A Computational Approach to Early Sepsis Detection

Jacob S. Calvert, BS; Daniel A. Price, MSc; Uli K. Chettipally, MD, MPH; Christopher W. Barton, MD; Mitchell D. Feldman, MD, MPhil; Jana L. Hoffman, PhD; Melissa Jay; Ritankar Das, MSc

aDascena Inc., Hayward, CA.
bKaiser Permanente South San Francisco Medical Center, South San Francisco, CA.
cDepartment of Emergency Medicine, University of California San Francisco, San Francisco, CA.
dDivision of General Internal Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA.

*Corresponding author: jana@dascena.com

Key Words: sepsis; severe sepsis; medical informatics; early diagnosis; computer-assisted diagnosis; clinical decision support systems

Financial Support: This material is based upon work supported by the National Science Foundation under Grant No. 1549867. The funder had no role in the conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, and approval of the manuscript; and decision to submit the manuscript for publication.
ABSTRACT

Objective: To develop high-performance early sepsis prediction technology for the general patient population.

Methods: Retrospective analysis of adult patients admitted to the intensive care unit (from the MIMIC II dataset) who were not septic at the time of admission.

Results: A sepsis early warning algorithm, InSight, was developed and applied to the prediction of sepsis up to 3 hours prior to a patient’s first five hour Systemic Inflammatory Response Syndrome (SIRS) episode. When applied to a never-before-seen set of test patients, InSight predictions demonstrated a sensitivity of 0.90 (95% CI: 0.89-0.91) and a specificity of 0.81 (95% CI: 0.80-0.82), exceeding or rivaling that of existing biomarker detection methods. Across predictive times up to three hours before a sustained SIRS event, InSight maintained an average area under the ROC curve of 0.83 (95% CI: 0.80-0.86). Analysis of patient sepsis risk showed that contributions from the coevolution of multiple risk factors were more important than the contributions from isolated individual risk factors when making predictions further in advance.

Conclusions: Sepsis can be predicted at least 3 hours in advance of onset of the first five hour SIRS episode, using only nine commonly available vital signs, with better performance than methods in standard practice today. High-order correlations of vital sign measurements are key to this prediction, which improves the likelihood of early identification of at-risk patients.
INTRODUCTION

Severe sepsis and septic shock are among the leading causes of death in the United States.\(^1\)\(^2\) Sepsis is most frequently caused by a systemic bacterial infection, but can also be caused by fungal, viral, and microbial endotoxin infections.\(^3\)\(^4\) A nonspecific indicator of risk for developing sepsis is Systemic Inflammatory Response Syndrome (SIRS).\(^5\) SIRS is defined as two or more of the following variables: temperature of more than 38°C or less than 36°C, heart rate of more than 90 beats per minute, respiratory rate of more than 20 breaths per minute or arterial carbon dioxide tension of less than 32 mm Hg, or abnormal white blood cell count (>12,000/µL or < 4,000/µL or >10% immature band forms).\(^5\) Sepsis is defined as SIRS with the addition of a known or suspected infection. Severe sepsis is sepsis with associated organ dysfunction, and septic shock additionally includes refractory hypotension.\(^5\)\(^6\) Approximately 750,000 patients are diagnosed with severe sepsis annually, and roughly one third of them die.\(^6\)\(^7\) The cost of treating sepsis is estimated to be $16.7 billion per year, making sepsis one of the most expensive conditions to diagnose and treat.\(^7\)\(^8\)

Despite this, sepsis detection methods have changed little since 1991 and include screening labs, which may be slow or inaccurate. Multiple studies have shown that early diagnosis and treatment, such

---

as Early Goal-Directed Therapy (EGDT), can reduce the risk of adverse patient outcome from severe sepsis and septic shock,\textsuperscript{9,10,11} though recent studies have questioned the effectiveness of existing treatment methods.\textsuperscript{12,13,14} Regardless, earlier and more accurate diagnosis of patients at high risk of developing severe sepsis or septic shock would provide a valuable window for identifying the most effective sepsis treatments or preventative measures. To fill the need for earlier and higher performance sepsis screening technology, we have developed a machine learning workflow for sepsis prediction, called InSight. InSight computes, in real-time, the risk that a patient will develop sepsis. The goal of InSight is to provide clinicians with accurate advance notice that a patient is ”trending septic”.

The increasing availability of Electronic Health Records (EHR) in clinical settings has inspired several attempts to identify patient conditions and trends through the automated analysis of medical records, with varying success. Alarm indicators for sepsis and septic shock have been shown to reduce mortality in hospital settings.\textsuperscript{15} Several systems have been validated against the detection of existing severe sepsis or septic shock, but lack predictive value.\textsuperscript{16,17,18} In this study, we assess the sensitivity

\begin{footnotesize}
\textsuperscript{11} Kumar, Anand et al. "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*." \textit{Critical care medicine} 34.6 (2006): 1589-1596.
\textsuperscript{15} Ferreras, JM et al. "[Implementation of an automatic alarms system for early detection of patients with severe sepsis]." \textit{Enfermedades infecciosas y microbiologia clinica} 33.8 (2015): 508-515.
\end{footnotesize}
and specificity of the InSight algorithm in the prediction of sepsis, three hours prior to an extended SIRS episode. This prediction is achieved through the analysis of correlations between nine common vital sign measurements.

**MATERIALS AND METHODS**

**Data Collection and Inclusion Criteria**

This is a retrospective study using the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II Clinical Database (Version 3).\(^{20}\) The MIMIC II database is composed of anonymized clinical documentation from approximately 32,000 patients at the Beth Israel Deaconess Medical Center (BIDMC) collected between 2001 and 2008. The BIDMC and the Massachusetts Institute of Technology Institutional Review Boards waived individual patient consent requirements, as the study did not affect clinical care and all data were anonymized.

Inclusion criteria for this study were (Figure 1):

I. Adult patient (i.e. age ≥ 18 years) admitted to the medical Intensive Care Unit (ICU).

II. Patient does not meet SIRS criteria at time of admission to the ICU or within first four hours of stay.

III. Documented measurements available for (i) systolic blood pressure, (ii) pulse pressure, (iii) heart rate, (iv) temperature, (v) respiration rate, (vi) white blood cell count, (vii) pH, (viii) blood oxygen saturation and (ix) age.\(^{21}\)

In order to analyze time series data more easily, beginning with ICU admission, the patient ICU stay was divided into one-hour intervals and measurement timestamps were rounded up to the nearest

---


hour. For intervals without observations for all nine measurements, missing values were taken to be the most recent available observation.

Gold Standard

After selection of the patients in the retrospective dataset for inclusion, each of the patients underwent a binary classification process to designate them as positive or negative for having acquired in-hospital sepsis. This classification was made based on the patient meeting both of the following criteria:

1. The patient record contains an ICD9 code (995.9) indicating in-hospital contraction of sepsis.
2. The patient meets the 1991 Systemic Inflammatory Response Syndrome (SIRS) criteria for sepsis for a persistent 5-hour period of time. The beginning of the patient's first 5-hour SIRS event is defined as the zero hour.

Training and Testing

1,394 patients satisfied inclusion criteria I-III, of which 159 (11.4%) also met gold standard criteria (1) and (2). The 1,394 patients were partitioned into mutually exclusive sets for training and testing the prediction algorithm. In order to ensure that training and testing set selections did not erroneously influence results, 4-fold cross validation was used. The 4-fold cross validation was done with a built-in MATLAB (MathWorks, Natick, MA, R2014a) function, which randomized the patients being placed in each group based on their anonymized medical record number (AMRN) provided in MIMIC II.

Analysis of Patient Time-Series Data

In order to capture trends in patient measurements and to emulate the analysis that would be performed for a prospective study, patient data were analyzed as a causal time-series. In particular, correlations between the following nine measurements (labeled as \( i \) below) – systolic blood pressure, pulse pressure, heart rate, temperature, respiration rate, white blood cell count, pH, blood oxygen
saturation, and age – were classified within a sliding, 5-hour observation window. These nine measurements were selected for their standard availability, medical relevance to sepsis, and the reliable likelihood of their frequent determination in a clinical setting.

I. At time of admission ($T_0$), age was recorded.

II. At the $N$th hour after admission, each measurement (labeled as $i$) was averaged over the time window $[T_{N-5}, T_N]$. These averages were assigned to $M_i$ and the changes in each measurement $i$ between $T_{N-5}$ and $T_N$ were assigned to $D_i$. Because we used 9 measurements, there were 9 $M_i$ and 9 $D_i$.

III. To classify measurement $i$ as increasing, roughly constant, or decreasing, $D_i$ was classified as positive, negligible, or negative. This classification, called $\hat{D}_i$, was made according to thresholds of the median of $|D_i|$.

IV. To classify trends among pairs of measurements, an indicator of positive, negative, or negligible correlation between pairs of measurements $\{i,j\}$, was stored in $\hat{D}_{i,j}$. Similarly, correlations between triplets of measurements $\{i,j,k\}$ were assigned to $\hat{D}_{i,j,k}$.

Doublet and triplet trend classifications provide information on the coupling of organ systems. The human body maintains healthy vital physiologies through complex, inhibitory feedback mechanisms. Serious illnesses like sepsis can initiate combative feedback cycles, exhausting the body's reserve capacity to maintain homeostasis. This development can be observed in the tightly coupled patterns of organ systems, and the coupled measurement classifications are designed to illuminate these patterns.\(^5\)

**Assigning InSight Scores for the Prediction of Sepsis Onset**
Each patient’s measurement and trend information stored in $M_i$, $D_i$, $D_{ij}$, and $D_{ijk}$, was translated into a dimensionless score, according to Eq. 1.

$$S_{\text{score}} = a \sum_{i \in \mathcal{M}} p(M_i) + b \sum_{i \in \mathcal{D}} p(D_i) + c \sum_{(i,j) \in \mathcal{L}} p(D_{ij}) + d \sum_{(i,j,k) \in \mathcal{L}} p(D_{ijk})$$  \hspace{1cm} (1)

Eq. 1 is somewhat reminiscent of a Modified Early Warning Score (MEWS) calculation\textsuperscript{22} -- summing numbers from reference tables, based on the range into which a measurement falls. Here, however, the summed numbers resemble probabilities and the equation incorporates measurement trends (the singlet $D_i$, doublet $D_{ij}$, and triplet $D_{ijk}$ trend terms). The functions, $p(\cdot)$, are related to the probability of a particular measurement or combination of measurements leading to sepsis development, and thereby combine diverse measurements and indicators into a single score. Here, $A-D$ allow the sums to be written compactly by representing the sets of all statistically significant indicators in the training library for each analysis type ($M_i$, $D_i$, $D_{ij}$, and $D_{ijk}$). The calibration constants $a-d$ were chosen to maximize the area under the training set receiver operator characteristic (ROC) curve (AUROC), using a standard optimization technique. The score calibration and assignment were handled with custom scripts written in MATLAB. We have previously utilized similar techniques in the application of a related algorithm for the prediction of patient stability.\textsuperscript{23} For this study, calibration constant values typically fell within [0,2] and scores ranged within [-1.18,3.50].

RESULTS

\textsuperscript{22} Subbe, CP et al. "Validation of a modified Early Warning Score in medical admissions." \textit{Qjm} 94.10 (2001): 521-526.

InSight was used to predict which patients would develop sepsis 3 hours before the zero hour; zero hour was defined by the patient’s first sustained SIRS episode of at least five hours (the Gold Standard criterion 2). Sepsis risk scores ranged from -1.18 to 3.50, with an average of -0.211 (95% CI: -0.25 to -0.17).

InSight demonstrates an AUROC of 0.92 (95% CI: 0.86 - 0.93) at three hours before a sustained SIRS episode, the zero hour (Figure 2). We compared this performance against the documented performance of procalcitonin (PCT), a common sepsis biomarker test, derived from a review and meta-analysis of 31 studies in the literature. The AUROC of the averaged PCT blood tests is 0.85. We have assumed here that the PCT blood test is ordered at the first sign of SIRS and that laboratory results are returned and analyzed in three hours. This is a conservative estimate, assuming a rapid laboratory turnaround time. By designating a score of 0.30 as the cutoff (i.e. scores higher than 0.30 indicate a prediction of sepsis), InSight achieved a sensitivity of 90% and a specificity of 81%, compared with only 63% sensitivity at a comparable 80% specificity for the PCT assay.

Along with PCT ≥ 2.05 ng/mL, we further compared InSight sensitivity and specificity against results from the literature for lactate ≥ 2 mmol/L, a common sepsis stratification biomarker, and the Systemic Inflammatory Response Syndrome (SIRS) criteria (Figure 3). InSight rivaled the sensitivity of the SIRS criteria (90% vs. 93%) and the specificity of the referenced lactate assay (81% vs. 82%).

However, the SIRS criteria alone have poor specificity, resulting in a 90% false positive rate, and the lactate assay is reported to have a low sensitivity of only 34%.  

The *InSight* results were robust under several random, mutually exclusive training and testing set selections. We summarize the *InSight* patient classification 3 hours before zero hour in a confusion matrix, for one quarter of the patient population used as a test run (Table 1). Here, $\hat{Y}$ indicates the number of patients predicted to become septic, while $\bar{Y}$ denotes the set of patients satisfying the gold standard criteria for sepsis. For example, the top-left table entry $(\hat{Y}, \bar{Y})$ lists the number of true positives. *InSight* had an overall accuracy of 82.7% (95% CI: 78.2-86.4).

<table>
<thead>
<tr>
<th></th>
<th>$Y$</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Y}$</td>
<td>36</td>
<td>56</td>
</tr>
<tr>
<td>$\bar{Y}$</td>
<td>4</td>
<td>252</td>
</tr>
</tbody>
</table>

Table 1: True positives, false positives, true negatives and false negatives for one four-fold cross validation test.

This process was repeated for 0-, 1-, and 2- hours before zero hour, each time with four fold cross validation, as outlined above. Averaging across all hours before zero hour (not including the zero hour results themselves), the AUROC was 0.83 (95% CI: 0.80-0.86). For each time-before-zero, *InSight* sensitivity and specificity were calculated for a variety of optimized calibration constants resulting from different training and test set partitions. *InSight* performance quality was maintained across all tested hours preceding zero hour.
To identify the key parameters driving accurate sepsis prediction before onset, the InSight patient scores were broken down into components from single vital sign $M_i$ terms and trend $\tilde{D}_i$, $\tilde{D}_{ij}$, and $\tilde{D}_{ijk}$ terms, and the relative contribution of each component was analyzed. On average, single vital sign $M_i$ and triplet trend $\tilde{D}_{ijk}$ terms contributed a combined 78% of each patient score. For predictions made 3 hours before the zero hour, InSight scores were almost entirely determined by $M_i$ and $\tilde{D}_{ijk}$ components (97% of each score), with 63% from $\tilde{D}_{ijk}$ alone. Conversely, $\tilde{D}_{ijk}$ components played a minor role in determining patient score closer to onset, which is made clear through the comparison of $\tilde{D}_{ijk}$ contribution histograms (Figure 4).

DISCUSSION

We have developed an algorithm, InSight, which predicts patient sepsis development three hours prior to a sustained SIRS episode, and which demonstrates a sensitivity of 90% at a specificity of 81%. This prediction is calculated using only nine very common clinical measurements, and outperforms the existing biomarker detection methods. While the SIRS criteria are sensitive to sepsis, SIRS suffers from a high false positive rate. Contrastingly, lactate assays are specific, but often miss septic patients. The lack of adequate prediction for patients at risk of septic shock in the ICU prevents early intervention and inhibits the ability to study and develop effective treatment methods for patients prior to organ dysfunction or hypotension. InSight’s high sensitivity and specificity, along with the ability to provide 3-hour advance notice before a sustained SIRS event, provides early and accurate identification of patients at risk for sepsis development and would be a useful clinical tool for the early prediction of sepsis evolution. Early identification of sepsis risk would allow clinicians to implement supportive treatments,

---

determine appropriate antibiotic administration, and potentially reduce associated complications and extended patient hospitalizations.

*InSight*’s key feature is the ability to combine diverse measurements and find correlations of these aggregate measurements with patient outcomes of interest. This feature is critical for sepsis prediction, as multi-organ diseases evolve in complicated ways that elude existing diagnostic methods. The correlations of measurements, as well as their trends over time, could provide valuable information about current homeostatic conditions. This is demonstrated here through the analysis of triplet trend $\mathcal{D}_{i;ik}$ contributions to patient scores. In particular, the dependence of $\mathcal{D}_{i;ik}$ score fraction on time suggests that higher order correlations played an essential role in early sepsis detection, but mattered less closer to zero hour. Existing tools, which consider only low-dimensional correlations, have yet to tap into this wealth of higher-order predictive power.

Surprisingly, first- and second-order trend indicators ($\mathcal{D}_i$ and $\mathcal{D}_{i;i}$) mattered little in scoring patient sepsis risk. In fact, those two indicators account for less than 10% of patient scores, 2 and 3 hours before zero hour. This may be because the triplet trend $\mathcal{D}_{i;ik}$ indicator is able to quantify the majority of patient risk for 2- and 3-hour predictions. Further, it is possible that, as the time before zero hour decreases, $\mathcal{D}_{i;ik}$-type trends are replaced by more apparent septic trends in patient measurements, which can be captured by single vital sign $\mathbf{M}_i$ analysis. These lower-order trends could account for the high sensitivity of SIRS. However, variations in single vital sign $\mathbf{M}_i$-type trends may readily overlap with different adverse medical events than sepsis, leading to increased false positive rates.

Because this is a retrospective study on data available through MIMIC II (ICU records from 2001 to 2008), the clinical measurements are necessarily limited. Requiring complete records for all possible measurements would diminish the available patient pool to statistically irrelevant levels. However, within the bounds of those limitations, we have utilized nine very commonly available, clinically relevant measurements to predict patient sepsis development with high accuracy, three hours prior to an extended
SIRS episode. The power of this algorithm derives from analyzing the higher order correlations and trends between sets of these common clinical measurements.

In the future, we anticipate implementing InSight prospectively, ideally in an Emergency Department setting, to aid in the early identification of patients at risk of sepsis. InSight is designed to be integrated into a hospital’s existing EHR, and trained on the data set and patient population available at the site of implementation. Different clinical variables may be available in a prospective study because of geographic variations in clinical policy and practice and advances in medical technology, which will affect the frequency and types of measurements recorded.\textsuperscript{31} InSight’s performance and the relative contributions of the M- and D-type indicators were robust to changes in the training and testing sets, which suggests that these results will generalize to other populations and data sets. However, a period of training and statistical analysis to determine which available clinical variables are most relevant in a new setting will benefit the predictive power of the algorithm. Further, redundant factors can be eliminated through the use of LASSO to prevent overfitting to irrelevant measurements.

CONCLUSION

We have described a novel machine learning technique for early sepsis detection in the Intensive Care Unit. Vital signs, lab tests, patient demographics, and their changes over time, were processed into dimensionless indicators. These indicators were then aggregated into higher-dimensional classes of measurement behavior ($\mathbf{M}_i$, $\mathbf{D}_i$, $\mathbf{D}_{ij}$, and $\mathbf{D}_{ijk}$), which were combined according to an equation learned from patient data. The results demonstrated that InSight is capable of sepsis prediction up to 3 hours prior to the zero hour, with sensitivity and specificity that rival or exceed the individual strengths of existing clinical detection tools.

ACKNOWLEDGEMENTS

The authors would like to thank Drew Birrenkott (Dascena Inc.) for helpful discussions.

FIGURE CAPTIONS

Figure 1. Patient inclusion flow chart.

Figure 2. Receiver operating characteristic (ROC) comparison of InSight and procalcitonin results averaged from 31 studies in the literature. InSight provides highly sensitive and specific sepsis predictions an estimated six hours prior to procalcitonin, the common sepsis biomarker test. "t=0" represents the onset of the patient’s first five hour SIRS episode.

Figure 3. Sensitivity and specificity comparison between InSight and standard sepsis diagnostic methods. SIRS is the Systemic Inflammatory Response Syndrome. Common sepsis biomarkers lactate (≥ 2 mmol/L), and PCT (procalcitonin, ≥ 2.05 ng/mL) sensitivity and specificity are assessed from previously published studies in the literature. The error bars on InSight represent a 95% confidence interval.

Figure 4. Histogram sequence of third-order $D_{t_{ij}}$ trend scores. As sepsis onset approaches, high order trends become less important, with most third-order trend score contributions collapsing onto zero, one hour before the patient’s first five hour SIRS episode.
MIMIC II patients

...in the MICU

...for patients age $\geq 18$

...not meeting SIRS and with an observation for each measurement

To classifier
Figure 4