Which Biomarkers Reveal Neonatal Sepsis?

Kun Wang1, Vineet Bhandari2, Sofya Chepustanova1, Greg Huber4, Stephen O’Hara1, Corey S. O’Hern5, Mark D. Shattuck6, Michael Kirby1
1 Department of Mathematics, Colorado State University, Fort Collins, CO, USA
2 Division of Perinatal Medicine, Department of Pediatrics, Yale University School of Medicine, New Haven, CT, USA
3 Department of Mathematics, Colorado State University, Fort Collins, CO, USA
4 Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA, USA
5 Department of Mechanical Engineering & Materials Science, Department of Applied Physics, and Department of Physics, Yale University, New Haven, CT, USA
6 Benjamin Levich Institute and Physics Department, The City College of New York, New York, NY, USA
∗ E-mail: kirby@math.colostate.edu

Abstract

We address the identification of optimal biomarkers for the rapid diagnosis of neonatal sepsis. We employ both canonical correlation analysis (CCA) and sparse support vector machine (SSVM) classifiers to select the best subset of biomarkers from a large hematological data set collected from infants with suspected sepsis from Yale-New Haven Hospital’s Neonatal Intensive Care Unit (NICU). CCA is used to select sets of biomarkers of increasing size that are most highly correlated with sepsis infection. The effectiveness of these biomarkers is then validated by constructing a sparse support vector machine diagnostic classifier. We find that the following set of five biomarkers capture the essential diagnostic information (in order of importance): bands, platelets, neutrophil CD64, white blood cells, and segs. Further, the diagnostic performance of the optimal set of biomarkers is significantly higher than that of isolated individual biomarkers. These results suggest an enhanced sepsis scoring system for neonatal sepsis that includes these five biomarkers.

Introduction

The identification and treatment of sepsis continues to be a major health issue. The incidence of sepsis is particularly high in the neonatal population, where low birth weight and other compromising factors make it a primary cause of morbidity and death [4,10,11]. Early identification and treatment are critically important to healthy patient outcomes
given the inconsistent presentation of sepsis in terms of body temperature, which may be either above or below normal [12,13,26].

The most reliable diagnostic of neonatal sepsis, often referred to as the gold standard, is a blood culture test for bacteria. While this test is the most reliable available, it can take 48 hours to obtain the results. As a result, treatment must often begin before the results are known. An additional complication is the fact that the blood culture test can be negative for one in five subjects with sepsis [1,10]. Thus, it is of critical importance to identify new biomarkers that will enable fast and reliable hematological scoring systems for sepsis in its earliest stages.

The current hematological scoring system was first proposed by Rodwell, et al. in 1988 and is based on the following seven quantities: total leukocyte (or white blood cell, WBC) count, mature neutrophil count (also named Segs, absolute neutrophil count, or ANC), immature neutrophil count (also named Bands, absolute band count, or ABC), ratio of immature to total neutrophil count (IT-ratio), ratio of immature to mature total neutrophil count, platelet count (Plt), and adverse changes in the total neutrophil count [27]. Another scoring system was proposed in Ref. [13] that characterizes a patient as septic if any two of the following four criteria are satisfied:

- ANC < 7500 or ANC > 14500/mm³
- ABC > 1500/mm³
- IT-ratio > 0.16
- Plt < 150,000/mm³.

These hematological scores are supplemented by other observational evidence and measurements collected by physicians including body temperature, blood pressure, and clinical presentation in determining the course of treatment before the blood culture results are available.

Additional diagnostic hematological biomarkers have been studied such as C-reactive protein [8,25] and procalcitonin [14,20]. While these biomarkers have shown to be correlated with sepsis, they are considered to have limited diagnostic information [17,26]. More recently, the blood biomarker neutrophil CD64 has proved to be particularly promising for early detection of sepsis [3,23,24]. Neutrophil surface CD64 expression is a high affinity Fc receptor (FcγRI) for immunoglobulin G (IgG) expressed on neutrophils (and other white blood cells). Quantities of CD64 increase markedly when neutrophils are activated by the human body’s response to infection, and in particular, to sepsis.

The challenge of biomarker identification is reflected by the fact that over 3000 sepsis biomarker studies have been published with almost 200 candidate biomarkers evaluated [26]. Nonetheless, clinicians are unsatisfied with the diagnostic tools currently available for making accurate and timely sepsis diagnosis that also support treatment therapies. The challenge is not to identify single biomarkers that pass a univariate test
for diagnostic efficacy, but to determine which sets of biomarkers, when considered as a group, yield the most accurate prognosticator.

In this investigation we integrate two tools for discovering information in large data sets. Embedded feature selection using a sparse support vector machine classifier [7, 18] and canonical correlation analysis [19], a tool for identifying relationships between two sets of variables. This two-pronged analysis provides a powerful general tool for the identification of biomarkers useful for multivariate scoring systems.

In this manuscript, we present a systematic study of the multivariate diagnostic capacity of a set of ten hematological biomarkers. Our goal is to establish a general approach that can be used effectively on potentially much larger sets of biomarkers. We develop an approach to identify a minimum set of predictive biomarkers with the ultimate goal of improving the early detection of sepsis. We verify the results by conducting an exhaustive evaluation of all possible combinations of biomarkers. We envision that the algorithms proposed here will be helpful tools as advances in biomedicine produce additional candidate biomarkers arising from new proteomic and metabolomic tests [22, 28].

**Results**

A total of 1383 sepsis evaluations were performed on 749 neonates during the study period. Blood cultures, complete blood counts (CBC), and neutrophil CD64 data were obtained for \( n_{all} = 674 \) of the sepsis evaluations.\(^1\) Evaluations were partitioned into three groups: (1) blood culture positive septic group \( (n_1 = 37) \), (2) clinically probable septic group \( (n_2 = 290) \), and (3) nonseptic group \( (n_3 = 347) \). In this study, we combined groups 1 and 2 and labeled these subjects as having sepsis. Our analysis is based on the comparison between this combined septic group \( (n_s = 327) \) and nonseptic group \( (n_n = 347) \). See Materials and Methods for details.

Data for ten hematological biomarkers were analyzed in this study including: (1) Age, (2) WBC count, (3) Hemoglobin count (Hgb), (4) Hematocrit percentage (Hct), (5) Platelets (Plt), (6) Segs, (7) Bands, (8) Lymphocyte (Lymph) count as a percentage of WBC, (9) Monocyte (Mono) count as a percentage of WBC, and (10) neutrophil CD64 expression. Following Ref. [12], P-values were computed for the biomarker data and all ten biomarkers were determined to have predictive capacity.

**Optimal subsets of biomarkers**

Multivariate correlation analysis is a general tool for exploring how variables are interrelated. Canonical correlation analysis provides a powerful tool for discovering relationships between two sets of variables. Given two sets of variables, CCA can identify subsets

\(^1\)One evaluation was excluded due to the erroneously high neutrophil CD64 value that skewed the results.
Table 1: By applying CCA for all possible $k$-combinations ($k = 1, \ldots, 10$), the subset of $k$ biomarkers with the highest correlation with the sepsis score is determined. The ‘Enter’ column indicates which biomarker is added to achieve the highest correlation at each $k$. The ‘Leave’ column indicates which biomarker is eliminated from the combination at that particular $k$. A biomarker will stay in the combination until it occurs in ‘Leave’ column. For instance, for the 5-combination, the most correlated biomarkers include Bands, Plt, CD64, Segs, and WBC. Hgb, which was present in the 4-combination, is replaced by Segs and WBC at level 5.

<table>
<thead>
<tr>
<th>$k$-combination</th>
<th>Correlation</th>
<th>Enter</th>
<th>Leave</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.563</td>
<td>Bands</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.615</td>
<td>Plt</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.633</td>
<td>Hgb</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.643</td>
<td>CD64</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.653</td>
<td>Segs, WBC</td>
<td>Hgb</td>
</tr>
<tr>
<td>6</td>
<td>0.660</td>
<td>Hgb</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.663</td>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.664</td>
<td>Lymph</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.666</td>
<td>Mono</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.668</td>
<td>Hct</td>
<td></td>
</tr>
</tbody>
</table>

of each set, which when combined as latent variables, produce the maximum correlation between the two sets. In this study, we choose one set of variables to be the sepsis score, and the second set is taken from all possible subsets of the ten biomarkers. CCA can thus generate an ordered list of biomarkers that are most correlated with the sepsis score. See Materials and Methods for details.

Here we discuss the results of applying CCA to select the best combinations of sepsis biomarkers. We first consider the single biomarker with highest correlation to the sepsis score. As shown in Table 1, this biomarker is bands. If we consider all pairs of biomarkers, bands and Plt possess the highest correlation with sepsis score. We note that CD64 has the second highest correlation with sepsis score, in the univariate sense, but improves the correlation of bands to sepsis score less than Plt, which has a lower univariate correlation with sepsis score. This is due to the fact that bands and CD64 are more correlated than bands and Plt, and so less information is provided by adding CD64. Hgb enters at $k = 3$ even though it has a very weak pairwise correlation with the sepsis score given it also has very weak pairwise correlation with Bands and Plt. The correlation saturates at $k = 5$ with the following combination set of biomarkers: Bands, CD64, Segs, WBC, and Plt. The rest of the biomarkers do not provide significant additional information about the sepsis score. The above analysis suggests that these five variables should be included in our sepsis scoring system. In the next section, we validate this result using a classifier to predict the sepsis score in terms of these biomarkers.
The diagnostic classifier

We seek to construct a decision function from the biomarker data that serves as a hematological scoring system, i.e. a function that maps a sample vector of biomarkers to a positive or negative sepsis diagnosis. Using the biomarkers identified by CCA above, WBC, Plt, Segs, bands, and CD64, we propose the linear decision function

\[ d(x) = w_1WBC + w_2Plt + w_3Segs + w_4Bands + w_5CD64 + b \]

From the sparse support vector machine approach described in Materials and Methods, we determined the optimal decision function to be

\[ \text{Score} = 0.37WBC - 0.88Plt - 0.7Segs + 2.7Bands + 0.45CD64 - 0.66. \] (1)

With this decision function, if the Score is greater than or equal to zero the diagnosis is positive for sepsis, whereas if the Score is less than zero, the diagnosis is healthy or aseptic disease. We note that since the range of values of the biomarkers varies widely, all values of the biomarkers are normalized by subtracting the mean over all cases and then dividing by the standard deviation.

The results of applying the classifier in Equation (1) to the full sepsis dataset are shown in Table 3. We calculated the true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), and accuracy (ACC) for these five biomarkers. We emphasize that there are two remaining questions of interest. How good is the classifier? Did we identify the most predictive biomarkers from the original set of ten? We focus on the validation of these biomarkers in the next section.
Biomarker validation

In this section we have two goals. First, we will verify that the number of biomarkers suggested by CCA, \( k = 5 \), is optimal. Secondly, we seek to provide evidence that the CCA-selected biomarkers are optimal. To do this, we will perform an exhaustive analysis of all possible scoring systems for the ten biomarkers. Clearly this approach is not feasible for large sets of biomarkers, but we exploit the fact that we only have ten to illustrate the power of CCA biomarker selection by constructing all possible SVM classifiers. We used the accuracy of the resulting decision functions for our validation.

Validation of the \( k = 5 \) classifier

For each \( k \), we select the \( k \)-combination set of biomarkers as identified by CCA and shown in Table 1. We construct a decision function for each \( k \) from 1 to 10 and evaluate several measures of the quality of the scoring system in Fig. 1. We find that each measure begins to saturate near \( k = 5 \), although one could argue that some slight improvement could be obtained by adding one or two more biomarkers for the given model. (We note that this particular model was not optimized over variations in the parameter \( b \).)

The receiver operating characteristic (ROC) curves for true positive versus false positive rate provide additional insight into the determination of the minimal number of biomarkers that provide predictive information about sepsis infection. In Figure 2, we show that the ROC curves become independent of \( k \) for \( k \geq 5 \), and thus \( k = 5 \) is indeed the appropriate number of biomarkers. In Figure 3, we show the ROC curve for \( k = 5 \) averaged over 100 SSVM models.

We provide further evidence that our biomarker selection was in fact the optimal one by applying SSVM to all possible combinations of biomarkers for each \( k \). We show the TPR for the top 20 of all possible combinations in Figure 4. It is clear that the CCA-selected biomarkers possess the largest TPR for each \( k \).

Materials and Methods

The data sets were obtained from a prospective study conducted in Neonatal Intensive Care Unit at Yale-New Haven Hospital [3]. Consecutive patients, who underwent a sepsis work-up as deemed necessary by the attending neonatologist during the time period 5/2005-7/2006, were enrolled in the study [3].

Sepsis Evaluations

The clinical and historical features used to identify patients at risk for sepsis include one or more of the following, as determined by the attending neonatologist [2,13,16]: (1) respira-

\(^2\)This study was approved by the Yale University School of Medicine Human Investigation Committee.
Figure 1: Prediction measures, true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), and accuracy (ACC), are shown for each k-combination of biomarkers selected by CCA.

...atory compromise (e.g. tachypnea, increase in frequency or severity of apnea, or increased ventilator support); (2) cardiovascular compromise (e.g. increased frequency or severity of bradycardic episodes, pallor, decreased perfusion, or hypotension); (3) metabolic changes (e.g. temperature instability, feeding intolerance, glucose instability, or metabolic acidosis); (4) neurological changes (e.g. lethargy, hypotonia, or irritability); and (5) antenatal risk factors (e.g. maternal Group B Streptococcus (GBS) colonization without adequate intrapartum prophylaxis, unknown maternal GBS status, maternal temperature, chorioamnionitis, preterm labor, or prolonged rupture of membranes). After the sepsis evaluation was performed, we utilized the following values derived from the complete blood count (CBC) to assign a sepsis score [3, 27]: (1) absolute neutrophil count (ANC) < 7500 or > 14500/mm³; (2) absolute band count (ABC) > 1500/mm³; (3) immature to total neutrophil ratio (IT-ratio) > 0.16; and (4) platelet (Plt) count < 150,000/mm³. Infants who met 2 or more of these laboratory criteria were categorized as having a positive sepsis score. Hemoglobin was measured in the clinical hematology laboratory using a calorimetric method. The hematocrit was calculated after measuring the total red blood cell count (RBC) and the mean corpuscular volume (MCV) of the RBCs. All blood cultures were collected using standard sterile techniques. As per unit protocol, we attempt to obtain 2 blood cultures with a minimum of 0.5 ml. The BACTEC (Becton Dickinson and Co., Sparks, MD) microbial detection system was used to detect positive blood cultures.
Figure 2: Receiver operating characteristic (ROC) curves of true positive rate versus false positive rate for optimal sets of $k$ biomarkers where $k = 1, \ldots, 10$ averaged over 100 SSVM models. The shaded region in the inset shows the standard deviation for $k = 5$.

Neutrophil CD64 expression was measured using 50 $\mu$l of whole blood incubated for 10 minutes at room temperature with a saturating amount of fluorescein isothiocyanate (FITC)-conjugated anti-CD64 monoclonal antibody or isotype control (Leuko64 kit, Trillium Diagnostics, Scarborough, ME), followed by ammonium chloride-based red cell lysis. Samples were washed once and re-suspended in 0.5 ml of phosphate-buffered saline with 0.1% bovine serum albumin. Flow cytometric analysis was accomplished using a Becton-Dickinson FACScan (Mountainview, CA) to collect log FITC fluorescence, log right-angle side scatter and forward scatter on a minimum of 50,000 leukocytes. Interassay standardization and neutrophil CD64 quantification were performed using FITC calibration beads (Leuko64 kit). Data analysis was performed using light scatter gating to define the neutrophil population, and neutrophil CD64 Index was quantified as mean equivalent soluble fluorescence units using QuickCal for Winlist (Verity Software House, Topsham, ME) with a correction for nonspecific antibody binding by subtracting values for the isotype control [3]. This was expressed as an absolute value. Investigators checking and confirming the neutrophil CD64 results were blinded to the clinical data, including the blood culture results. Clinicians did not have access to the neutrophil CD64 values and these were not used to decide initiation or duration of antibiotic therapy.
Sepsis Evaluations

Evaluations were obtained by accessing the electronic medical record from January 2008 through June 2009. Each evaluation typically included a complete blood count (CBC), two peripheral blood cultures, and other optional cultures. A patient could undergo multiple sepsis evaluations during admission. Since a single evaluation represented a separate episode of suspected sepsis and could be treated independently, we therefore treated all evaluations equivalently in this manuscript. Evaluations were excluded if the CBC, neutrophil CD64, or blood culture tests were not provided in the patient record. A total of 674 sepsis evaluations with complete hematologic, neutrophil CD64, and blood culture data were used for the analyses. (One evaluation, which was positive for *Candida albicans* was excluded due to the high neutrophil CD64 value that skewed the results.) Information about each sepsis evaluation included (1) sepsis diagnosis type, (2) day of life that the evaluation was performed, and (3) CBC data and neutrophil CD64 expression. Ten biomarkers were included in the analysis: Age, WBC, Hgb, Hct, Plt, Segs, Bands, Lymph, Mono, and CD64. Additional details about the laboratory and clinical data were recently

![ROC curve](image)

Figure 3: The ROC curve for the optimal 5 biomarker SSVM model. The shaded area around the curve represents the standard deviation of the ROC curve over 100 models.
Figure 4: The 20 highest TPR values when SSVM was applied for all possible combinations of \( k \) biomarkers (blue circles) from \( k = 1, \ldots, 10 \). The solid red circles are the TPR values for models built using the best \( k \) biomarkers selected by CCA.

published [29, 30].

**Defining Sepsis Outcome**

Individual sepsis evaluations with positive blood cultures were diagnosed as culture-proven sepsis according to the current National Healthcare Safety Network definitions for laboratory-confirmed bloodstream infections [15]. Individual sepsis evaluations with positive sepsis scores were categorized as clinical sepsis [3, 27]. This might include infants with other infectious diagnoses that were not accompanied by a positive blood culture, such as pneumonia, urinary tract infection, and necrotizing enterocolitis.

Three groups of evaluations were defined, and each evaluation was assigned to only one of the three groups. Group 1 consisted of 37 evaluations with a positive blood culture according to the current National Healthcare Safety Network definitions for laboratory confirmed bloodstream infections. Group 2 with “suspected sepsis” consisted of 290 evaluations, where the patients lacked a definitive positive blood culture, but the clinical
diagnosis was unable to rule out bacterial infection. Group 3 consisted of 347 evaluations, for which both the blood culture and clinical diagnosis showed no evidence of infection.

Data Pre-Processing
First, each evaluation \( i = 1, \ldots, n \), with data \( x_i \), is categorized as septic (groups 1 and 2 above) and nonseptic (group 3), where \( x_i \) is a real-valued vector with \( p = 10 \) components (biomarkers) and \( n = 674 \) is the total number of evaluations. For convenience, we labeled each evaluation using the variable \( y_i \), where \( y_i = +1 \) is the label for the septic group and \( y_i = -1 \) is given to each in the nonseptic group. To standardize the range of independent biomarkers, we normalized the real-valued data \( x \), a \( n \times p \) matrix, to have zero mean and unit standard deviation for each biomarker: [21]:

\[
x_{\text{norm}} = \frac{x - \bar{x}}{\sigma(x)},
\]

where \( x \) and \( x_{\text{norm}} \) are \( n \times p \) matrices, \( \bar{x} \) and \( \sigma(x) \) are the mean value and standard deviation of \( x \) for each biomarker.

Canonical Correlation Analysis
CCA is a multivariate statistical tool that facilitates the study of interrelationships among multiple variables [19,31]. A linear combination of variables can be chosen by CCA such that the correlation between two sets of data is maximized [5]. In our studies, the two sets of data are the sepsis score \( y \) and each distinct \( k \)-combination of the \( p = 10 \) biomarkers in the data matrix \( x \). In this investigation, CCA is used to identify the set of \( k \)-biomarkers most correlated as a group to the sepsis score.

To explore the redundancy among the biomarkers, we calculated the correlations between each possible \( k \)-combination of \( x \) and \( y \) using CCA. By varying \( k \) from 1 (single biomarker) to \( p \) (all biomarkers), we selected the specific set of biomarkers that possessed the highest correlation with \( y \) for each \( k \). The sets of biomarkers that had the largest correlation with the sepsis score and their corresponding correlation coefficients are shown in Table 1.

Sparse Support Vector Machines (SSVM)
We applied the SSVM ensemble method to build a classifier for each of the CCA-selected \( k \)-combination of biomarkers selected by CCA. A linear support vector machine (SVM) is a widely used classifier, which finds the hyperplane that separates high dimensional data with maximum margin by categories. The search of this hyperplane can be translated
into the following optimization problem:

\[
\text{Minimize } \|w\|_1 + C_+ \sum_{i:y_i=+1} \xi_i + C_- \sum_{j:y_j=-1} \xi_j
\]

subject to

\[
w^T x_i + b + \xi_i \geq 1, \quad y_i = +1, \\
w^T x_j + b - \xi_j \leq -1, \quad y_j = -1, \quad \text{and} \quad \xi \geq 0,
\]

where \(\|w\| = \sum_i |w_i|\) is the 1-norm of a vector, which induces the sparsity in the weight vector \(w\) [18]. We refer to the solution of Equation (3) as a sparse support vector machine (SSVM) following Ref. [7]. Note that splitting the classes in the objective function allows for unbalanced sample sizes.

Due to the limit size and noise of our data, a bootstrap aggregation method was applied to build an ensemble of SSVM classifiers using the following procedure [6, 9]:

1. The data set \(x(k)\) is randomly divided into a learning set \(L\) and a test set \(T\). \(T\) is one third of the data.

2. Based on the bootstrap aggregation method, a bootstrap training set \(L_B\) is randomly selected from the original learning set \(L\) with replacement. That is, a training set \(L_B\) consists of the same sample size as the original training set \(L\), but with several training samples appearing multiple times. Each bootstrap set \(L_B\) contains 63.2% unique samples of the original training set \(L\). By repeating this process 50 times, an ensemble of classifiers \(f_i(x)\), with \(i = 1, \ldots, 50\), is built by the SSVM. To have the same total cost for both false positives and false negatives, the parameters \(C_+\) and \(C_-\) of the SSVM are chosen according to

\[
\frac{C_+}{C_-} = \frac{\text{number of nonseptic training evaluations}}{\text{number of septic training evaluations}}
\]

with \(C_- = 1.0\) since the results are not sensitive to the overall scale of \(C_\pm\).

3. The final classification is obtained by calculating the mean of the ensemble of 50 classifiers.

4. The random division of the data into \(L\) and \(T\) is repeated 100 times, after which we calculate the mean and standard deviation. We used the same 100 random divisions of the training and test sets for each \(k\).

**Acknowledgments**

This material is based on work partially supported by DARPA (Space and Naval Warfare Systems Center Pacific) under Award No. N66001-11-1-4184. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of DARPA. Additional support for this work was provided by the Infectious Disease Supercluster, Colorado State University.
References


