actual ESI level and confusion matrices were developed. Table 6 shows the overall confusion matrix for the NBN model.

3.3. Artificial neural networks

All linearly separable problems can be solved using a multilayer perceptron (MLP) that consists of one input layer, one hidden layer, and one output layer. The inclusion of NNS in this work is to determine whether there are high-order interactions not detected by regression or NBNs. For this reason, two hidden layers are used despite the unconventionality of this structure. Indicator variables were coded for each categorical variable. There were 57 indicator variables for the categorical variables and 2 continuous variables for a total of 59 input variables. Ten neurons were used in both the first hidden layer and in the second hidden layer. There were 4 neurons used in the output layer. Each output represented a binary classification into one of the classes ESI 2 through ESI 5. The Neurol Network Toolbox in MATLAB was used to train the neural networks (Demuth et al., 2007). Table 7 shows an overall confusion matrix for the NN model using 100% of the data in training and validation. Specifically, 20% of the data is used in validation and 80% of the data is used in training.

3.4. Model comparison

Models are compared on two measures of performance and three measures of non-performance similar to that of Zhang and Bivens (2007). The two measures of performance are model accuracy and speed of model evaluation. The three measures of non-performance are model structure, model interpretability, and
model flexibility. These five measures of comparison are presented in order below, followed by a discussion of results.

3.4.1. Accuracy

Accuracy in classification models is measured using confusion matrices. Confusion matrices are often presented separately for training data, testing data, and all data. For NNs there often is an additional confusion matrix for the validation data. We also consider overall analytical accuracy for the different sizes of training dataset considered. Table 8 illustrates that on average, NNs and NBNs perform similarly when using 50% or 100% of the data for training. However, regression, performed poorly with smaller training dataset ratios. When all data is used in training no substantial difference in performance was found.

Table 8 indicates that OLR was not robust to small datasets. This is seen by the low average accuracy of prediction on the testing data. The increasing trend in the regression suggests that accuracy for OLR was good with a sufficiently large training dataset. The performance of the NBNs is slightly more accurate on average than the performance of the NNs. This process was repeated starting from 50% of the data \((n = 435)\). The table of classification accuracy contained similar results. Thus, with respect to model accuracy, the NN approach was most preferred, followed by BN, and then OLR.

3.4.2. Model evaluation speed

The models were compared based on their speed of evaluation in training over training dataset sizes of 50–100% using the cleaned patient dataset. The speed of evaluation was averaged over 60 simulation runs for each model type with 10 runs at each decile between 50% and 100% training dataset size. Random training data was selected without replacement for each of the runs with less than 100% training dataset size. This selection was not included in speed calculations. Fig. 4 indicates the average results for all trials.

These results suggest that there is a slight linear increase in time for the regression model with the amount of data but the NBNs and NNs have constant time over the dataset sizes considered. Also, NN’s took about 0.5 s on average, regression took about 0.075 s, and NBN took at most 0.01 s. While, all times are relatively short for the dataset considered in this study, the most preferred model for model with respect to evaluation speed is the NBN, followed by OLR, and then NN.

3.4.3. Data utilization

Of the nine recorded patient variables, univariate regression found only SaO2 level, age, and chief complaint to be statistically significant. The multivariate model was selected using all univariate regression models that had statistically significant predictors. Thus, the multivariate ordinal logistic models in this study use only one-third of the collected data. This is a limitation of the logistic regression model that may ultimately limit its predictive abilities.

Data mining techniques such as NNs and BNs are designed to learn relationships among data that are not necessarily obvious and might remain unnoticed by conventional statistical methods such as regression. It is not surprising that these two approaches incorporate all nine recorded variables. These two data mining techniques can be represented as graphs, which permit incorporation of causal information as indicated by arcs of the graph. BNs have a straight forward interpretation as casually linked DAGs (Zhang & Bivens, 2007). This implies that if specific model knowledge is available from an expert, then it may be readily integrated into a BN.

Specific domain knowledge might help improve a BN’s performance. However, such knowledge will make the network more complex than the naïve networks used in this work. NNs do not possess easy structures for domain experts to incorporate expertise; therefore, with regards to data utilization BNs are preferred to NNs. With respect to the utilization of the overall dataset, BN’s were most preferred, followed by NN, and OLR.

3.4.4. Model interpretability

Model interpretability, which can be both visual and numerical, is particularly valuable because it allows users of a model to easily understand the underlying principles and relationships of the covariates to the output. Also, models that are easily understood are more likely to be adopted.

Both the NBN and NN models have straightforward graphical representations of variables. The NBN model is the easiest to interpret visually because it is a directed graph with arcs that convey a casual, conditional relationship (Zhang & Bivens, 2007). The NN model is often referred to as a black box. The interpretability of a trained network is difficult and requires understanding of the trained weights, which may have no literal meaning (Drew & Monson, 2000). Visualization methods have been developed but require additional processing and time in order to explain the covariate-response relationships. For additional details on these methods refer to the reference of Zhang and Bivens (2007).

The OLR model is the easiest to interpret. OLR is a linear model. It does not have a graphical representation similar to that of NBN and NN, but one is not necessary for explanation. One may plot the 95% confidence intervals of the odds ratios of each level in an OLR and assess if the confidence interval includes the 1.0 mark. If it crosses this mark then the variable is not substantively useful at the 95% confidence level because one is unsure if the variable

---

Table 8. Average accuracy of predictions by training size.

<table>
<thead>
<tr>
<th>Training size (%)</th>
<th>Naïve Bayes network</th>
<th>Logistic regression</th>
<th>Neural network</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training (%)</td>
<td>Testing (%)</td>
<td>Overall (%)</td>
</tr>
<tr>
<td>50</td>
<td>67.0</td>
<td>58.9</td>
<td>63.0</td>
</tr>
<tr>
<td>60</td>
<td>64.9</td>
<td>57.9</td>
<td>62.1</td>
</tr>
<tr>
<td>70</td>
<td>64.3</td>
<td>59.2</td>
<td>62.7</td>
</tr>
<tr>
<td>80</td>
<td>64.5</td>
<td>60.1</td>
<td>63.6</td>
</tr>
<tr>
<td>90</td>
<td>64.6</td>
<td>60.2</td>
<td>64.2</td>
</tr>
<tr>
<td>100</td>
<td>–</td>
<td>–</td>
<td>63.6</td>
</tr>
</tbody>
</table>

a NNs use 20% of data allotted to training as data for validation.
increases the odds of the ordinal response (greater than 1.0 odds) or decreases the odds of the ordinal response (less than 1.0 odds). An introduction to odds and the proportional odds assumption may be appropriate, but regression is such a common technique it generally requires only a brief explanation. Thus, with respect to the interpretability of the approaches, OLR was most preferred, followed by BNs, and then NNs.

3.4.5. Model maintainability and flexibility

Data handling in EDs often relies on manually input records. With this, there can be data errors, so it is important to select a methodology that is maintainable and flexible in this environment. The results of both NBNs and NNs above indicate that they are robust to smaller data sizes while regression was not. In this sense NBNs and NNs are flexible modeling methodologies of interest. NNs are especially adept at capturing high-order interactions among variables. Also, it has been said that Bayesian networks have a long memory where they do not easily forget previous data (Zhang & Bivens, 2007). This can be problematic when data is input over time and serial correlations exist. Certain Bayesian networks can handle this; however, the NBN model used in this work does not handle serial correlation. With respect to maintainability and flexibility, NNs were most preferred, followed by BNs and OLR.

3.4.6. Summary of comparisons

We discussed two performance based characteristics and three non-performance based characteristics of BNs, NNs, and OLR on our comparison of these on triage decision-making. The two performance characteristics are perhaps more important than the others because a model that does not perform well will have less overall utility, and may not be used at all for prediction. The non-performance based characteristics are secondary tie-breakers for approaches that are given full consideration based on performance (Table 9).

The experiment in this study shows that NNs using two hidden layers are more accurate, yet slower than NBNs; yet, both models are acceptable with respect to accuracy and speed. Regression with the entire data set has acceptable accuracy and evaluation speed, indicating that regression may not be as useful for small datasets. NBNs and NNs are more robust to the data size and should be considered further. For completeness, regression was also considered for non-performance characteristics. Clearly NBNs have the fastest evaluation speed; however, the difference between all model speeds and the exhibited trends does not provide a reason to suggest one model over another. Instead, additional considerations of the models are summarized separately below:

Considerations of BNs:

- Relationships across variables in the network are not prescribed.
- Output is a probability of recommending each class rather than a deterministic prediction.
- Calculations are simple and have low evaluation speed with respect to other models.
- NBN learning systems using Laplace smoothing have low difficulty with absent factor levels in the training set and produce non-zero predictions.
- More complex BNs than the naive method used here, such as a Tree Augmented naive Bayesian Network (TAN) or the noisy-or BN, may outperform NBNs.

Considerations of NNs:

- Approach assumes linearity and proportional odds fit the data.
- Performs poorly on small datasets. Conversely, statistical techniques tend to perform poorly on “big data” datasets as well.
- Numerically simple and easy to interpret.
- Used as a standard comparison for machine learning techniques.

Considerations of OLR:

- Approach assumes linearity and proportional odds fit the data.
- Performs poorly on small datasets. Conversely, statistical techniques tend to perform poorly on “big data” datasets as well.
- Numerically simple and easy to interpret.
- Used as a standard comparison for machine learning techniques.

4. Conclusion

This paper presented a comparison of NBNs, NNs, and OLR in ESI level recommendation for nurses during medical triage. It was found that the NBNs were the fastest models and NNs were the slowest. The NNs and NBNs were found to be robust to various training dataset sizes while the ordinal logistic regression was found to perform poorly at low training dataset sizes. The difference in performance between the NNs and NBNs is not substantively significant.

No technique predicted the ESI level given by the nurses with an overall accuracy greater than 68%. This suggests there may be omitted variables that affect a nurse’s judgment as to the proper ESI categorization for a patient. The low accuracy may have occurred because the dataset did not include combinations of every factor for every variable, thus limiting the ability of the models to accurately estimate interaction effects. In particular, the regression model was unable to estimate any interaction effects due to rank deficiency.

In absence of physical law or significant structural knowledge, it is necessary to capture as much data as possible to model a system. The classification of patient acuity using the ESI is a complex problem that is not easily conducted without many streams of data. In a survey study conducted by Garbez et al. (2011) it was found that chief complaint, vital signs, medical history, and extraneous factors contribute most significantly to nurses’ judgment during triage. However, the survey results list that 20 total factors were indicated to contribute most significantly to nurses' judgment during triage. For this reason, it is recommended that ordinal logistic regression is revisited with a larger dataset that incorporates the other suggested variables. Additional data may also bolster the power of the ordinal logistic regression. For the given dataset, only three of nine covariates are statistically significant at the 0.05 significance level in univariate regression. It is possible that additional data may improve the significance of other covariates.

In future testing the NBN may be replaced with a more complicated BN such as the Tree Augmented naive BN or noisy-or
BN. Additionally, an integrated approach using logistic regression as a prior for a BN will be explored for its improvement in accuracy. NNs proved to be a robust, fast modeling methodology for this problem. However, their lack of interpretability may limit their acceptance in the medical community as a triage support tool.

References


