Predicting Sepsis: A Comparison of Analytical Approaches

Femida Gwadry-Sridhar[1], Ali Hamou[1], Benoit Lewden[1], Claudio Martin[2], and Michael Bauer[3]

[1] I-THINK Research Lab, Lawson Health Research Institute, London, ON Canada
[2] Dept of Medicine and Physiology, University of Western Ontario, London, ON Canada
[3] Dept of Computer Science, University of Western Ontario, London, ON Canada
femida.gwadry-sridhar@lhsc.on.ca, ali.hamou@sjhc.london.on.ca, benoit.lewden@lawsonresearch.com, cmartin@lhsc.on.ca, bauer@csd.uwo.ca

Abstract. Sepsis is a significant cause of mortality and morbidity and is often associated with increased hospital resource utilization, prolonged intensive care unit and hospital stay. With advances in medicine, there is now aggressive goal oriented treatments that can be used to help patients that may be at risk for sepsis. To predict this risk, we hypothesized that commonly used univariate and multivariate models could be enhanced by using multiple analytic methods to providing greater precision. As a first step, we analyze data about patients with and without sepsis using multiple regression, decision trees and cluster analysis. We compare the predictive accuracy of the three different approaches in predicting which patients are likely (or not likely) to develop sepsis. The precision analysis suggests that decision trees may provide a better predictive model than either regression methods or cluster analysis.

1 Introduction

Sepsis is defined as infection plus systematic manifestations of infection [1]. Severe sepsis is considered present when sepsis co-exists with sepsis-induced organ dysfunction or tissue hypo-perfusion [1]. Sepsis can result in mortality and morbidity, especially when associated with shock and/or organ dysfunction [2]. This can be associated with increased hospital resource utilization, prolonged intensive care unit (ICU) and hospital stay, decreased long-term health related quality of life and an economic burden estimated at US $17 billion each year in the United States alone [3]. In Canada, there is limited data on the burden of severe sepsis; however, costs in Quebec may be as high as $73M per year [4], which contribute to estimates of total Canadian cost of approximately $325M per year.

Patients with severe sepsis generally receive their care in the ICU. A multinational study of sepsis in teaching hospitals found that severe sepsis or septic shock is present or develops in 15% of ICU patients [5]. However, diagnosing sepsis is difficult because there is no “typical” presentation despite published definitions for sepsis [6-7]. In the Canadian Sepsis Treatment and Response (STAR) registry [8] the total rate for severe sepsis was 19%. Of these, 63% occurred after hospitalization.

M. Szomszor and P. Kostkova (Eds.): E-Health 2010, LNICST 69, pp. 95–102, 2011.

There are now aggressive goal oriented treatments that can be used to help these patients [9-10]. If researchers were able to predict which patients may be at risk for sepsis, treatment could start early and potentially reduce the risk of mortality and morbidity. Therefore, new methods to help with the early diagnosis of patients who are either present with sepsis or develop sepsis in hospital are needed.

Such methods include a variety of analysis techniques that can be used to identify relationships among a set of measures. We hypothesized that analytical methods currently used in clinical research to determine the risk of a patient developing sepsis may be further enhanced by using multi-modal methods that together could be used to provide greater insight and precision. Researchers commonly use univariate and multivariate regressions to gather information about variables that are associated with the dependent variable, which in this case is whether or not a patient developed sepsis. However, at times these models are constrained as we either use univariate analysis to guide our decision on which variable to include, or we rely on the literature to guide the variable selection.

Our previous work had looked at the use of regression techniques to develop a linear predictive model [11]. In that work, we explored which variables emerged as key in predicting the likelihood of sepsis. We also considered the use of decision tree analysis and cluster analysis. Decision trees provide a prescriptive approach for arriving at a decision with an associated probability. In contrast, cluster analysis takes a holistic approach to partition the data into similar but disjointed sets. In this paper, we report on the accuracy of these decision tree and cluster analysis models in predicting the likelihood of sepsis or not.

2 Data in Study

We obtained data[1] that was collected from 12 Canadian intensive care units that were geographically distributed and included a mix of medical and surgical patients [8]. Data were collected on all patients admitted to the ICU who had an ICU stay greater than 24 hours or who had severe sepsis at the time of ICU admission. Patients who were not anticipated to obtain to receive active treatment were excluded.

Hospitals collected a minimum data set on all eligible patients admitted to the ICU [12-13]. This included demographic information and data about their admission, source of admission, diagnosis, illness severity, outcome and length of ICU and hospital stay. Illness severity scores were calculated using data obtained during the first 24 hours in the ICU [14-15]. All patients were subsequently assessed on a daily basis for the presence of infection and severe sepsis. The characteristics of the patients have been described elsewhere and are summarized in Table 1.

---

[1] **Ethical Review, Funding and Data Ownership.** The study was approved by the University of Western Ontario Research Ethics Board and the need for informed consent was waived. Participating institutions submitted the study to their review process if local approval was required. All activities were compliant with the privacy and confidentiality practices of the participating institutions and the Federal and Provincial governments of Canada. Eli Lilly Canada provided a research grant to London Health Sciences Centre to support the trial. Data is owned by and resides with London Health Sciences Centre.

Table 1. Baseline characteristics of patients with severe sepsis

	All (n=1238)	Community cases (n=458)	Hospital cases (n=305)	Early ICU cases (n=195)	Late ICU cases (n=280)
Predisposition					
Age*	61.2 (16.5)	59.1 (16.6)	63.5 (15.3)	60.6 (17.8)	62.6 (16.2)
Sex – Female (%)	40.2	47.3	38.7	34.9	33.6
ICU Admission Diagnosis (%)					
Cardiac Arrest	3.8	5.2	3.9	3.1	1.8
Aspiration Pneumonia	3.3	3.1	5.6	2.1	2.1
Respiratory Arrest	5.0	2.8	10.2	6.7	2.5
Bacterial/Viral Pneumonia	9.5	14.4	10.5	6.7	2.5
COPD	2.0	2.2	1.3	2.1	2.5
Respiratory-Other	6.5	3.9	7.9	7.7	8.6
Sepsis-Non Urinary Tract	13.9	24.5	14.8	3.6	2.9
Sepsis-Urinary Tract	3.1	5.5	3.0	2.1	0
Trauma	6.6	3.5	1.7	10.8	13.9
Renal Disease	1.8	1.3	1.6	2.1	2.5
GI-Perforation / Obstruction	5.5	5.1	7.6	5.7	4.0
Other	39.0	28.6	32.1	48.7	56.8
Patient Classification – Surgical (%)	33.4	20.5	33.6	45.0	46.0
Co-morbid Conditions (%)					
Any	30.9	32.3	35.7	29.2	24.3
Multiple (>1)	4.5	4.8	4.6	4.6	3.9
Immunosuppression	13.3	12.0	17.7	14.4	9.6
Liver (Hepatic/Cirrhosis)	2.6	3.5	3.0	1.5	1.4
Congestive Heart Failure	6.9	6.6	9.2	5.1	6.1
Chronic Lung Disease	7.0	7.4	6.9	5.6	7.1
Renal Failure	6.1	8.1	4.3	7.2	4.3
Time Between Hospital & ICU Admission (Days)*	4.6 (12.1)	0.2 (0.4)	12.8 (19.5)	3.4 (8.5)	3.5 (9.1)
APACHE II Score*	24.9 (8.6)	26.0 (9.3)	25.5 (8.4)	24.0 (7.6)	22.9 (7.9)
Predicted Mortality (APACHE II)*	0.48 (0.26)	0.53 (0.27)	0.52 (0.26)	0.43 (0.25)	0.39 (0.24)

* [mean (SD)]

The management of severe sepsis requires prompt treatment within the first six weeks of resuscitation [1]. Intensive care specialists agree (supported by literature) that early goal-directed resuscitation has been shown to improve survival in patients presented to emergency rooms with septic shock [1]. The purpose of our study was to determine whether different analytic methods including cluster analysis and decision trees can provide markers for early diagnosis that can be used for early interventions.

3 Overview of Analytic Approaches

We provide a brief overview of the analytical approaches used in the analysis of the sepsis data and briefly summarize the key variables in each of the models; more details on the models and a discussion of the variables can be found in [11].

3.1 Decision Tree

A decision tree is a predictive model, that is, a mapping from observations about an item resulting in conclusions about its target value [16]. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. A decision tree is made from a succession of nodes, each splitting the dataset into branches. Generally, the algorithm begins by treating the entire dataset as a single large set and then proceeds to recursively split the set. Three popular rules are typically applied in the automatic creation of classification trees. The *Gini* rule splits off a single group of as large a size as possible, whereas the *entropy* and *two-ing* rules find multiple groups comprising as close to half the samples as possible. The algorithms construct the tree from the "top" down until some stopping criterion is met. In our current approach, we have used the gain in entropy, which accurately models the physical system evolving spontaneously toward equilibrium, in order to determine how to best create each node of the tree.

In order to define information gain precisely, we used a measure called entropy, that characterizes the "purity" (or, conversely, "impurity") of an arbitrary collection of examples. Generally, given a set *S*, containing only positive and negative examples of some target concept (the so-called two-class problem), the entropy of set *S* relative to this simple, binary classification is defined as:

$$Entropy(S) = - p_p \log_2 p_p - p_n \log_2 p_n \tag{1}$$

where p_p is the proportion of positive examples in *S* and p_n is the proportion of negative examples in *S*.

Given that entropy is a measure of the impurity in a collection of training examples, we can then define the effectiveness of an attribute in classifying the data. The measure we will use, called information gain, is simply the expected reduction in entropy caused by partitioning the examples accordingly. More precisely, the information gain, *Gain* (*S, A*) of an attribute *A*, relative to a collection of examples *S*, is defined as:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{v \varepsilon values\ (A)} \frac{|S_v|}{|S|} Entropy\ (S_v) \tag{2}$$

where *values(A)* is the set of all possible values for attribute *A*, and S_v is the subset of *S* for which attribute *A* has value *v* (i.e., $S_v = \{s\ \varepsilon\ S \mid A(s) = v\}$). In other words, *Gain*(*S,A*) is the information provided about the target attribute value, given the value of some other attribute *A*.

Continuous variables, such as temperature, require a somewhat special approach. This is accomplished by dynamically defining new discrete attributes that partition the continuous attributes into a discrete set of intervals. In particular, for an attribute *A*

that is continuous, the algorithm can dynamically create a new Boolean attribute *Ac* that is true if $A < c$ and false otherwise. The only question is how to best select the value for the threshold *c*. This is accomplished by selecting values for the threshold based on the existing values of the attribute *A* and computing the gain. The threshold *c* that produces the greatest information gain is then chosen.

The process of selecting a new attribute and partitioning the training examples is then repeated for each non-terminal descendant node in the tree, only the training examples associated with that node. Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree. This process continues for each new leaf node until either every attribute has already been included along this path through the tree, or the training examples associated with this leaf node all have the same target attribute value (i.e., their entropy is zero).

3.2 Cluster Analysis

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait - often proximity according to some defined distance measure. The K-means algorithm [17] assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster – that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. The k-means clustering algorithm was used in this work due to its flexibility and speed of execution on large datasets.

For our cluster analysis we developed a slightly modified version of the K-means clustering algorithm. The regular algorithm was not useful as we needed to introduce a distance measure relevant to our problem. Our data are grouped into categories, where each variable was essential a discrete point in a space having its own dimension and base (the dimension of each variable is simply the number of categories available). We used a distance measure roughly equivalent to the 2-norm of the cross product of two vectors. Basically, this produces a distance measure in which the distance between two points will be “zero” if they are in the same category and “one” otherwise. This way, we can quantify the distance between points without introducing a hierarchy between categories. The basic algorithm can therefore follow the standard K-means algorithm.

4 Accuracy of Methods

By examining multiple analytical approaches, we can help identify the key variables that should be used to develop predictive models of which patient may or may develop sepsis. However, it is also critical to look at the precision or accuracy of the models in prediction. In our earlier work [11], looking at regression analysis resulted in a model able to classify patients not likely to get sepsis with very good accuracy (99%) and a reasonable accuracy (66%) in predicting patients that were likely to get sepsis. We now examine the accuracy in the decision tree and cluster analysis methods.

Models were built from a subset of our patient database of 23,000 patients. For both the decision tree and cluster analyses, we randomly selected different groups of patients of different sizes and built decision trees and clusters. We then used the remaining patients to measure the accuracy of the respective predictive models, that is, for each of the remaining patients, we determined whether the decision tree or clusters would accurately predict whether the patient developed sepsis or not. In both cases, we randomly selected 10 different groups of each group size.

For the decision trees, as indicated, we randomly created groups of 50, 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000, and 6000 patients. For each of the ten groups at each size, we determined the accuracy. The results are presented in Table 2.

Table 2. Precision of Decision Tree Models

	Sepsis Patients		Non-Sepsis Patients	
Group Size	Mean	Std. Dev.	Mean	Std. Dev.
50	0.8204	0.3240	0.9870	0.0155
100	0.8762	0.1895	0.9938	0.0088
200	0.8843	0.1378	0.9878	0.0080
500	0.8839	0.0645	0.9881	0.0060
1000	0.9022	0.0459	0.9872	0.0033
2000	0.8685	0.0413	0.9880	0.0027
3000	0.8507	0.0382	0.9890	0.0020
4000	0.8706	0.0289	0.9871	0.0025
5000	0.7983	0.0325	0.9900	0.0030
6000	0.8688	0.0267	0.9875	0.0018

Table 3. Precision of Decision Tree Models

	Sepsis Patients		Non-Sepsis Patients	
Group Size	Mean	Std. Dev.	Mean	Std. Dev.
50	0.7852	0.2636	0.9667	0.0325
100	0.6947	0.2953	0.9672	0.0137
200	0.7862	0.0826	0.9653	0.0203
500	0.7551	0.1112	0.9608	0.0112
1000	0.7900	0.0849	0.9638	0.0053
2000	0.7071	0.0596	0.9645	0.0116
3000	0.7495	0.0488	0.9670	0.0101
4000	0.6988	0.0402	0.9670	0.0054
5000	0.7111	0.0523	0.9680	0.0069
6000	0.7288	0.0492	0.9615	0.0141
7000	0.7327	0.0599	0.9619	0.0129
8000	0.7270	0.0674	0.9655	0.0058
9000	0.6800	0.0657	0.9711	0.0083
10000	0.6954	0.0729	0.9684	0.0127
11000	0.6934	0.0474	0.9682	0.0094
12000	0.7187	0.0430	0.9631	0.0126
13000	0.6970	0.0686	0.9661	0.0091
14000	0.6590	0.0801	0.9668	0.0127
15000	0.7271	0.0748	0.9618	0.0145

In the case of the cluster analysis, we used group sizes of 50, 100, 200, 500, 1000, 2000, 3000, 4000, 5000, …, 15000 patients and randomly created ten groups of each size. Table 3 summarizes the accuracy.

It is interesting to note that both approaches provided prediction accuracy similar to the regression analysis in the case of non-sepsis patients. The cluster analysis seemed to be somewhat better at predicting patients likely to develop sepsis over the regression analysis. The decision tree, however, was substantially better, typically providing accuracy in the 80%-90% range for patients likely to develop sepsis.

Both approaches had relatively low standard deviations once the sample size reached 1000~2000. This makes is understandable since a random selection of patients for a group would have significantly more non-sepsis patients than sepsis patients, hence a sample large enough to include enough sepsis patients would be needed in order for the models to be accurately developed.

5 Conclusion and Future Work

From a methodological perspective, it appears that decision trees provide the most precision when determining sepsis risk. This is based on point estimates of variables (as described in [11]) that are easily measured at the time of hospital admission, which is an important consideration for decision support tools applied at bedside.

The cluster analysis also illustrated specific patient features that are very telling and ominous. The cluster analysis suggested that patients with poor urinary output and oxygenation and have a low temperature (<36) are very much at risk for sepsis. Although the argument could be made that clinicians with excellent acumen would recognize these signs regardless of mathematical models, the fact is that with hectic emergency rooms and the heterogeneity of staff expertise, decision support and algorithms can make the difference between life and death for the patient.

What is also clear is that traditional models used in medicine, such as regressions, may not provide sufficient precision and act primarily as a starting point for further analysis.

Future research will focus on the following:

- The utilization of multiple methods to develop a collective robust model that will be validated against observational cohort data and tested in clinical trials.
- The development of methods that use a “composite index” such that the estimate of relative risks of a poor outcome is quantifiable.
- The development of methods to test the “number needed to determine risk” – that is, how many clusters are required to determine risk, which at the same time minimizing the false negative rate.
- Determination of whether the frequency of variable present in multiple models adds to the precision.
- Investigation of temporal functions in key variables and the impact on precision.

References

1. Dellinger, R.P., Levy, M.M., Carlet, J.M., et al.: Surviving Sepsis Campaign: international guidelines for management of severe sepsis and septic shock: 2008. Crit. Care Med. 36(1), 296–327 (2008)
2. Angus, D.C., Linde-Zwirble, W.T., Lidicker, J., Clermont, G., Carcillo, J., Pinsky, M.R.: Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit. Care Med. 29(7), 1303–1310 (2001)
3. Brun-Buisson, C., Doyon, F., Carlet, J., et al.: Incidence, risk factors, and outcome of severe sepsis and septic shock in adults. A multicenter prospective study in intensive care units. French ICU Group for Severe Sepsis. JAMA 274(12), 968–974 (1995)
4. Letarte, J., Longo, C.J., Pelletier, J., Nabonne, B., Fisher, H.: Patient characteristics and costs of severe sepsis and septic shock in Quebec. J. Crit. Care 17(1), 39–49 (2002)
5. Alberti, C., Brun-Buisson, C., Burchardi, H., et al.: Epidemiology of sepsis and infection in ICU patients from an international multicentre cohort study. Intensive Care Med. 28(2), 108–121 (2002)
6. American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference: definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Crit. Care Med. 20(6), 864–74 (1992)
7. Levy, M.M., Fink, M.P., Marshall, J.C., et al.: SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. Crit. Care Med. 31(4), 1250–1256 (2001)
8. Martin, C., Priestap, F., Fisher, H., et al.: A prospective, observational registry of patients with severe sepsis: The Canadian Sepsis Treatment And Response (STAR) Registry. Crit. Care Med. (2009) (in press)
9. Rivers, E., Nguyen, B., Havstad, S., et al.: Early goal-directed therapy in the treatment of severe sepsis and septic shock. N. Engl. J. Med. 345(19), 1368–1377 (2001)
10. Minneci, P.C., Deans, K.J., Banks, S.M., Eichacker, P.Q., Natanson, C.: Meta-analysis: the effect of steroids on survival and shock during sepsis depends on the dose. Ann. Intern. Med. 141(1), 47–56 (2004)
11. Gwadry-Sridhar, F., Lewden, B., Mequanint, S., Bauer, M.: Multi-Analytical Approaches Informing the Risk of Sepsis. In: Biomedical Engineering Systems and Technologies. CCIS, pp. 394–406. Springer, Heidelberg (2010)
12. Critical Care Research Network: About CCR-Net (2005), `http://www.criticalcareresearch.net/`, `http://www.criticalcareresearch.net/`
13. Keenan, S.P., Martin, C.M., Kossuth, J.D., Eberhard, J., Sibbald, W.J.: The Critical Care Research Network: a partnership in community-based research and research transfer. J. Eval. Clin. Pract. 6(1), 15–22 (2000)
14. Knaus, W.A., Draper, E.A., Wagner, D.P., Zimmerman, J.E.: APACHE II: a severity of disease classification system. Crit. Care Med. 13(10), 818–829 (1985)
15. Knaus, W.A., Wagner, D.P., Draper, E.A., et al.: The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100(6), 1619–1636 (1991)
16. Quinlan, J.R.: Induction of decision trees. Machine Learning 1, 81–106 (1986)
17. Hartigan, J.A., Wong, M.A.: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28(1), 100–108 (1979)