

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email editorial.bmjopen@bmj.com

BMJ Open

Validation of a machine learning algorithm for the prediction and detection of sepsis using only vital sign data

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2017-017833
Article Type:	Research
Date Submitted by the Author:	18-May-2017
Complete List of Authors:	Mao, Qingqing; Dascena Jay, Melissa; Dascena hoffman, jana; Dascena, Calvert, Jake; Dascena Barton, Christopher; University of California San Francisco, Emergency Medicine Shimabukuro, David; University of California San Francisco, Anesthesia and Perioperative Care Shieh, Lisa; Stanford University School of Medicine, Medicine Chettipally, Uli ; University of California San Francisco, Emergency Medicine; Kaiser Permanente South San Francisco Medical Center Fletcher, Grant; University of Washington School of Medicine Kerem, Yaniv; Stanford University School of Medicine Zhou, Yifan; University of California Berkeley, Statistics Das, Ritankar; Dascena
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Diagnostics, Infectious diseases, Intensive care, Emergency medicine
Keywords:	sepsis, septic shock, clinical decision support, prediction, machine learning, electronic health records

SCHOLARONE™
Manuscripts

Validation of a machine learning algorithm for the prediction and detection of sepsis using only vital sign data

Qingqing Mao¹, Melissa Jay¹, Jana L. Hoffman^{1*}, Jacob Calvert¹, Christopher Barton², David Shimabukuro³, Lisa Shieh⁴, Uli Chettipally^{2,5}, Grant Fletcher⁶, Yaniv Kerem^{7,8}, Yifan Zhou⁹
Ritankar Das¹

¹ Dascena Inc., Hayward, CA

² Department of Emergency Medicine, University of California San Francisco, San Francisco, CA, United States

³ Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, CA, United States

⁴ Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

⁵ Kaiser Permanente South San Francisco Medical Center, South San Francisco, CA, United States

⁶ Division of Internal Medicine, University of Washington School of Medicine, Seattle, WA, United States

⁷ Department of Clinical Informatics, Stanford University School of Medicine, Stanford, CA, United States

⁸ Department of Emergency Medicine, Kaiser Permanente Redwood City Medical Center, Redwood City, CA, United States

⁹ Department of Statistics, University of California Berkeley, Berkeley, CA, United States

* Corresponding author

Email: Jana@Dascena.com
22710 Foothill Blvd., Suite #2
Hayward, CA 94541

Abstract

Objectives: We validate a machine learning-based sepsis prediction algorithm (*InSight*) for the detection and prediction of three sepsis-related gold standards, using only six common vital signs, and compare these results with other sepsis scoring systems commonly in clinical use. We also evaluate *InSight*'s robustness to missing data, and assess customization of the algorithm to site-specific data using transfer learning.

Design: We used a machine learning algorithm with gradient tree boosting, relying solely on data from six vital signs to train *InSight*. Relevant features for prediction were created from combinations of vital sign measurements and their changes over time.

Setting: A mixed-ward (emergency and inpatient) retrospective data set from the University of California, San Francisco (UCSF) Medical Center.

Participants: 90,353 adult emergency and inpatient encounters from June 2011 to March 2016.

Interventions: none

Primary and secondary outcome measures: Area under the receiver operating characteristic curve (AUROC) for detection and prediction of sepsis, severe sepsis, and septic shock.

Results: In the detection of sepsis and severe sepsis, *InSight* achieves an area under the receiver operating characteristic (AUROC) curve of 0.95 (95% CI 0.93 - 0.97) and 0.90 (95% CI 0.88 - 0.92), respectively. Four hours prior to onset, *InSight* predicts septic shock with an AUROC of 0.96 (95% CI 0.94 - 0.98), and severe sepsis onset with an AUROC of 0.85 (95% CI 0.79 - 0.91).

Conclusions: *InSight* outperforms existing sepsis scoring systems in both identification and prediction of sepsis, severe sepsis, and septic shock. This is the first sepsis screening system to exceed an AUROC of 0.90 using only vital sign inputs. *InSight* is robust to significant amounts of missing data, and can be customized to a novel hospital data set using a small fraction of site data.

Strengths and limitations of this study

- Machine learning is applied to the detection and prediction of three separate sepsis standards.
- Only six commonly measured vital signs are used as input for the algorithm.
- The algorithm is robust to randomly missing data.

- Transfer learning successfully leverages large dataset information to a target dataset.
- Retrospective nature of the study does not predict clinician reaction to information.

Introduction

Sepsis is a major health crisis and one of the leading causes of death in the United States [1]. Approximately 750,000 hospitalized patients are diagnosed with severe sepsis in the United States annually, with an estimated mortality rate of up to one-third [2,3]. The cost burden of sepsis is disproportionately high, with estimated costs of \$20.3 billion dollars annually, or \$55.6 million per day in US hospitals [4]. Additionally, the average hospital stay for sepsis is twice as expensive as other conditions [5], and the average incidence of severe sepsis is increasing by approximately 13% per year [6]. Early diagnosis and treatment have been shown to reduce mortality and associated costs [7-9]. Despite clear benefits, early and accurate sepsis detection remains a difficult clinical problem.

Sepsis has been defined as a dysregulated host response to infection. In practice, sepsis can be challenging to recognize because of the heterogeneity of the host response to infection, and the diversity of possible infectious insult. Sepsis has been traditionally recognized as two or more Systemic Inflammatory Response Syndrome (SIRS) [10] criteria together with a known or suspected infection; progressing to severe sepsis, in the event of organ dysfunction; and finally to septic shock, which additionally includes refractory hypotension [10]. However, ongoing debates over sepsis definitions and clinical criteria, as evidenced by the recent proposed redefinitions of sepsis [11], underscore a fundamental difficulty in the identification and accurate diagnosis of sepsis.

Various rule-based disease severity scoring systems are widely used in hospitals in an attempt to identify septic patients. These scores, such as the Modified Early Warning Score (MEWS) [12], the Systemic Inflammatory Response Syndrome (SIRS) criteria [13], and the Sequential Organ Failure Assessment (SOFA) [14], are manually tabulated at the bedside and lack accuracy in sepsis diagnosis. However, the increasing prevalence of Electronic Health Records (EHR) in clinical settings provides an opportunity for enhanced patient monitoring and increased early detection of sepsis.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

This study validates a machine learning algorithm *InSight*, which uses only six vital signs taken directly from the EHR, in the detection and prediction of sepsis, severe sepsis, and septic shock in a mixed-ward population at the University of California, San Francisco (UCSF). We investigate the effects of induced data sparsity on *InSight* performance, and compare all results with other scores that are commonly used in the clinical setting for the detection and prediction of sepsis. Furthermore, we apply a transfer learning scheme to customize a Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III-trained algorithm to the UCSF patient population using a minimal amount of UCSF-specific data.

Methods

Data sets

We used a data set from the UCSF Medical Center representing patient stays from June 2011 to March 2016 in all experiments. The UCSF data set contains 17,467,987 hospital encounters, including inpatient and outpatient visits to all units within the UCSF medical system. The data were de-identified to comply with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. For transfer learning, we used the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III v1.3 data set, compiled from the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA between 2001 and 2012, composed of 61,532 ICU stays [15]. This database was constructed by researchers at MIT's Laboratory for Computational Physiology, and the data were also de-identified in compliance with HIPAA. Data collection for the MIMIC-III and UCSF datasets did not impact patient safety. Therefore, this study constitutes non-human subjects research, which does not require Institutional Review Board approval.

Data Extraction and Imputation

The data were provided in the form of comma separated value (CSV) files and stored in a PostgreSQL [16] database. Custom SQL queries were written to extract measurements and patient outcomes of interest. The measurement files were then binned by hour for each patient. To be included, patients were required to have at least one of each type of measurement recorded

1
2
3 during the encounter. If a patient did not have a measurement in a given hour, the missing
4 measurement was filled in using carry-forward imputation. This imputation method applied the
5 patient's last measured value to the following hour. In the case of multiple measurements within
6 an hour, the mean was calculated and used in place of an individual measurement. After the data
7 were processed and imputed in Python [17], they were used to train the *InSight* classifier and test
8 its predictions at sepsis onset and at fixed time points prior to onset.
9
10
11
12
13

14 15 16 **Gold Standards**

17
18 In this study, we tested *InSight*'s performance according to various gold standards
19 (clinical indications). We investigated *InSight*'s ability to predict and detect sepsis, severe sepsis,
20 and septic shock. Further, we compared *InSight*'s performance to SIRS, MEWS, and SOFA, for
21 each of the following gold standards. For training and testing the algorithm, we conservatively
22 identified each septic condition by requiring that the ICD-9 code corresponding to the diagnosis
23 was coded for each positive case, in addition to meeting the clinical requirements for the
24 definition of each septic standard as defined below.
25
26
27
28
29
30
31
32

33 **Sepsis**

34 The sepsis gold standard was determined using the 2001 consensus sepsis definition [10]:
35 “the presence of two or more SIRS criteria paired with a suspicion of infection.” To identify a
36 case as positive for sepsis, we required ICD-9 code 995.91. The onset time was defined as the
37 first time two or more SIRS criteria were met within the same hour. SIRS criteria are defined as:
38
39
40

- 41 ● heart rate > 90 beats/ min,
 - 42 ● body temperature > 38 °C or < 36 °C,
 - 43 ● respiratory rate >20 breaths/min or PaCO₂ < 32 mmHg, and
 - 44 ● white blood cell count > 12,000 cells/μL or < 4,000 cells/μL. [10]
- 45
46
47
48
49

50 **Severe Sepsis**

51 The severe sepsis gold standard used the definition of severe sepsis as “organ dysfunction
52 caused by sepsis” which can be represented by one or more of the criteria below, and identified
53
54
55
56
57
58
59
60

1
2
3 for patients with the severe sepsis ICD-9 code 995.92. We assigned the severe sepsis onset time
4 to be the first instance during which one of the following organ dysfunction criteria were met.
5
6

- 7 ● Lactate > 2 mmol/L
- 8 ● Systolic blood pressure < 90 mmHg
- 9 ● Urine output < 0.5 mL/kg, over two hours, prior to organ dysfunction after fluid
10 resuscitation
- 11 ● Creatinine > 2 mg/dL without renal insufficiency or chronic dialysis
- 12 ● Bilirubin > 2 mg/dL without having liver disease or cirrhosis
- 13 ● Platelet count < 100,000 μ L
- 14 ● International normalized ratio > 1.5
- 15 ● PaO₂/FiO₂ < 200 in addition to pneumonia
16 < 250 with acute kidney injury but without pneumonia
17
18
19
20
21
22
23
24
25

26 **Septic Shock**

27
28 We identified as positive case septic shock those patients who received the septic shock ICD-9
29 code 785.52 and additionally demonstrated the following conditions:
30
31

- 32 ● systolic blood pressure of < 90 mmHg, defined as hypotension, for at least 30 minutes,
33 and
- 34 ● who were resuscitated with ≥ 20 ml/kg over a 24 hour period, or
- 35 ● who received ≥ 1200 ml in total fluids. [18]
36
37
38

39 The onset time was defined as the first hour when either the hypotension or fluid resuscitation
40 criterion was met.
41
42
43
44

45 **Calculating Comparators**

46
47 We compared *InSight* predictions for each gold standard to three common patient deterioration
48 scoring systems: SIRS, SOFA, and MEWS. The SIRS criteria, as explained in the sepsis
49 definition, were evaluated independently of the suspicion of infection. To calculate the SOFA
50 score, we collected each patient's PaO₂/FiO₂, Glasgow Coma Score, mean arterial blood pressure
51 or administration of vasopressors, bilirubin level, platelet counts, and creatinine level. Each of
52 the listed measurements is associated with a SOFA score of 1-4, based on severity level, as
53
54
55
56
57
58
59
60

1
2
3 described by Vincent et al. [14]. After receiving a score for each of the six organ dysfunction
4 categories, the overall SOFA score was computed as the sum of the category scores and used as a
5 comparator to *InSight*. Finally, the MEWS score, which ranges from 0 (normal) to 14 (high risk
6 of deterioration), was determined by tabulating subscores for heart rate, systolic blood pressure,
7 respiratory rate, temperature, and Glasgow Coma Score. We used the subscore system
8 presented in Fullerton et al. [19] to compute each patient's MEWS score.
9
10
11
12
13

14 15 16 **Measurements and Patient Inclusion**

17
18 In order to generate *InSight* scores, patient data were analyzed from each of the following
19 six clinical vital sign measurements: systolic blood pressure, diastolic blood pressure, heart rate,
20 respiratory rate, peripheral capillary oxygen saturation (SpO₂), and temperature. We used only
21 vital signs, which are frequently available and routinely taken in the ICU, ED, and floor units.
22 Patient data were used from the course of a patient's hospital encounter, regardless of the unit the
23 patient was in when the data were collected.
24
25
26
27
28

29 All patients over the age of 18 were considered for this study. For a given encounter, if
30 the patient was admitted to the hospital from the ED, the start of the ED visit is where the
31 analysis began. Patients in our final data sets were required to have at least one measurement for
32 each of the six vital signs. We further limited the study group to exclude patients whose septic
33 condition onset time was within seven hours after the start of their record, which was either the
34 time of admission to the hospital or the start of their ED visit; the latter was applicable only if the
35 patient was admitted through the ED. This window of sepsis onset time was selected to enable
36 sepsis prediction up to four hours prior to onset. Patients with sepsis onset after 2,000 hours post-
37 admission were also excluded, to limit the data analysis matrix size. The final UCSF data set
38 included 90,353 patients (Fig.1) and the MIMIC-III data set contained 21,604 patients, following
39 the same inclusion criteria.
40
41
42
43
44
45
46
47

48 After patient exclusion, our final group of UCSF patients was composed of 55% women
49 and 45% men with a median age of 55. The median hospital length of stay was 4 days, IQR =
50 (2,6). Of the 90,353 patients, 1,179 were found to have sepsis (1.30%), 349 were identified as
51 having severe sepsis without shock (0.39%), and 614 were determined to have septic shock
52 (0.68%). The in-hospital mortality rate was 1.42%. Patient encounters spanned a variety of
53 wards. The most common units represented in our study were perioperative care, the emergency
54
55
56
57
58
59
60

department, the neurosciences department, and cardiovascular and thoracic transitional care. In the MIMIC-III data set, approximately 44% of patients were women and 56% were men. Stays were typically shorter in this data set, since each encounter included only an ICU stay. The median length of stay was 2 days. Furthermore, due to the nature of intensive care, there was a higher prevalence of sepsis (1.91%), severe sepsis (2.82%), and septic shock (4.36%). A full summary of baseline characteristics for both data sets is presented in Table 1.

Table 1: Demographic and clinical characteristics for UCSF patient population analyzed (N=90,353) and MIMIC-III patient population analyzed (N=21,604).

Demographic Overview	Characteristic	UCSF		MIMIC-III	
		Count	Percentage	Count	Percentage
Gender	Female	49,763	55.08%	9,499	43.97%
	Male	40,590	44.92%	12,105	56.03%
Age UCSF: median 55, IQR (38-67) MIMIC-III: median 65, IQR (53-77)	18-29	10,652	11.79%	978	4.53%
	30-39	14,202	15.72%	1,114	5.16%
	40-49	11,888	13.16%	2,112	9.78%
	50-59	16,856	18.66%	3,880	17.96%
	60-69	19,056	21.09%	4,906	22.71%
	70+	17,699	19.59%	8,614	39.87%
Length of Stay (days) UCSF: median 4, IQR (2-6)	0-2	28,258	31.26%	11,054	51.17%
	3-5	35,128	38.88%	7,004	32.42%
	6-8	12,664	14.02%	1,673	7.74%
	9-11	4,934	5.46%	734	3.40%

MIMIC-III: median 2, IQR (2-4)	12+	9,369	10.37%	1,139	5.27%
Death During Hospital Stay	Yes	1,279	1.42%	1,328	6.15%
	No	89,074	98.58%	20,276	93.85%
ICD-9 Code	Sepsis	1,179	1.30%	413	1.91%
	Severe Sepsis	349	0.39%	609	2.82%
	Septic Shock	614	0.68%	943	4.36%

Machine Learning

We used gradient tree boosting to construct our classifier. Gradient tree boosting is an ensemble technique which combines the results from multiple weak decision trees in an iterative fashion. We created features from combinations of the six vital sign measurements and the changes in these measurements over time. Each decision tree was built by discretizing features into two categories. For example, one node of the decision tree might have stratified a patient based on whether their respiratory rate was greater than 20 breaths per minute, or not. Depending on the answer for a given patient, a second, third, etc., vital sign may be checked. A risk score was generated for the patient based on their path along the decision tree. We limited each tree to split no more than six times; no more than 1000 trees were aggregated in the iteration through gradient boosting to generate a robust risk score.

We performed four-fold cross validation to validate *InSight's* performance and minimize potential model overfit. We randomly split the UCSF data set into a training set, comprised of 80% of UCSF's encounters, and an independent test set with the remaining 20% of encounters. Of the training set, data were divided into four groups, three of which were used to train *InSight*, and one of which was used to test. After cycling through all combinations of train and test set, we then tested each of the four models on the independent test set. Mean performance metrics were calculated based on these four models.

Missing Data

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

After assessing *InSight*'s performance on complete data sets, we used a random deletion process to simulate the algorithm's robustness to missing measurements. Individual measurements from the test set were deleted according to a probability of deletion, P . We set $P = \{0, 0.1, 0.2, 0.4, \text{ and } 0.6\}$ for each of our missing data experiments and tested the *InSight* algorithm on the sparse data sets.

Transfer Learning

To evaluate *InSight*'s performance on a minimal amount of UCSF data, we used a transfer learning approach [20]. There are clear dissimilarities in patient demographics, clinical characteristics, and average measurement frequencies between the UCSF and MIMIC-III data sets (see Table 1). Partially this is because the UCSF data involves a variety of hospital wards, whereas the MIMIC-III data set provides only measurements taken in the ICU. We sought to determine improved performance metrics on the UCSF target data set, when the algorithm is primarily trained on MIMIC-III. Using MIMIC-III data as the source, and UCSF as the target, we trained the *InSight* classifier according to the severe sepsis gold standard. Variable amounts of UCSF training data were incrementally added to the MIMIC-III training data set, and the resulting model was then validated on the separate UCSF test data set. Specifically, we left 50% of the UCSF patients as test data, and we randomly selected different fractions of the remaining UCSF data and combined them with the entire MIMIC-III data set as the training data. For each fraction used, we adjusted the relative weighting between the UCSF data and the MIMIC-III data to determine the best performance. Ten-fold cross validations were performed to determine the best weights and to calculate the uncertainties.

Results

InSight's performance with respect to MEWS, SOFA, and SIRS is summarized in Figures 2A-C. Figures 2A, 2B, and 2C demonstrate *InSight*'s ability to accurately detect the onset of sepsis and severe sepsis, and to accurately predict septic shock four hours prior to onset, compared to the performance of common sepsis scoring systems. Each figure presents *InSight*'s receiver operating characteristic (ROC) curve together with the ROC curves for MEWS, SOFA, and SIRS. *InSight* achieves an area under the receiver operating characteristic (AUROC) curve

for sepsis onset of 0.95 (95% confidence interval (CI), 0.93 - 0.97), for severe sepsis onset 0.90 (95% CI 0.88 - 0.92), and for septic shock 0.99 (95% CI 0.98 - 1.00); compared to SIRS, which demonstrates an AUROC of 0.75, 0.72, and 0.84, respectively.

Comparing *InSight's* performance across the three sepsis-related gold standards, it is clear that the septic shock criteria are relatively less challenging to anticipate, as its four hour prediction metrics are stronger than those for the detection of both sepsis and severe sepsis. Accordingly, we display the four hour prior to onset prediction case for septic shock (Fig 2C), where existing tools fail to adequately meet prediction standards relevant for sound clinical use. Four hours in advance of septic shock onset, *InSight* achieved an AUROC of 0.96 (95% CI 0.94 - 0.98)

Additional comparison metrics at time of detection for each gold standard are available in Table 2. Sensitivity was fixed near 0.80, in order to compare the specificities from each gold standard and comparator. Across all gold standards, a sensitivity of 0.80 results in a high specificity for *InSight*; however, the sensitivities for MEWS, SOFA, and SIRS are significantly lower. Notably, at 0.80 sensitivity, *InSight* achieves a specificity of 0.95 for sepsis, 0.84 for severe sepsis, and 0.99 for septic shock detection.

Table 2: Performance metrics for three sepsis gold standards at time of onset (zero hour), with sensitivities fixed at or near 0.80. The three values per cell correspond to sepsis (green), severe sepsis (orange), and septic shock (red), respectively.

	<i>InSight</i> (95% CI)	MEWS	SOFA	SIRS
AUROC	0.95 (0.93 - 0.97)	0.76	0.71	0.75
	0.90 (0.88 - 0.92)	0.77	0.76	0.72
	0.99 (0.98 - 1.00)	0.94	0.88	0.84
Sensitivity (Fixed near 0.80 for comparison)	0.80	0.72	0.79	0.82
	0.80	0.72	0.79	0.80
	0.80	0.87	0.75	0.70
Specificity	0.95	0.72	0.45	0.51
	0.84	0.72	0.51	0.51

	0.99	0.90	0.80	0.86
--	------	------	------	------

In addition to *InSight's* ability to detect sepsis, severe sepsis, and septic shock, Figure 3A illustrates the ROC of severe sepsis detection and prediction four hours prior to severe sepsis onset. Even four hours in advance, the *InSight* severe sepsis AUROC is 0.85 (95% CI 0.79 - 0.91), significantly higher than the onset time SIRS result of 0.75 AUROC. Figure 3B summarizes *InSight's* predictive advantage, using the severe sepsis gold standard, over MEWS, SOFA, and SIRS at the same time points in the hours leading up to onset. *InSight* maintains a high AUROC in the continuum up to four hours preceding severe sepsis onset. *InSight's* predictions four hours in advance produce a sensitivity and specificity that are greater than the at-onset time sensitivity and specificity of each MEWS, SOFA, and SIRS (Table 2, Fig. 3B).

In our second set of experiments, we validated *InSight's* performance in the presence of missing data. We tested *InSight's* ability to detect severe sepsis at time of onset with various rates of data dropout. Table 3 presents the results of these experiments. After randomly deleting data from the test set with a probability of 0.10, *InSight's* AUROC for severe sepsis detection is 0.82. Dropping approximately 60% of the test set measurements results in an AUROC of 0.75, demonstrating *InSight's* robustness to missing data. Of note, the AUROC of *InSight* at 60% data dropout achieves comparable performance to SIRS with no missing data. These results are useful in estimating *InSight's* performance in institutions or specific care units where measurements may be taken less frequently or have reduced availability.

Table 3: *InSight's* severe sepsis screening performance at time of onset in the presence of data sparsity, compared to SIRS with a full data complement.

	<i>InSight</i>					SIRS
% Data Missing	0%	10%	20%	40%	60%	0%
AUROC	0.90	0.82	0.79	0.76	0.75	0.72

Sensitivity	0.80	0.80	0.80	0.80	0.80	0.80
Specificity	0.84	0.66	0.57	0.50	0.49	0.51

Transfer Learning

InSight is flexible by design, and can be easily trained on an appropriate retrospective data set before being applied to a new patient population. However, sufficient historical patient data is not always available for training on the target population. We evaluated *InSight*'s performance when trained on a mixture of the MIMIC-III data together with increasing amounts of UCSF training data, and then tested on a separate hold-out UCSF patient population using transfer learning. In Figure 4, we show that the performance of the algorithm improves as the fraction of UCSF target population data used in training increases.

Discussion

We have validated the machine learning algorithm, *InSight*, on the mixed-ward data of UCSF, which includes patients from the ED and floor units as well as the ICU, with varying types and frequencies of patient measurements. *InSight* outperformed commonly-used disease severity scores such as SIRS, MEWS, and SOFA for the screening of sepsis, severe sepsis, and septic shock (Figure 2). These results, shown in Table 2, confirm *InSight*'s strength in predicting these sepsis-related gold standard outcomes. To the authors' knowledge, *InSight* is first sepsis screening system to meet or exceed an AUROC of 0.90 using only vital sign inputs, on each of the sepsis gold standards evaluated in this study. Additionally, *InSight* provides predictive capabilities in advance of sepsis onset, aided by the analysis of trends and correlations between vital sign measurements. This advantage is apparent in the comparison with SIRS made in Figure 3A. Up to four hours prior to severe sepsis onset, *InSight* maintains a high AUROC above 0.85 (Figure 3). This advance warning of patients trending toward severe sepsis could extend the window for meaningful clinical intervention.

InSight uses only six common vital signs derived from a patient's EHR to detect sepsis onset, as well as to predict those patients most at risk for developing sepsis. The decreased performance of *InSight* for recognition of severe sepsis relative to sepsis onset may be in part

1
2
3 because the organ failure characteristic of severe sepsis is more easily recognized through
4 laboratory tests for organ function. Because we have not incorporated metabolic function panels
5 in this validation of *InSight*, the detection of organ failure using only six common vital signs may
6 be more difficult. In practice, *InSight* is adaptable to different inputs and is able to incorporate
7 laboratory results as they become available. Inclusion of these results may well increase the
8 performance of *InSight* for the detection and prediction of severe sepsis. However, in this work
9 we have chosen to benchmark the performance of *InSight* using only six commonly measured
10 vital signs. The ordering of metabolic panel laboratory tests are often predicated on clinician
11 suspicion of severe sepsis, and therefore, early or developing cases may be missed. Additionally,
12 because these vital sign inputs do not require time-dependent laboratory results or additional
13 manual data entry, surveillance by *InSight* is frequent, and as a result, sepsis conditions are
14 detected in a more timely manner. Minimal data requirements also lighten the burden of
15 implementation in a clinical setting and broaden the potential clinical applications of *InSight*.

16
17 Although *InSight* uses only a handful of clinical variables, it maintains a high level of
18 performance in experiments with randomly missing data. We demonstrate in Table 3 that for the
19 detection of severe sepsis, even with up to 60% of randomized test patient data missing, *InSight*
20 still achieves comparable performance to SIRS calculated with complete data availability. These
21 results validate the accuracy and clinical value of our machine learning algorithm even under
22 conditions of data scarcity.

23
24 Additionally, we have investigated the customizability of *InSight* to local hospital
25 demographics and measurements. The incorporation of site-specific data into the training set
26 using transfer learning improves performance on test sets, over that of a training set comprised
27 entirely of an independent population. This indicates that it may be possible to adequately train
28 *InSight* for use in a new clinical setting, while still predominantly using existing retrospective
29 data from other institutions.

30
31 Our previous studies have investigated *InSight* applied to individual sepsis standards such
32 as the SIRS standard for sepsis [21], severe sepsis [22], and septic shock [23], on the MIMIC
33 retrospective datasets. We have also developed a related algorithm to detect patient stability [24]
34 and predict mortality [23, 25]. However, this study is the first to apply *InSight* to all three
35 standard sepsis definitions simultaneously, and to validate the algorithm on a mixed ward
36 population, including ED, ICU and floor wards from UCSF. This study is also the first to use
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 only six minimal vital signs, without utilizing a mental status evaluation such as Glasgow Coma
4 Score, or even age, in the detection and prediction of those sepsis standards. Additionally, this
5 study demonstrates the adaptability of the machine learning algorithm to an entirely new patient
6 data set with markedly different demographics and outcomes.
7
8
9

10 11 12 **Limitations**

13
14 While we incorporated data from both UCSF and MIMIC-III, we cannot claim
15 generalizability of our results to other populations on the basis of this study alone. However, we
16 are aided by the minimality of data used to make predictions; because *InSight* requires only six
17 of the most basic and widely-available clinical measurements, it is likely that it will perform
18 similarly in other settings if vital sign data is available. The gold standard references we use to
19 determine sepsis, severe sepsis and septic shock rely on ICD-9 codes from the hospital database.
20 The administrative coding procedures may vary by hospital and do not always precisely
21 reproduce results from manual chart review for sepsis diagnosis. although ICD-9 codes have
22 been previously validated for accuracy in the detection of severe sepsis [26]. This study was
23 conducted retrospectively, and so we are unable to make claims regarding performance in a
24 prospective setting, which involves the interpretation and use of *InSight*'s predictions by
25 clinicians. We intend to evaluate these algorithms in prospective clinical studies in future work.
26
27
28
29
30
31
32
33
34
35
36
37

38 **Conclusions**

39
40 We have validated the machine learning algorithm, *InSight*, in a multicenter study in a
41 mixed-ward population from UCSF and an ICU population from BIDMC. *InSight* provides high
42 sensitivity and specificity for the detection and prediction of sepsis, severe sepsis, and septic
43 shock using the analysis of only six common vital signs taken from the electronic health record.
44 *InSight* outperforms scoring systems in current use for the detection of sepsis, is robust to a
45 significant amount of missing patient data, and can be customized to novel sites using a limited
46 amount of site-specific data. Our results indicate that *InSight* outperforms tools currently used
47 for sepsis detection and prediction, which may lead to improvements in sepsis-related patient
48 outcomes.
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7 **Acknowledgments:** We acknowledge the assistance of Siddharth Gampa and Emily Huynh for
8 editing contributions. We acknowledge Yifan Zhou for assistance with data calculations and
9 figure generation. We thank Dr. Hamid Mohamadlou and Dr. Thomas Desautels for
10 contributions to the development of the machine learning algorithm *InSight*.
11

12 **Author Statement:** QM, JC, and RD conceived the described experiments. DS acquired the
13 UCSF data. QM and YZ executed the experiments. QM, RD, JC, and MJ interpreted the results.
14 QM, MJ, and JH wrote the manuscript. QM, RD, MJ, JH, JC, CB, DS, LS, UC, GF, and YK
15 revised the manuscript, with assistance from Emily Huynh and Siddharth Gampa. All authors
16 approved the version to be published and agree to be accountable for all aspects of the work in
17 ensuring that questions related to the accuracy or integrity of any part of the work are
18 appropriately investigated and resolved.
19

20 **Competing Interests:** All authors who have affiliations listed with Dascena (Hayward, CA,
21 USA) are employees of Dascena. Dr. Barton reports receiving consulting fees from Dascena. Dr.
22 Barton, Dr. Shimabukuro and Dr. Fletcher report receiving grant funding from Dascena.
23

24 **Funding:** Research reported in this publication was supported by the National Institute of
25 Nursing Research, of the National Institutes of Health, under award number R43NR015945. The
26 content is solely the responsibility of the authors and does not necessarily represent the official
27 views of the National Institutes of Health. The funder had no role in the conduct of the study;
28 collection, management, analysis, and interpretation of data; preparation, review, and approval of
29 the manuscript; and decision to submit the manuscript for publication.
30

31 **Data Sharing:** Data from UCSF may be made available to qualified researchers upon request.
32 MIMIC-III is a publicly available database.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

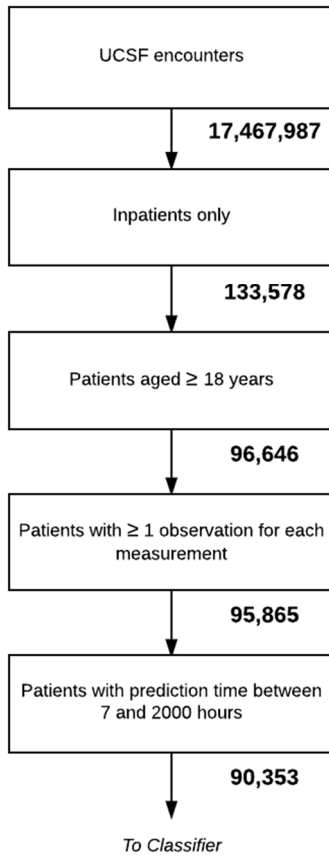
49 **References**

- 50
51 1. Murphy SL, Xu J, Kochanek KD. Deaths: final data for 2010. National vital statistics
52 reports: from the Centers for Disease Control and Prevention, National Center for Health
53 Statistics, National Vital Statistics System. 2013;61: 1-17.
54
55
56
57
58
59
60

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
2. Angus, Derek C et al. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Crit Care Med*. 2001;29: 1303-1310.
 3. Stevenson, Elizabeth K et al. Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis. *Crit Care Med*. 2014;42: 625.
 4. Pfunter A, Wier LM, Steiner C. Costs for Hospital Stays in the United States, 2010: Statistical Brief #146. In: *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs* [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2006 Feb-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK121966/>
 5. O'Brien, J. The Cost of Sepsis. *CDC Safe Healthcare Blog*. 2015. Available from: <https://blogs.cdc.gov/safehealthcare/the-cost-of-sepsis/#ref>
 6. Gaieski DF, Edwards JM, Kallan MJ, Carr BG. Benchmarking the incidence and mortality of severe sepsis in the United States. *Crit Care Med*. 2013;41: 1167-1174.
 7. Rivers, Emanuel et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New Engl J Med*. 2001;345: 1368-1377.
 8. Nguyen, H Bryant et al. Implementation of a bundle of quality indicators for the early management of severe sepsis and septic shock is associated with decreased mortality. *Crit Care Med*. 2007;35: 1105-1112.
 9. Kumar, Anand et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med*. 2006;34: 1589-1596.
 10. Levy, Mitchell M et al. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Intensive Care Med*. 2003;29: 530-538.
 11. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016;315: 801-10.
 12. Subbe CP, Slater A, Menon D, Gemmell L. Validation of physiological scoring systems in the accident and emergency department. *Emerg Med J*. 2006;23:841-845.
 13. Rangel-Frausto MS, Pittet D, Costigan M, Hwang T, Davis CS, Wenzel RP. The natural history of the systemic inflammatory response syndrome (SIRS): a prospective study. *JAMA*. 1995;273:117-123.

14. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med.* 1996;22: 707-710.
15. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data.* 2016. doi: 10.1038/sdata.2016.35.
16. PostgreSQL Global Development Group, <https://www.postgresql.org/>
17. G. Van Rossum. The Python Language Reference Manual. Network Theory Ltd. Python Software Foundation. 2003. Available from: <https://www.python.org/>
18. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med.* 2015;7:299ra122. doi: 10.1126/scitranslmed.aab3719.
19. Fullerton JN, Price CL, Silvey NE, Brace SJ, Perkins GD. Is the Modified Early Warning Score (MEWS) superior to clinician judgement in detecting critical illness in the pre-hospital environment? *Resuscitation.* 2012;83: 557-562.
20. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. *Mach Learn.* 2010;79: 151-175.
21. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. *Comp Biol Med.* 2016;74: 69-73.
22. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform.* 2016;4: e28.
23. Calvert J, Desautels T, Chettipally U, Barton C, Hoffman J, Jay M, Mao Q, Mohamadlou H, Das R. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Ann Med Surg.* 2016;8: 50-55.
24. Calvert JS, Price DA, Barton CW, Chettipally UK, Das R. Discharge recommendation based on a novel technique of homeostatic analysis. *J Am Med Inform Assoc.* 2016;24: 24-29.
25. Calvert J, Mao Q, Hoffman JL, Jay M, Desautels T, Mohamadlou H, et al. Using electronic health record collected clinical variables to predict medical intensive care unit mortality. *Ann Med Surg.* 2016;11: 52-57.

26. Iwashyna TJ, Odden A, Rohde J, Bonham C, Kuhn L, Malani P, et al. Identifying patients with severe sepsis using administrative claims: patient-level validation of the angus implementation of the international consensus conference definition of severe sepsis.



Med Care. 2014;52:e39.

Figure 1: Patient inclusion flow diagram for the UCSF data set.

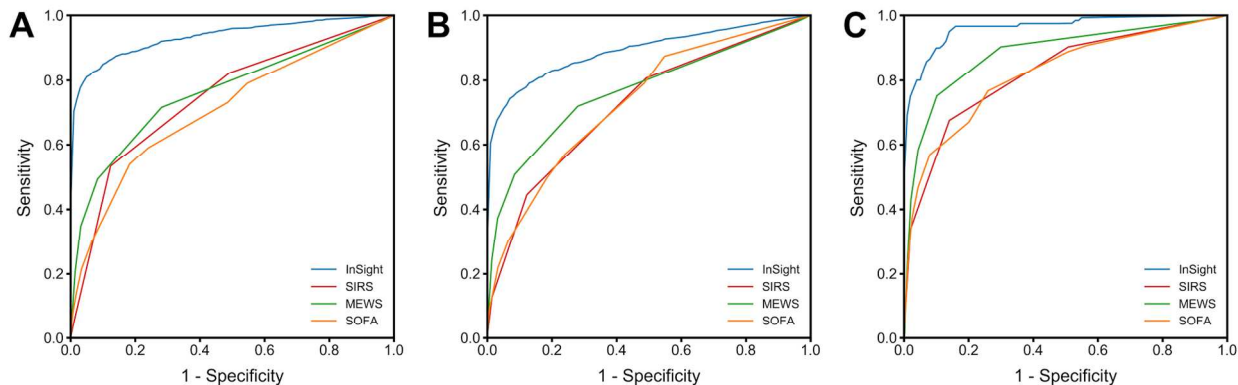


Figure 2: ROC curves for *InSight* and common scoring systems at time of (A) sepsis onset, (B) severe sepsis onset, and (C) four hours before septic shock onset.

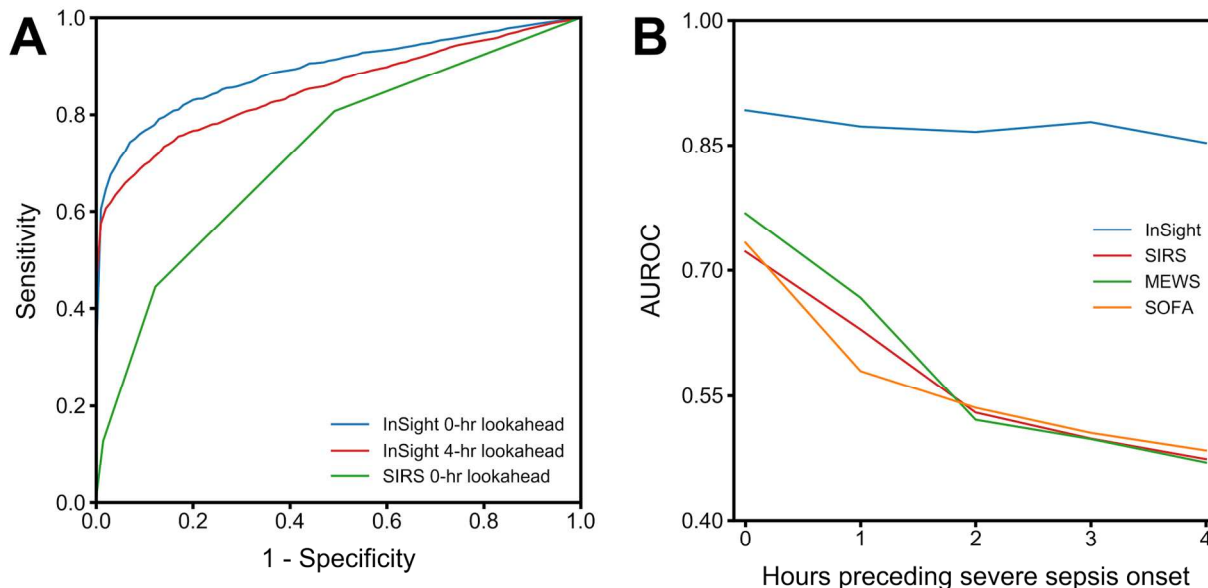


Figure 3: A) ROC detection (zero hour, blue) and prediction (four hour prior to onset, red) curves using *InSight* and ROC detection (zero hour, green) curve for SIRS, with the severe sepsis gold standard. B) Predictive performance of *InSight* and comparators, using the severe sepsis gold standard, as a function of time prior to onset.

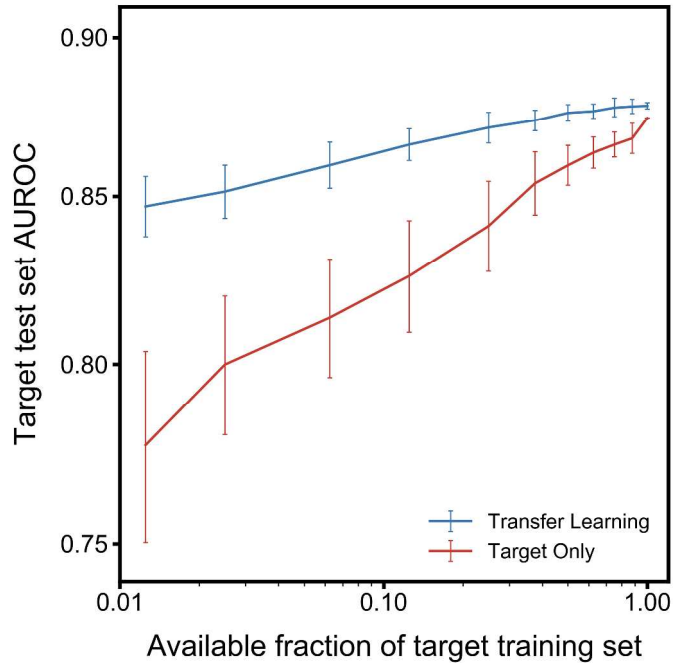


Figure 4. Learning curves (mean AUROC on the UCSF target data set) with increasing number of target training examples. Error bars represent the standard deviation. When data availability of the target set is low, target-only training exhibits lower AUROC values and high variability.

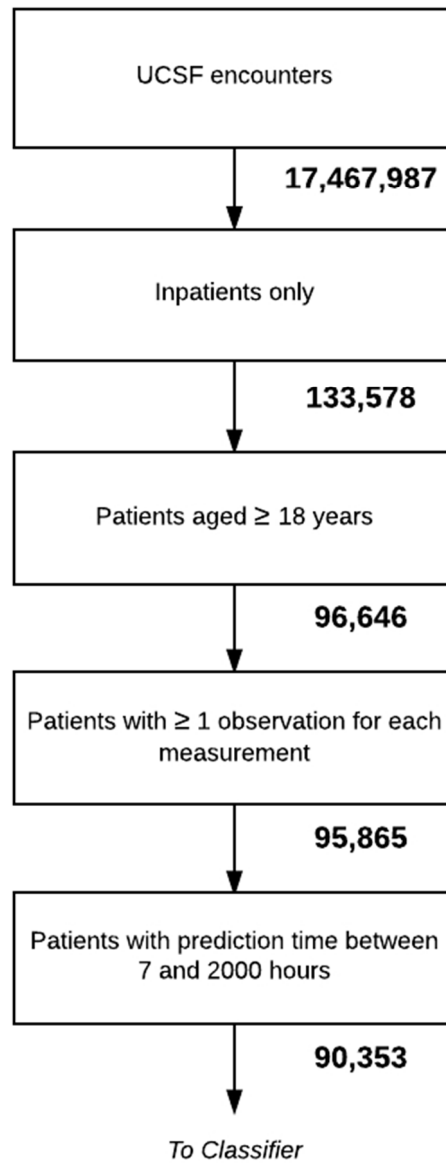


Figure 1: Patient inclusion flow diagram for the UCSF data set.

94x230mm (96 x 96 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

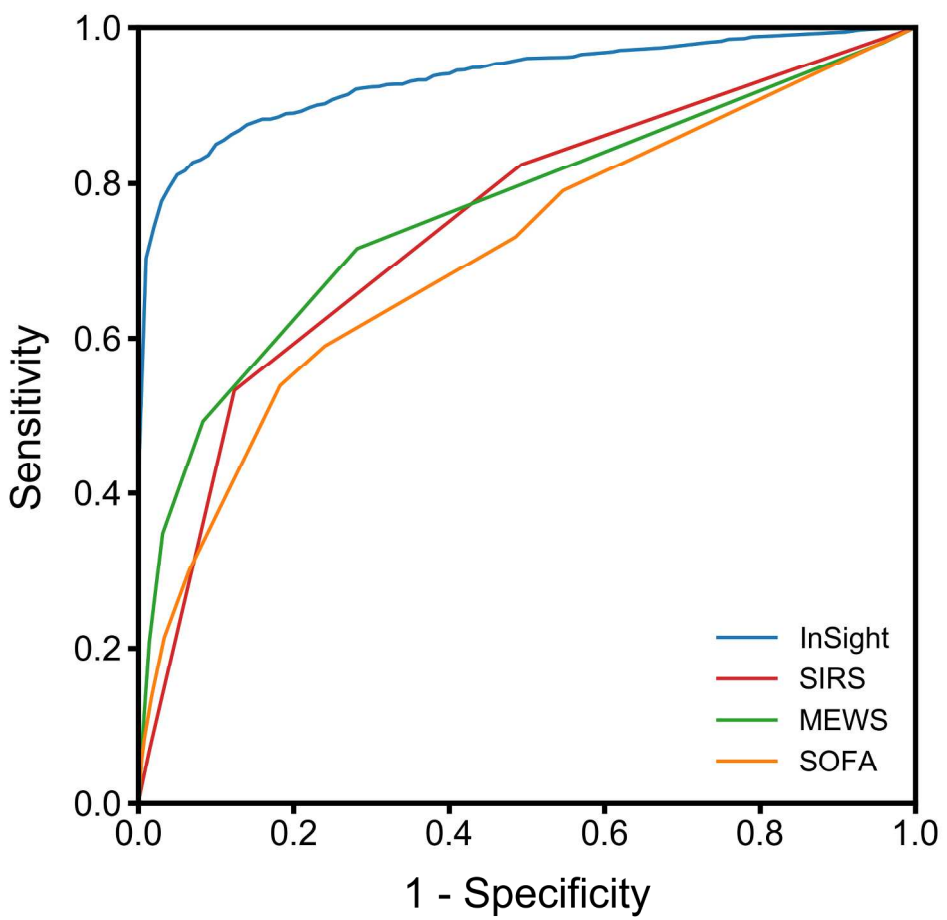


Figure 2A: ROC curves for InSight and common scoring systems at time of (A) sepsis onset, (B) severe sepsis onset, and (C) four hours before septic shock onset.

203x203mm (300 x 300 DPI)



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

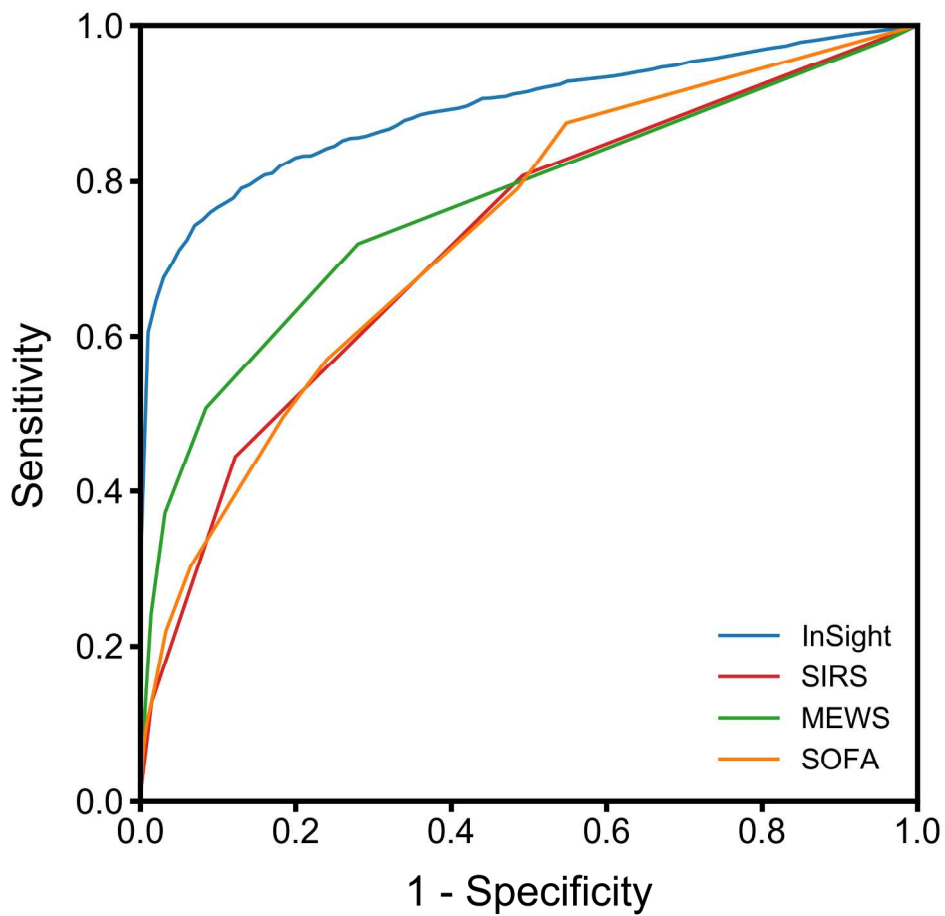


Figure 2B: ROC curves for InSight and common scoring systems at time of (A) sepsis onset, (B) severe sepsis onset, and (C) four hours before septic shock onset.

203x203mm (300 x 300 DPI)



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

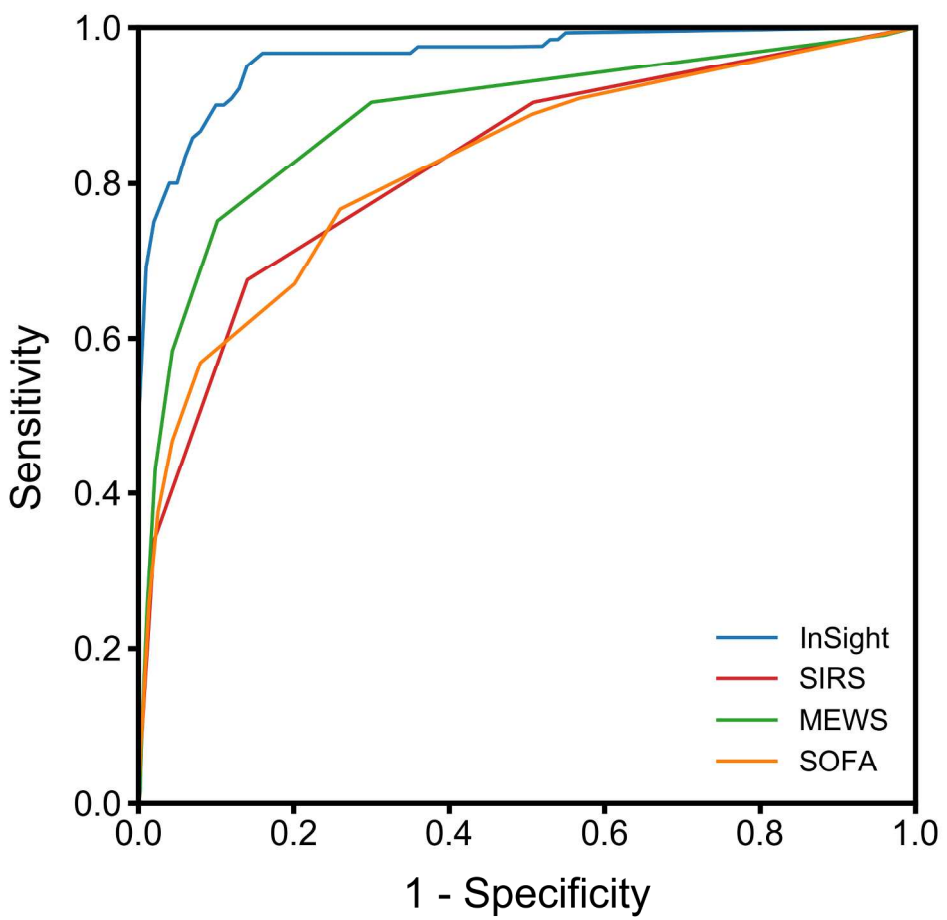


Figure 2C: ROC curves for InSight and common scoring systems at time of (A) sepsis onset, (B) severe sepsis onset, and (C) four hours before septic shock onset.

203x203mm (300 x 300 DPI)



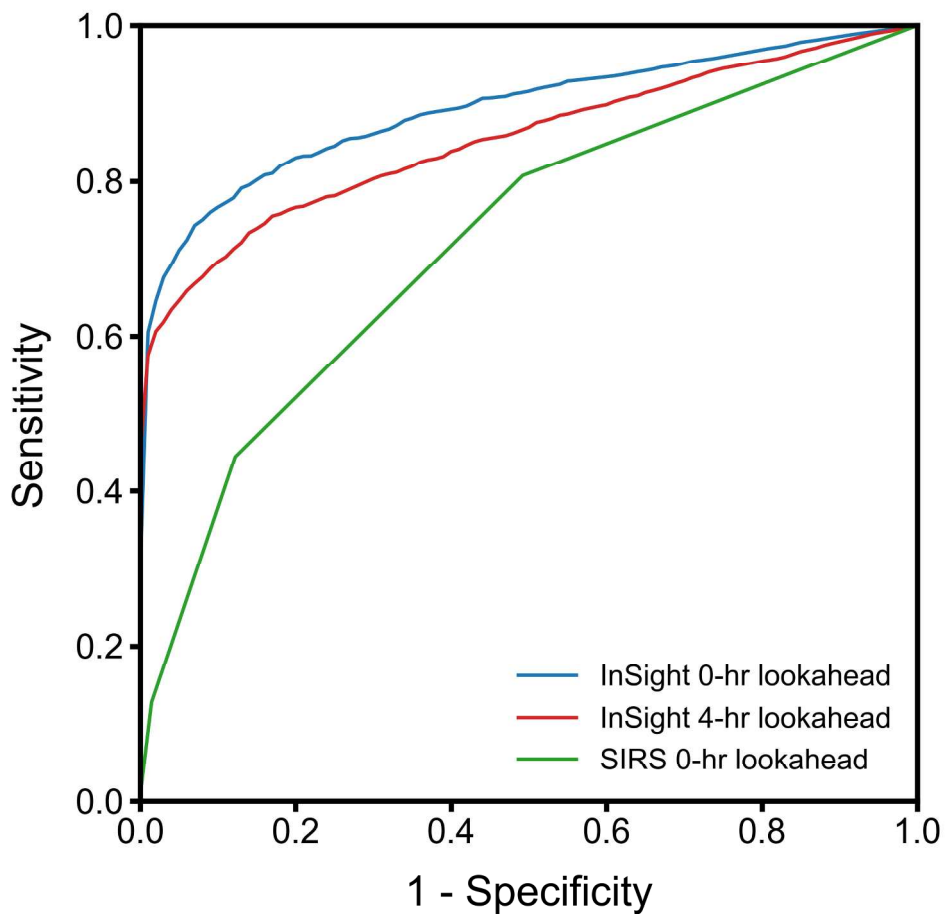


Figure 3A: A) ROC detection (zero hour, blue) and prediction (four hour prior to onset, red) curves using InSight and ROC detection (zero hour, green) curve for SIRS, with the severe sepsis gold standard. B) Predictive performance of InSight and comparators, using the severe sepsis gold standard, as a function of time prior to onset.

203x203mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

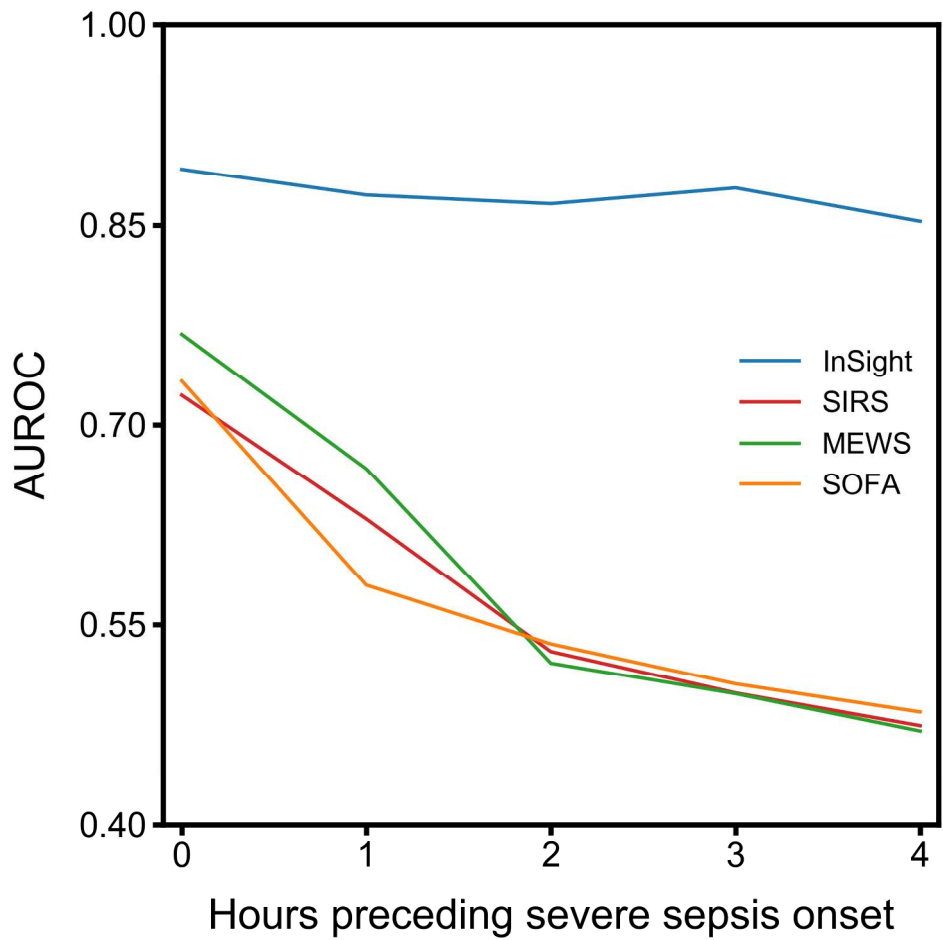


Figure 3B: A) ROC detection (zero hour, blue) and prediction (four hour prior to onset, red) curves using InSight and ROC detection (zero hour, green) curve for SIRS, with the severe sepsis gold standard. B) Predictive performance of InSight and comparators, using the severe sepsis gold standard, as a function of time prior to onset.

203x203mm (300 x 300 DPI)

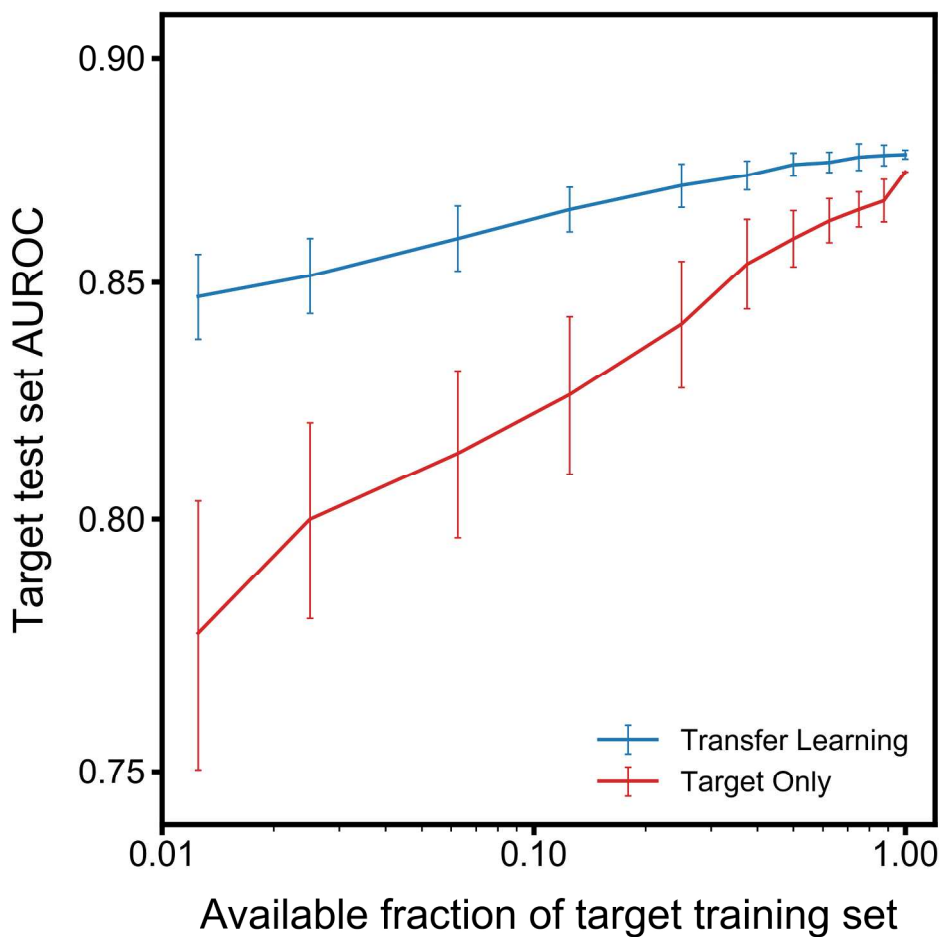


Figure 4. Learning curves (mean AUROC on the UCSF target data set) with increasing number of target training examples. Error bars represent the standard deviation. When data availability of the target set is low, target-only training exhibits lower AUROC values and high variability.!! †

203x203mm (300 x 300 DPI)



BMJ Open

Validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2017-017833.R1
Article Type:	Research
Date Submitted by the Author:	10-Aug-2017
Complete List of Authors:	Mao, Qingqing; Dascena Jay, Melissa; Dascena hoffman, jana; Dascena Calvert, Jake; Dascena Barton, Christopher; University of California San Francisco, Emergency Medicine Shimabukuro, David; University of California San Francisco, Anesthesia and Perioperative Care Shieh, Lisa; Stanford University School of Medicine, Medicine Chettipally, Uli ; University of California San Francisco, Emergency Medicine; Kaiser Permanente South San Francisco Medical Center Fletcher, Grant; University of Washington School of Medicine Kerem, Yaniv; Stanford University School of Medicine Zhou, Yifan; University of California Berkeley, Statistics Das, Ritankar; Dascena
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Diagnostics, Infectious diseases, Intensive care, Emergency medicine
Keywords:	sepsis, septic shock, clinical decision support, prediction, machine learning, electronic health records

SCHOLARONE™
Manuscripts

Validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU

Qingqing Mao¹, Melissa Jay¹, Jana L. Hoffman^{1*}, Jacob Calvert¹, Christopher Barton², David Shimabukuro³, Lisa Shieh⁴, Uli Chettipally^{2,5}, Grant Fletcher⁶, Yaniv Kerem^{7,8}, Yifan Zhou^{1,9}
Ritankar Das¹

¹ Dascena Inc., Hayward, CA, United States

² Department of Emergency Medicine, University of California San Francisco, San Francisco, CA, United States

³ Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, CA, United States

⁴ Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

⁵ Kaiser Permanente South San Francisco Medical Center, South San Francisco, CA, United States

⁶ Division of Internal Medicine, University of Washington School of Medicine, Seattle, WA, United States

⁷ Department of Clinical Informatics, Stanford University School of Medicine, Stanford, CA, United States

⁸ Department of Emergency Medicine, Kaiser Permanente Redwood City Medical Center, Redwood City, CA, United States

⁹ Department of Statistics, University of California Berkeley, Berkeley, CA, United States

* Corresponding author

Email: Jana@Dascena.com
22710 Foothill Blvd., Suite #2
Hayward, CA 94541

Keywords: sepsis, septic shock, clinical decision support, prediction, machine learning, electronic health records

Word count: 5,250

Abstract

Objectives: We validate a machine learning-based sepsis prediction algorithm (*InSight*) for the detection and prediction of three sepsis-related gold standards, using only six common vital signs. We also evaluate *InSight's* robustness to missing data, and assess customization of the algorithm to site-specific data using transfer learning.

Design: We used a machine learning algorithm with gradient tree boosting, relying solely on data from six vital signs to train *InSight*. Relevant features for prediction were created from combinations of vital sign measurements and their changes over time.

Setting: A mixed-ward (emergency and inpatient) retrospective data set from the University of California, San Francisco (UCSF) Medical Center, and an intensive care unit data set from the Beth Israel Deaconess Medical Center, as a transfer learning data source.

Participants: 90,353 adult emergency and inpatient encounters from June 2011 to March 2016.

Interventions: none

Primary and secondary outcome measures: Area under the receiver operating characteristic curve (AUROC) for detection and prediction of sepsis, severe sepsis, and septic shock.

Results: In the detection of sepsis and severe sepsis, *InSight* achieves an area under the receiver operating characteristic (AUROC) curve of 0.92 (95% CI 0.90 - 0.93) and 0.87 (95% CI 0.86 - 0.88), respectively. Four hours prior to onset, *InSight* predicts septic shock with an AUROC of 0.96 (95% CI 0.94 - 0.98), and severe sepsis onset with an AUROC of 0.85 (95% CI 0.79 - 0.91).

Conclusions: *InSight* outperforms existing sepsis scoring systems in both identification and prediction of sepsis, severe sepsis, and septic shock. This is the first sepsis screening system to exceed an AUROC of 0.90 using only vital sign inputs. *InSight* is robust to significant amounts of missing data, and can be customized to a novel hospital data set using a small fraction of site data.

Strengths and limitations of this study

- Machine learning is applied to the detection and prediction of three separate sepsis standards in the emergency department, general ward and intensive care settings.
- Only six commonly measured vital signs are used as input for the algorithm.
- The algorithm is robust to randomly missing data.
- Transfer learning successfully leverages large dataset information to a target dataset.
- Retrospective nature of the study does not predict clinician reaction to information.

Introduction

Sepsis is a major health crisis and one of the leading causes of death in the United States [1]. Approximately 750,000 hospitalized patients are diagnosed with severe sepsis in the United States annually, with an estimated mortality rate of up to one-third [2,3]. The cost burden of sepsis is disproportionately high, with estimated costs of \$20.3 billion dollars annually, or \$55.6 million per day in US hospitals [4]. Additionally, the average hospital stay for sepsis is twice as expensive as other conditions [5], and the average incidence of severe sepsis is increasing by approximately 13% per year [6]. Early diagnosis and treatment have been shown to reduce mortality and associated costs [7-9]. Despite clear benefits, early and accurate sepsis detection remains a difficult clinical problem.

Sepsis has been defined as a dysregulated host response to infection. In practice, sepsis can be challenging to recognize because of the heterogeneity of the host response to infection, and the diversity of possible infectious insult. Sepsis has been traditionally recognized as two or more Systemic Inflammatory Response Syndrome (SIRS) [10] criteria together with a known or suspected infection; progressing to severe sepsis, in the event of organ dysfunction; and finally to septic shock, which additionally includes refractory hypotension [10]. However, ongoing debates over sepsis definitions and clinical criteria, as evidenced by the recent proposed redefinitions of sepsis [11], underscore a fundamental difficulty in the identification and accurate diagnosis of sepsis.

Various rule-based disease severity scoring systems are widely used in hospitals in an attempt to identify septic patients. These scores, such as the Modified Early Warning Score (MEWS) [12], the Systemic Inflammatory Response Syndrome (SIRS) criteria [13], and the Sequential Organ Failure Assessment (SOFA) [14], are manually tabulated at the bedside and lack accuracy in sepsis diagnosis. However, the increasing prevalence of Electronic Health Records (EHR) in clinical settings provides an opportunity for enhanced patient monitoring and increased early detection of sepsis.

This study validates a machine learning algorithm *InSight*, which uses only six vital signs taken directly from the EHR, in the detection and prediction of sepsis, severe sepsis, and septic shock in a mixed-ward population at the University of California, San Francisco (UCSF). We

1
2
3 investigate the effects of induced data sparsity on *InSight* performance, and compare all results
4 with other scores that are commonly used in the clinical setting for the detection and prediction
5 of sepsis. Furthermore, we apply a transfer learning scheme to customize a Multiparameter In-
6 telligent Monitoring in Intensive Care (MIMIC)-III-trained algorithm to the UCSF patient popu-
7 lation using a minimal amount of UCSF-specific data.
8
9
10
11

12 13 **Methods**

14 15 **Data sets**

16
17
18
19
20 We used a data set provided by the UCSF Medical Center representing patient stays from
21 June 2011 to March 2016 in all experiments. The UCSF data set contains 17,467,987 hospital
22 encounters, including inpatient and outpatient visits to all units within the UCSF medical system.
23 The data were de-identified to comply with the Health Insurance Portability and Accountability
24 Act (HIPAA) Privacy Rule. For transfer learning, we used the Multiparameter Intelligent Moni-
25 toring in Intensive Care (MIMIC)-III v1.3 data set, compiled from the Beth Israel Deaconess
26 Medical Center (BIDMC) in Boston, MA between 2001 and 2012, composed of 61,532 ICU
27 stays [15]. This database is a publicly available database constructed by researchers at MIT's
28 Laboratory for Computational Physiology, and the data were also de-identified in compliance
29 with HIPAA. Data collection for the MIMIC-III and UCSF datasets did not impact patient safety.
30 Therefore, this study constitutes non-human subjects research, which does not require Institu-
31 tional Review Board approval.
32
33
34
35
36
37
38
39
40
41
42
43

44 45 **Data Extraction and Imputation**

46
47 The data were provided in the form of comma separated value (CSV) files and stored in a
48 PostgreSQL [16] database. Custom SQL queries were written to extract measurements and pa-
49 tient outcomes of interest. The measurement files were then binned by hour for each patient. To
50 be included, patients were required to have at least one of each type of measurement recorded
51 during the encounter. If a patient did not have a measurement in a given hour, the missing meas-
52 urement was filled in using carry-forward imputation. This imputation method applied the pa-
53 tient's last measured value to the following hour (a causal procedure). In the case of multiple
54
55
56
57
58
59
60

1
2
3 measurements within an hour, the mean was calculated and used in place of an individual meas-
4
5 urement. Because patient data was standardized into single hourly measurements before being
6
7 fed into the classifier, any information related to frequency of data collection was lost before
8
9 predictions were made. After the data were processed and imputed in Python [17], they were
10
11 used to train the *InSight* classifier and test its predictions at sepsis onset and at fixed time points
12
13 prior to onset.

14 15 16 **Gold Standards**

17
18 In this study, we tested *InSight*'s performance according to various gold standards (clini-
19
20 cal indications). We investigated *InSight*'s ability to predict and detect sepsis, severe sepsis, and
21
22 septic shock. Further, we compared *InSight*'s performance to SIRS, MEWS, and SOFA, for each
23
24 of the following gold standards. For training and testing the algorithm, we conservatively identi-
25
26 fied each septic condition by requiring that the ICD-9 code corresponding to the diagnosis was
27
28 coded for each positive case, in addition to meeting the clinical requirements for the definition of
29
30 each septic standard as defined below.

31 32 **Sepsis**

33
34 The sepsis gold standard was determined using the 2001 consensus sepsis definition [10]:
35
36 “the presence of two or more SIRS criteria paired with a suspicion of infection.” To identify a
37
38 case as positive for sepsis, we required ICD-9 code 995.91. The onset time was defined as the
39
40 first time two or more SIRS criteria were met within the same hour. SIRS criteria are defined as:

- 41 ● heart rate > 90 beats/ min,
- 42 ● body temperature > 38 °C or < 36 °C,
- 43 ● respiratory rate >20 breaths/min or PaCO₂ < 32 mmHg, and
- 44 ● white blood cell count > 12,000 cells/μL or < 4,000 cells/μL. [10]

45 46 47 **Severe Sepsis**

48
49 The severe sepsis gold standard used the definition of severe sepsis as “organ dysfunction
50
51 caused by sepsis” which can be represented by one or more of the criteria below, and identified
52
53 for patients with the severe sepsis ICD-9 code 995.92. We assigned the severe sepsis onset time
54
55
56
57
58
59
60

1
2
3 to be the first instance during which two SIRS criteria as described above and one of the follow-
4
5 ing organ dysfunction criteria were met within the same hour.
6

- 7 ● Lactate > 2 mmol/L
- 8 ● Systolic blood pressure < 90 mmHg
- 9 ● Urine output < 0.5 mL/kg, over two hours, prior to organ dysfunction after fluid resusci-
10 tation
- 11 ● Creatinine > 2 mg/dL without renal insufficiency or chronic dialysis
- 12 ● Bilirubin > 2 mg/dL without having liver disease or cirrhosis
- 13 ● Platelet count < 100,000 μ L
- 14 ● International normalized ratio > 1.5
- 15 ● PaO₂/FiO₂ < 200 in addition to pneumonia
- 16 ● PaO₂/FiO₂ < 250 with acute kidney injury but without pneumonia
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26

27 **Septic Shock**

28 We identified as positive cases for septic shock those patients who received the septic shock
29 ICD-9 code 785.52 and additionally demonstrated the following conditions:
30

- 31 ● systolic blood pressure of < 90 mmHg, defined as hypotension, for at least 30 minutes,
32 and
33
- 34 ● who were resuscitated with \geq 20 ml/kg over a 24 hour period, or
- 35 ● who received \geq 1200 ml in total fluids. [18]
- 36
- 37
- 38

39 The onset time was defined as the first hour when either the hypotension or fluid resuscitation
40 criterion was met.
41

42 **Calculating Comparators**

43 We compared *InSight* predictions for each gold standard to three common patient deterioration
44 scoring systems: SIRS, SOFA, and MEWS. Area under the receiver operating characteristic
45 (AUROC) curve, sensitivity, and specificity were compared across all prediction models. The
46 SIRS criteria, as explained in the sepsis definition, were evaluated independently of the suspicion
47 of infection. To calculate the SOFA score, we collected each patient's PaO₂/FiO₂, Glasgow Co-
48 ma Score, mean arterial blood pressure or administration of vasopressors, bilirubin level, platelet
49 counts, and creatinine level. Each of the listed measurements is associated with a SOFA score of
50
51
52
53
54
55
56
57
58
59
60

1
2
3 1-4, based on severity level, as described by Vincent et al. [14]. After receiving a score for each
4 of the six organ dysfunction categories, the overall SOFA score was computed as the sum of the
5 category scores and used as a comparator to *InSight*. Finally, the MEWS score, which ranges
6 from 0 (normal) to 14 (high risk of deterioration), was determined by tabulating subscores for
7 heart rate, systolic blood pressure, respiratory rate, temperature, and Glasgow Coma Score. We
8 used the subscore system presented in Fullerton et al. [19] to compute each patient's MEWS
9 score.
10
11
12
13
14

15 16 17 18 **Measurements and Patient Inclusion**

19
20 In order to generate *InSight* scores, patient data were analyzed from each of the following
21 six clinical vital sign measurements: systolic blood pressure, diastolic blood pressure, heart rate,
22 respiratory rate, peripheral capillary oxygen saturation (SpO₂), and temperature. We used only
23 vital signs, which are frequently available and routinely taken in the ICU, ED, and floor units.
24 Patient data were used from the course of a patient's hospital encounter, regardless of the unit the
25 patient was in when the data were collected.
26
27
28
29

30 All patients over the age of 18 were considered for this study. For a given encounter, if
31 the patient was admitted to the hospital from the ED, the start of the ED visit is where the analy-
32 sis began. Patients in our final data sets were required to have at least one measurement for each
33 of the six vital signs. In order to ensure enough data to accurately characterize sepsis predictions
34 at four hours pre-onset, we further limited the study group to exclude patients whose septic con-
35 dition onset time was within seven hours after the start of their record, which was either the time
36 of admission to the hospital or the start of their ED visit; the latter was applicable only if the pa-
37 tient was admitted through the ED. A smaller window to sepsis onset time would have resulted
38 in insufficient testing data to make 4-hour prediction possible in some cases, which would inap-
39 propriately affect performance metrics such as sensitivity and specificity. Patients with sepsis
40 onset after 2,000 hours post-admission were also excluded, to limit the data analysis matrix size.
41 The final UCSF data set included 90,353 patients (Fig.1) and the MIMIC-III data set contained
42 21,604 patients, following the same inclusion criteria.
43
44
45
46
47
48
49
50
51
52

53 After patient exclusion, our final group of UCSF patients was composed of 55% women
54 and 45% men with a median age of 55. The median hospital length of stay was 4 days, IQR =
55 (2,6). Of the 90,353 patients, 1,179 were found to have sepsis (1.30%), 349 were identified as
56
57
58
59
60

having severe sepsis without shock (0.39%), and 614 were determined to have septic shock (0.68%). The in-hospital mortality rate was 1.42%. Patient encounters spanned a variety of wards. The most common units represented in our study were perioperative care, the emergency department, the neurosciences department, and cardiovascular and thoracic transitional care. In the MIMIC-III data set, approximately 44% of patients were women and 56% were men. Stays were typically shorter in this data set, since each encounter included only an ICU stay. The median length of stay was 2 days. Furthermore, due to the nature of intensive care, there was a higher prevalence of sepsis (1.91%), severe sepsis (2.82%), and septic shock (4.36%). A full summary of baseline characteristics for both data sets is presented in Table 1.

Table 1: Demographic and clinical characteristics for UCSF patient population analyzed (N=90,353) and MIMIC-III patient population analyzed (N=21,604).

Demographic Overview	Characteristic	UCSF		MIMIC-III	
		Count	Percentage	Count	Percentage
Gender	Female	49,763	55.08%	9,499	43.97%
	Male	40,590	44.92%	12,105	56.03%
Age UCSF: median 55, IQR (38-67) MIMIC-III: median 65, IQR (53-77)	18-29	10,652	11.79%	978	4.53%
	30-39	14,202	15.72%	1,114	5.16%
	40-49	11,888	13.16%	2,112	9.78%
	50-59	16,856	18.66%	3,880	17.96%
	60-69	19,056	21.09%	4,906	22.71%
	70+	17,699	19.59%	8,614	39.87%
Length of Stay	0-2	28,258	31.26%	11,054	51.17%

(days) UCSF: median 4, IQR (2-6) MIMIC-III: median 2, IQR (2-4)	3-5	35,128	38.88%	7,004	32.42%
	6-8	12,664	14.02%	1,673	7.74%
	9-11	4,934	5.46%	734	3.40%
	12+	9,369	10.37%	1,139	5.27%
Death During Hospital Stay	Yes	1,279	1.42%	1,328	6.15%
	No	89,074	98.58%	20,276	93.85%
ICD-9 Code	Sepsis	1,179	1.30%	413	1.91%
	Severe Sepsis	349	0.39%	609	2.82%
	Septic Shock	614	0.68%	943	4.36%

Feature Construction

We minimally processed raw vital sign data to generate features. Following EHR data extraction and imputation as described above, we obtained three hourly values for each of the six vital sign measurement channels from that hour, the hour prior, and two hours prior. We also calculated two difference values between the current hour and the prior hour, and between the prior hour and the hour before that. We concatenated these five values from each vital sign into a causal feature vector x with 30 elements (five values from each of six measurement channels).

Machine Learning

We used gradient tree boosting to construct our classifier. Gradient tree boosting is an ensemble technique which combines the results from multiple weak decision trees in an iterative fashion. Each decision tree, was built by discretizing features into two categories. For example, one node of the decision tree might have stratified a patient based on whether their respiratory

1
2
3 rate was greater than 20 breaths per minute, or not. Depending on the answer for a given patient,
4 a second, third, etc., vital sign may be checked. A risk score was generated for the patient based
5 on their path along the decision tree. We limited each tree to split no more than six times; no
6 more than 1000 trees were aggregated in the iteration through gradient boosting to generate a
7 robust risk score. Training was performed separately for each distinct task and prediction win-
8 dow, and observations were accordingly labeled positive for model fitting for each specific pre-
9 diction task. Patient measurements were not used after the onset of a positive clinical indication.

10
11 We performed ten-fold cross validation to validate *InSight's* performance and minimize
12 potential model overfit. We randomly split the UCSF data set into a training set, comprised of
13 80% of UCSF's encounters, and an independent test set with the remaining 20% of encounters.
14 Of the training set, data were divided into ten groups, nine of which were used to train *InSight*,
15 and one of which was used to test. After cycling through all combinations of train and test set,
16 we then tested each of the ten models on the independent test set. Mean performance metrics
17 were calculated based on these ten models.

18
19 Additionally, we trained and validated *InSight's* performance in identifying sepsis, severe
20 sepsis, and septic shock after removing all features which were used in our gold standard defini-
21 tions for each condition. This resulted in the removal of vital sign SIRS criteria measurements
22 for sepsis and severe sepsis predictions, and the removal of systolic and diastolic blood pressure
23 measurements for septic shock. We also trained and validated the algorithm for each of the three
24 gold standards for randomly selected, up- and down-sampled subpopulations with positive class
25 prevalence between zero and one hundred percent.

26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 **Missing Data**

43
44 After assessing *InSight's* performance on complete data sets, we used a random deletion
45 process to simulate the algorithm's robustness to missing measurements. Individual measure-
46 ments from the test set were deleted according to a probability of deletion, P . We set $P = \{0, 0.1,$
47 $0.2, 0.4, \text{ and } 0.6\}$ for each of our missing data experiments and tested the *InSight* algorithm on
48 the sparse data sets.

49 50 51 52 53 54 55 **Transfer Learning**

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

To evaluate *InSight's* performance on a minimal amount of UCSF data, we used a transfer learning approach [20]. There are clear dissimilarities in patient demographics, clinical characteristics, and average measurement frequencies between the UCSF and MIMIC-III data sets (see Table 1). Partially this is because the UCSF data involves a variety of hospital wards, whereas the MIMIC-III data set provides only measurements taken in the ICU. We sought to determine improved performance metrics on the UCSF target data set, when the algorithm is primarily trained on MIMIC-III. Using MIMIC-III data as the source, and UCSF as the target, we trained the *InSight* classifier according to the severe sepsis gold standard. Variable amounts of UCSF training data were incrementally added to the MIMIC-III training data set, and the resulting model was then validated on the separate UCSF test data set. Specifically, we left 50% of the UCSF patients as test data, and we randomly selected different fractions of the remaining UCSF data and combined them with the entire MIMIC-III data set as the training data. For each fraction used, we trained 100 models with different random relative weights on the UCSF and MIMIC-III training data. Then, the mean and standard deviation of AUROC values for each of these models were calculated on 20 randomly sampled sets, and the model with highest mean AUROC value among these 100 was used.

Results

InSight's performance with respect to MEWS, SOFA, and SIRS is summarized in Figures 2A-C. Figures 2A, 2B, and 2C demonstrate *InSight's* ability to accurately detect the onset of sepsis and severe sepsis, and to accurately predict septic shock four hours prior to onset, compared to the performance of common sepsis scoring systems. Each figure presents *InSight's* receiver operating characteristic (ROC) curve together with the ROC curves for MEWS, SOFA, and SIRS. *InSight* achieves an area under the receiver operating characteristic (AUROC) curve for sepsis onset of 0.92 (95% confidence interval (CI), 0.90 - 0.93), for severe sepsis onset of 0.87 (95% CI 0.86 - 0.88), and for septic shock of 0.99 (95% CI 0.9991 - 0.9994); compared to SIRS, which demonstrates an AUROC of 0.75, 0.72, and 0.84, respectively. Even when all gold standard involved measurements were removed from model training, *InSight* continued to demonstrate improved accuracy over SIRS, MEWS, and SOFA, with AUROC values of 0.84 (95% CI 0.83-0.85) for sepsis onset, 0.80 (95% CI 0.79-0.81) for severe sepsis onset, and 0.96 (95% CI 0.96-0.97) for septic shock onset.

Comparing *InSight*'s performance across the three sepsis-related gold standards, it is clear that the septic shock criteria are relatively less challenging to anticipate, as its four hour prediction metrics are stronger than those for the detection of both sepsis and severe sepsis. Accordingly, we display the four hour prior to onset prediction case for septic shock (Fig 2C), where existing tools fail to adequately meet prediction standards relevant for sound clinical use. Four hours in advance of septic shock onset, *InSight* achieved an AUROC of 0.96 (95% CI 0.94 - 0.98). The resulting confusion matrix from the ten-fold cross validation of *InSight* can be found in Supplementary Tables 1 and 2.

Additional comparison metrics at time of detection for each gold standard are available in Table 2. In order to compare the specificities from each gold standard, we fixed sensitivities near 0.80; that is, we fixed a point on the ROC curve (i.e. set a specific threshold) after model development and tested algorithm performance under the chosen conditions in order to present data as consistently as possible. We similarly fixed specificities near 0.80 in order to compare sensitivities. Across all gold standards, a sensitivity of 0.80 results in a high specificity for *InSight*; however, the sensitivities for MEWS, SOFA, and SIRS are significantly lower. Notably, at 0.80 sensitivity, *InSight* achieves a specificity of 0.95 for sepsis, 0.84 for severe sepsis, and 0.99 for septic shock detection.

Table 2: Performance metrics for three sepsis gold standards at time of onset (zero hour), with sensitivities fixed at or near 0.80 in the first instance, and specificities fixed at or near 0.80 in the second instance.

	Gold Standard	<i>InSight</i> (95% CI)	<i>InSight</i> , label definitions removed (95% CI)	MEWS	SOFA	SIRS
AUROC	Sepsis	0.92 (0.90, 0.93)	0.84 (0.83, 0.85)	0.76	0.63	0.75
	Severe Sepsis	0.87 (0.86, 0.88)	0.80 (0.79, 0.81)	0.77	0.65	0.72

	Septic Shock	0.9992 (0.9991, 0.9994)	0.963 (0.959, 0.968)	0.94	0.86	0.82
Sensitivity	Sepsis	0.98 (0.96, 1.00)	0.99 (0.97, 1.00)	0.98	0.82	0.82
(Specificity fixed near 0.80)	Severe Sepsis	0.996 (0.989, 1.000)	1.00 (1.00, 1.00)	0.98	0.90	0.81
	Septic Shock	1.00 (1.00, 1.00)	0.994 (0.992, 0.997)	1.00	0.99	0.91
Specificity	Sepsis	0.95 (0.93, 0.97)	0.75 (0.73, 0.77)	0.72	0.32	0.51
(Sensitivity fixed near 0.80)	Severe Sepsis	0.85 (0.84, 0.86)	0.68 (0.62, 0.75)	0.72	0.37	0.50
	Septic Shock	0.9990 (0.9987, 0.9993)	0.95 (0.94, 0.96)	0.91	0.58	0.49

In addition to *InSight's* ability to detect sepsis, severe sepsis, and septic shock, Figure 3A illustrates the ROC of severe sepsis detection and prediction four hours prior to severe sepsis onset. Even four hours in advance, the *InSight* severe sepsis AUROC is 0.85 (95% CI 0.79 - 0.91), which is significantly higher than the onset time SIRS result of 0.75 AUROC. Figure 3B summarizes *InSight's* predictive advantage, using the severe sepsis gold standard, over MEWS, SOFA, and SIRS at the same time points in the hours leading up to onset. *InSight* maintains a high AUROC in the continuum up to four hours preceding severe sepsis onset. *InSight's* predictions four hours in advance produce a sensitivity and specificity that are greater than the at-onset time sensitivity and specificity of each MEWS, SOFA, and SIRS (Table 2, Fig. 3B).

We ranked feature importance for the classifiers developed in this experiment, and determined that systolic blood pressure at the time of prediction was consistently the most im-

portant feature in making accurate model predictions. The relative importance of other features varied significantly based on the specific prediction task.

In our second set of experiments, we validated *InSight's* performance in the presence of missing data. We tested *InSight's* ability to detect severe sepsis at time of onset with various rates of data dropout. Table 3 presents the results of these experiments. After randomly deleting data from the test set with a probability of 0.10, *InSight's* AUROC for severe sepsis detection is 0.82. Dropping approximately 60% of the test set measurements results in an AUROC of 0.75, demonstrating *InSight's* robustness to missing data. Of note, the AUROC of *InSight* at 60% data dropout achieves slightly better performance than SIRS with no missing data. Further, our experiments on applying *InSight* to up- and down-sampled sets showed that AUROC was largest when the set was chosen such that around half the patients met the gold standard. Moving lower on prevalence from 50% down to 0%, the AUROC values were only slightly lower while they dropped steeply when moving higher on prevalence from 50% up to 100% (a clinically unrealistic range).

Table 3: *InSight's* severe sepsis screening performance at time of onset in the presence of data sparsity, compared to SIRS with a full data complement.

	<i>InSight</i>					SIRS
% Data Missing	0%	10%	20%	40%	60%	0%
AUROC	0.90	0.82	0.79	0.76	0.75	0.72
Sensitivity	0.80	0.80	0.80	0.80	0.80	0.80
Specificity	0.84	0.66	0.57	0.50	0.49	0.51

Transfer Learning

InSight is flexible by design, and can be easily trained on an appropriate retrospective data set before being applied to a new patient population. However, sufficient historical patient data

1
2
3 is not always available for training on the target population. We evaluated *InSight*'s performance
4 when trained on a mixture of the MIMIC-III data together with increasing amounts of UCSF
5 training data, and then tested on a separate hold-out UCSF patient population using transfer
6 learning. In Figure 4, we show that the performance of the algorithm improves as the fraction of
7 UCSF target population data used in training increases.
8
9

10
11
12 Feature importance was quite stable across transfer learning experiments, with systolic
13 blood pressure measurements consistently playing an important role. Systolic blood pressure at
14 two hours before onset, at time of onset, and at one hour before onset, in that order, were the
15 most important features for accurate prediction in all tasks. Heart rate and diastolic blood pres-
16 sure at time of onset were consistently the fourth and fifth most important features, though order
17 of importance of the two features varied between tasks.
18
19
20
21
22
23

24 Discussion

25
26
27 We have validated the machine learning algorithm, *InSight*, on the mixed-ward data of
28 UCSF, which includes patients from the ED and floor units as well as the ICU, with varying
29 types and frequencies of patient measurements. *InSight* outperformed commonly-used disease
30 severity scores such as SIRS, MEWS, and SOFA for the screening of sepsis, severe sepsis, and
31 septic shock (Figure 2). These results, shown in Table 2, confirm *InSight*'s strength in predicting
32 these sepsis-related gold standard outcomes. To the authors' knowledge, *InSight* is first sepsis
33 screening system to meet or exceed an AUROC of 0.90 using only vital sign inputs, on each of
34 the sepsis gold standards evaluated in this study. Additionally, *InSight* provides predictive capa-
35 bilities in advance of sepsis onset, aided by the analysis of trends and correlations between vital
36 sign measurements. This advantage is apparent in the comparison with SIRS made in Figure 3A.
37 Up to four hours prior to severe sepsis onset, *InSight* maintains a high AUROC above 0.85 (Fig-
38 ure 3). This advance warning of patients trending toward severe sepsis could extend the window
39 for meaningful clinical intervention.
40
41
42
43
44
45
46
47
48
49

50
51 *InSight* uses only six common vital signs derived from a patient's EHR to detect sepsis
52 onset, as well as to predict those patients most at risk for developing sepsis. The decreased per-
53 formance of *InSight* for recognition of severe sepsis relative to sepsis onset may be in part be-
54 cause the organ failure characteristic of severe sepsis is more easily recognizable through labora-
55 tory tests for organ function. Because we have not incorporated metabolic function panels in this
56
57
58
59
60

1
2
3 validation of *InSight*, the detection of organ failure using only six common vital signs may be
4 more difficult. In practice, *InSight* is adaptable to different inputs and is able to incorporate la-
5 boratory results as they become available. Inclusion of these results may well increase the per-
6 formance of *InSight* for the detection and prediction of severe sepsis. However, in this work we
7 have chosen to benchmark the performance of *InSight* using only six commonly measured vital
8 signs. The ordering of metabolic panel laboratory tests are often predicated on clinician suspi-
9 cion of severe sepsis, and therefore, early or developing cases may be missed. Additionally, be-
10 cause these vital sign inputs do not require time-dependent laboratory results or additional manu-
11 al data entry, surveillance by *InSight* is frequent, and as a result, sepsis conditions are detected in
12 a more timely manner. Minimal data requirements also lighten the burden of implementation in a
13 clinical setting and broaden the potential clinical applications of *InSight*.

14
15 Although *InSight* uses only a handful of clinical variables, it maintains a high level of
16 performance in experiments with randomly missing data. We demonstrate in Table 3 that for the
17 detection of severe sepsis, even with up to 60% of randomized test patient data missing, *InSight*
18 still achieves slightly better performance to SIRS calculated with complete data availability.

19
20 Additionally, we have investigated the customizability of *InSight* to local hospital de-
21 mographics and measurements. The incorporation of site-specific data into the training set using
22 transfer learning improves performance on test sets, over that of a training set comprised entirely
23 of an independent population. This indicates that it may be possible to adequately train *InSight*
24 for use in a new clinical setting, while still predominantly using existing retrospective data from
25 other institutions. Further, the results of our up- and down-sampling experiments indicate that
26 *InSight* is likely to only be slightly less effective (in AUROC terms) in settings with lower preva-
27 lence of sepsis, severe sepsis or septic shock, than UCSF or slightly more effective if the preva-
28 lence is higher than UCSF.

29
30 Our previous studies, performed on earlier versions of the model, have investigated *In-*
31 *Sight* applied to individual sepsis standards such as the SIRS standard for sepsis [21], severe sep-
32 sis [22], and septic shock [23], on the MIMIC retrospective datasets. We have also developed a
33 related algorithm to detect patient stability [24] and predict mortality [23, 25]. However, this
34 study, which evaluates a significantly improved algorithm, is the first to apply *InSight* to all three
35 standard sepsis definitions simultaneously, and to validate the algorithm on a mixed ward popu-
36 lation, including ED, ICU and floor wards from UCSF. This study is also the first to use only six
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 minimal vital signs, without utilizing a mental status evaluation such as Glasgow Coma Score, or
4 even age, in the detection and prediction of those sepsis standards. Additionally, this study
5 demonstrates the adaptability of the machine learning algorithm to an entirely new patient data
6 set with markedly different demographics and outcomes.
7
8
9

10 11 12 **Limitations**

13
14 While we incorporated data from both UCSF and MIMIC-III, we cannot claim generali-
15 zability of our results to other populations on the basis of this study alone. However, we are aid-
16 ed by the minimality of data used to make predictions; because *InSight* requires only six of the
17 most basic and widely-available clinical measurements, it is likely that it will perform similarly
18 in other settings if vital sign data is available. The gold standard references we use to determine
19 sepsis, severe sepsis and septic shock rely on ICD-9 codes from the hospital database; this stand-
20 ard potentially limits our ability to capture all septic patients in the dataset, should any have been
21 undiagnosed or improperly recorded. The administrative coding procedures may vary by hospi-
22 tal and do not always precisely reproduce results from manual chart review for sepsis diagnosis,
23 although ICD-9 codes have been previously validated for accuracy in the detection of severe sep-
24 sis [26]. The vital sign measurements abstracted from the EHR are basic measurements routinely
25 collected from all patients regardless of diagnosis and independent of physician judgement, and
26 therefore this input to *InSight* is not dependent on the time of clinical diagnosis. However, the
27 ordering of laboratory tests is contingent on physician suspicion, and the timing of these inputs
28 may reflect clinician judgement rather than true onset time, potentially limiting the accuracy of
29 our analysis.
30
31
32
33
34
35
36
37
38
39
40
41
42

43 It is important to note that we designed the study as a classification task rather than a
44 time-to-event modeling experiment, because the former is significantly more common in the lit-
45 erature [27-30]. The alternative would not allow for the use of an established, standard set of per-
46 formance metrics such as AUROC and specificity without custom modification, and would make
47 it more difficult to compare the present study to prior work in the field. This study was conduct-
48 ed retrospectively, and so we are unable to make claims regarding performance in a prospective
49 setting, which involves the interpretation and use of *InSight*'s predictions by clinicians. Addi-
50 tionally, our inclusion criteria requiring at least seven hours of patient data preceding sepsis on-
51 set also limits generalizability to a clinical setting, where the predictor would receive data in real
52
53
54
55
56
57
58
59
60

1
2
3 time and not based on these criteria. Finally, our random deletion of data is not necessarily repre-
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

time and not based on these criteria. Finally, our random deletion of data is not necessarily representative of data scarcity as it would occur in clinical settings where the rate of missing measurements would depend on the standard rate of data collection, which can vary widely, especially between the emergency department, general ward, and intensive care units. We intend to evaluate these algorithms in prospective clinical studies in future work.

Conclusions

We have validated the machine learning algorithm, *InSight*, in a multicenter study in a mixed-ward population from UCSF and an ICU population from BIDMC. *InSight* provides high sensitivity and specificity for the detection and prediction of sepsis, severe sepsis, and septic shock using the analysis of only six common vital signs taken from the electronic health record. *InSight* outperforms scoring systems in current use for the detection of sepsis, is robust to a significant amount of missing patient data, and can be customized to novel sites using a limited amount of site-specific data. Our results indicate that *InSight* outperforms tools currently used for sepsis detection and prediction, which may lead to improvements in sepsis-related patient outcomes.

Acknowledgments: We acknowledge the assistance of Siddharth Gampa and Emily Huynh for editing contributions. We thank Dr. Hamid Mohamadlou and Dr. Thomas Desautels for contributions to the development of the machine learning algorithm *InSight*.

Contributorship Statement: QM, JC, and RD conceived the described experiments. DS acquired the UCSF data. QM and YZ executed the experiments. QM, RD, JC, and MJ interpreted the results. QM, MJ, and JH wrote the manuscript. QM, RD, MJ, JH, JC, CB, DS, LS, UC, GF, and YK revised the manuscript, with assistance from Emily Huynh and Siddharth Gampa. All authors approved the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing Interests: All authors who have affiliations listed with Dascena (Hayward, CA, USA) are employees or contractors of Dascena. Dr. Barton reports receiving consulting fees

1
2
3 from Dascena. Dr. Barton, Dr. Shieh, Dr. Shimabukuro and Dr. Fletcher report receiving grant
4 funding from Dascena.
5

6
7 **Funding:** Research reported in this publication was supported by the National Science Founda-
8 tion under Grant No. 1549867. The content is solely the responsibility of the authors and does
9 not necessarily represent the official views of the National Science Foundation. The funder had
10 no role in the conduct of the study; collection, management, analysis, and interpretation of data;
11 preparation, review, and approval of the manuscript; and decision to submit the manuscript for
12 publication.
13

14
15
16
17 **Data Sharing:** No data obtained from UCSF in this study can be shared or made available for
18 open access. MIMIC-III is a publicly available database. Please visit <https://mimic.physionet.org/>
19 for information on using the MIMIC-III database.
20
21
22
23

24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

1. Murphy SL, Xu J, Kochanek KD. Deaths: final data for 2010. National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System. 2013;61: 1-17.
2. Angus, Derek C et al. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med. 2001;29: 1303-1310.
3. Stevenson, Elizabeth K et al. Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis. Crit Care Med. 2014;42: 625.
4. Pfunter A, Wier LM, Steiner C. Costs for Hospital Stays in the United States, 2010: Statistical Brief #146. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2006 Feb-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK121966/>
5. O'Brien, J. The Cost of Sepsis. CDC Safe Healthcare Blog. 2015. Available from: <https://blogs.cdc.gov/safehealthcare/the-cost-of-sepsis/#ref>
6. Gaieski DF, Edwards JM, Kallan MJ, Carr BG. Benchmarking the incidence and mortality of severe sepsis in the United States. Crit Care Med. 2013;41: 1167-1174.
7. Rivers, Emanuel et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. New Engl J Med. 2001;345: 1368-1377.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
8. Nguyen, H Bryant et al. Implementation of a bundle of quality indicators for the early management of severe sepsis and septic shock is associated with decreased mortality. *Crit Care Med.* 2007;35: 1105-1112.
9. Kumar, Anand et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med.* 2006;34: 1589-1596.
10. Levy, Mitchell M et al. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Intensive Care Med.* 2003;29: 530-538.
11. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315: 801-10.
12. Subbe CP, Slater A, Menon D, Gemmell L. Validation of physiological scoring systems in the accident and emergency department. *Emerg Med J.* 2006;23:841-845.
13. Rangel-Frausto MS, Pittet D, Costigan M, Hwang T, Davis CS, Wenzel RP. The natural history of the systemic inflammatory response syndrome (SIRS): a prospective study. *JAMA.* 1995;273:117-123.
14. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med.* 1996;22: 707-710.
15. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data.* 2016. doi: 10.1038/sdata.2016.35.
16. PostgreSQL Global Development Group, <https://www.postgresql.org/>
17. G. Van Rossum. The Python Language Reference Manual. Network Theory Ltd. Python Software Foundation. 2003. Available from: <https://www.python.org/>
18. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med.* 2015;7:299ra122. doi: 10.1126/scitranslmed.aab3719.
19. Fullerton JN, Price CL, Silvey NE, Brace SJ, Perkins GD. Is the Modified Early Warning Score (MEWS) superior to clinician judgement in detecting critical illness in the pre-hospital environment? *Resuscitation.* 2012;83: 557-562.

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
20. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. *Mach Learn*. 2010;79: 151-175.
21. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. *Comp Biol Med*. 2016;74: 69-73.
22. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform*. 2016;4: e28.
23. Calvert J, Desautels T, Chettipally U, Barton C, Hoffman J, Jay M, Mao Q, Mohamdlou H, Das R. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Ann Med Surg*. 2016;8: 50-55.
24. Calvert JS, Price DA, Barton CW, Chettipally UK, Das R. Discharge recommendation based on a novel technique of homeostatic analysis. *J Am Med Inform Assoc*. 2016;24: 24-29.
25. Calvert J, Mao Q, Hoffman JL, Jay M, Desautels T, Mohamdlou H, et al. Using electronic health record collected clinical variables to predict medical intensive care unit mortality. *Ann Med Surg*. 2016;11: 52-57.
26. Iwashyna TJ, Odden A, Rohde J, Bonham C, Kuhn L, Malani P, et al. Identifying patients with severe sepsis using administrative claims: patient-level validation of the angus implementation of the international consensus conference definition of severe sepsis. *Med Care*. 2014;52:e39.
27. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *J Am Med Inform Assoc*. 2013; 315-325. doi: 10.1136/amiajnl-2013-001815
28. Brause R, Hamker F, Paetz J. Septic shock diagnosis by neural networks and rule based systems. In: Schmitt M, Teodorescu HN, Jain A, et al., editors. *Computational intelligence techniques in medical diagnosis and prognosis*. New York: Springer, 2002; 323-356.
29. Thiel SW, Rosini JM, Shannon W, Doherty JA, Micek ST, Kollef MH, Early Prediction of Septic Shock. *J. Hosp. Med* 2010;1;19-25. doi:10.1002/jhm.530

- 1
2
3 30. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an au-
4 tomated trigger for sepsis clinical decision support at emergency department triage using
5 machine learning. PLOS ONE 2017; 12(4): e0174708 Doi:
6
7 10.1371/journal.pone.0174708
8
9

10
11 **Figure 1.** Patient inclusion flow diagram for the UCSF data set.
12

13
14
15 **Figure 2.** ROC curves for InSight and common scoring systems at time of (A) sepsis onset, (B)
16 severe sepsis onset, and (C) four hours before septic shock onset.
17
18

19
20
21 **Figure 3.** A) ROC detection (zero hour, blue) and prediction (four hour prior to onset, red)
22 curves using InSight and ROC detection (zero hour, green) curve for SIRS, with the severe sep-
23 sis gold standard. B) Predictive performance of InSight and comparators, using the severe sepsis
24 gold standard, as a function of time prior to onset.
25
26
27

28
29
30 **Figure 4.** Learning curves (mean AUROC on the UCSF target data set) with increasing number
31 of target training examples. Error bars represent the standard deviation. When data availability of
32 the target set is low, target-only training exhibits lower AUROC values and high variability.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

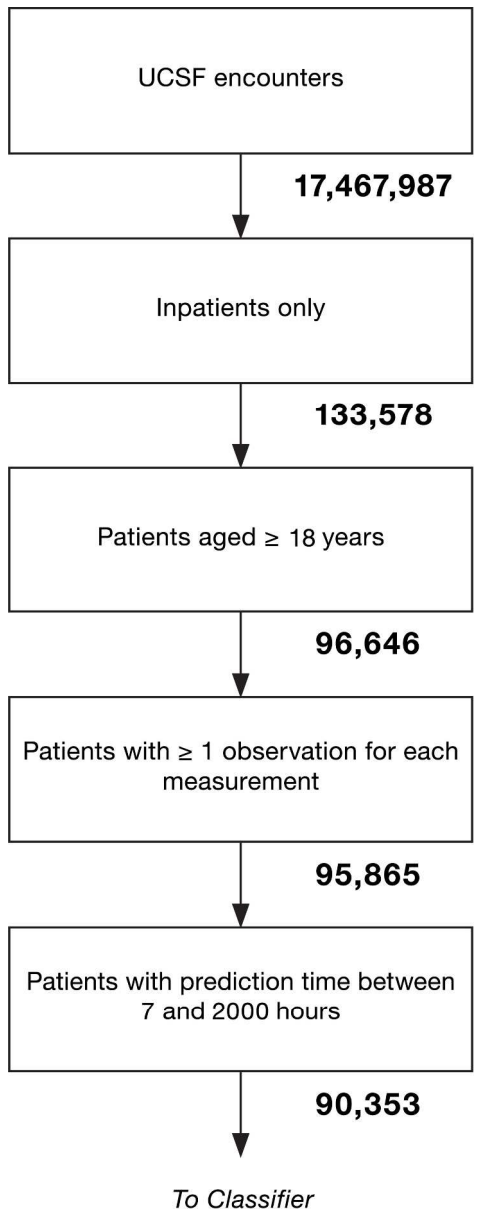


Figure 1. Patient inclusion flow diagram for the UCSF data set.

109x282mm (300 x 300 DPI)

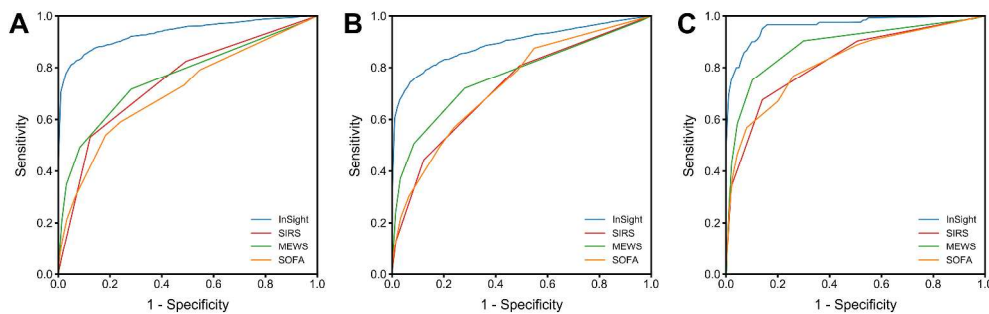


Figure 2. ROC curves for InSight and common scoring systems at time of (A) sepsis onset, (B) severe sepsis onset, and (C) four hours before septic shock onset.

609x203mm (300 x 300 DPI)

Peer review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

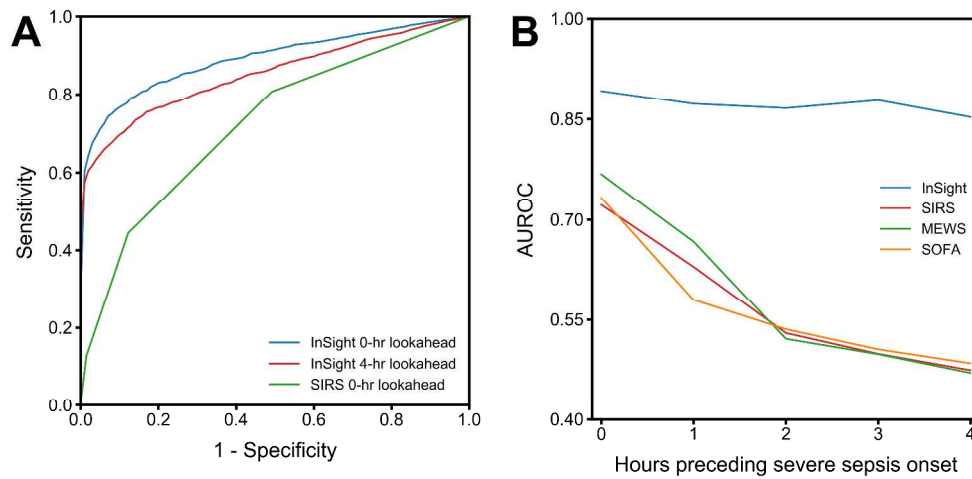


Figure 3. A) ROC detection (zero hour, blue) and prediction (four hour prior to onset, red) curves using InSight and ROC detection (zero hour, green) curve for SIRS, with the severe sepsis gold standard. B) Predictive performance of InSight and comparators, using the severe sepsis gold standard, as a function of time prior to onset.

406x203mm (300 x 300 DPI)

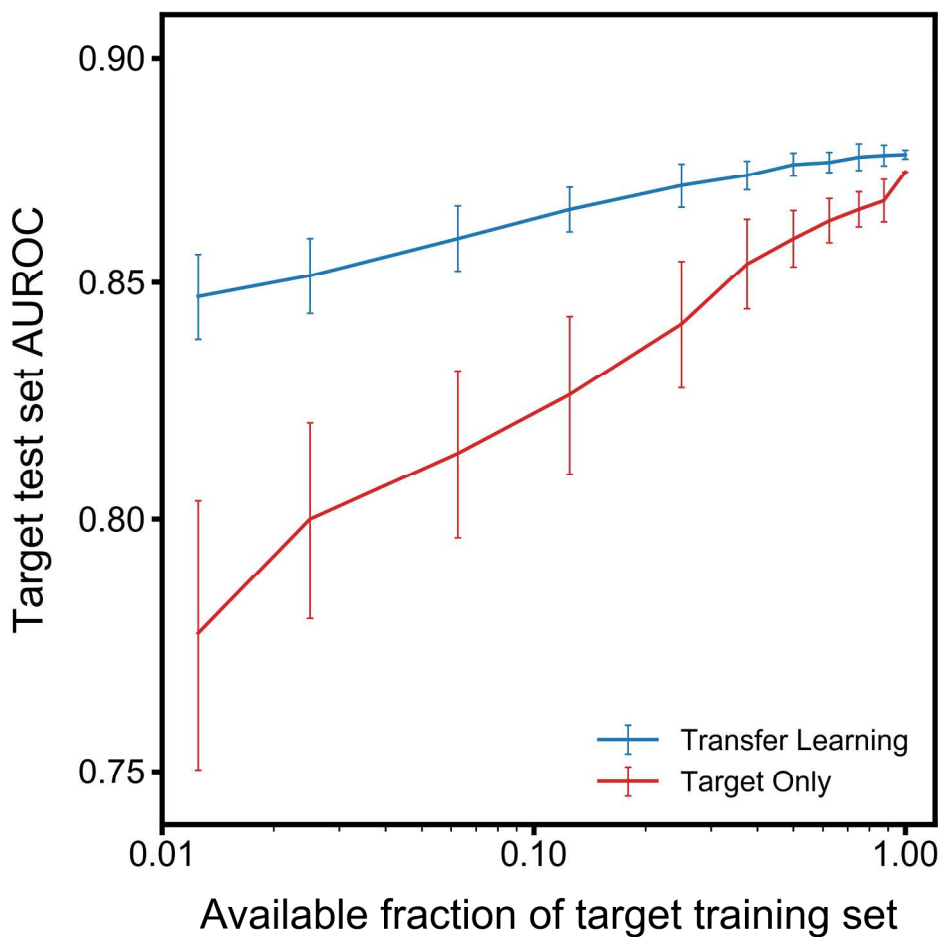


Figure 4. Learning curves (mean AUROC on the UCSF target data set) with increasing number of target training examples. Error bars represent the standard deviation. When data availability of the target set is low, target-only training exhibits lower AUROC values and high variability.!! †

203x203mm (300 x 300 DPI)



Supplemental Table 1. Confusion matrix of ten-fold cross validation results, using all features.

Values in the table represent averages \pm standard deviations.

Sepsis:

	Predicted Positive	Predicted Negative
Actual Positive	17109 \pm 424	774 \pm 424
Actual Negative	30 \pm 2	110 \pm 2

Severe Sepsis:

	Predicted Positive	Predicted Negative
Actual Positive	15322 \pm 369	2531 \pm 369
Actual Negative	48 \pm 4	171 \pm 4

Septic Shock:

	Predicted Positive	Predicted Negative
Actual Positive	18801 \pm 9	19 \pm 9
Actual Negative	67 \pm 3	265 \pm 3

Supplemental Table 2. Confusion matrix of ten-fold cross validation results, with gold standard definition associated inputs removed. Values in the table represent averages \pm standard deviations.

Sepsis:

	Predicted Positive	Predicted Negative
Actual Positive	13655 \pm 488	4231 \pm 488
Actual Negative	34 \pm 3	120 \pm 3

Severe Sepsis:

	Predicted Positive	Predicted Negative
Actual Positive	13855 \pm 2381	4015 \pm 2381
Actual Negative	72 \pm 25	147 \pm 25

Septic Shock:

	Predicted Positive	Predicted Negative
Actual Positive	17972 \pm 320	878 \pm 320
Actual Negative	66 \pm 0	260 \pm 0

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.
Introduction			
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.
Methods			
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.
	5b	D;V	Describe eligibility criteria for participants.
	5c	D;V	Give details of treatments received, if relevant.
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	D;V	Explain how the study size was arrived at.
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10c	V	For validation, describe how the predictions were calculated.
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.
Risk groups	11	D;V	Provide details on how risk groups were created, if done.
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.
Results			
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).
Model development	14a	D	Specify the number of participants and outcome events in each analysis.
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	D	Explain how to use the prediction model.
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).
Discussion			
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.
Other information			
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

BMJ Open

Multicenter validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2017-017833.R2
Article Type:	Research
Date Submitted by the Author:	17-Oct-2017
Complete List of Authors:	Mao, Qingqing; Dascena, Research Jay, Melissa; Dascena, Research hoffman, jana; Dascena, Research Calvert, Jake; Dascena, Research Barton, Christopher; University of California San Francisco, Emergency Medicine Shimabukuro, David; University of California San Francisco, Anesthesia and Perioperative Care Shieh, Lisa; Stanford University School of Medicine, Medicine Chettipally, Uli ; University of California San Francisco, Emergency Medicine; Kaiser Permanente South San Francisco Medical Center Fletcher, Grant; University of Washington School of Medicine Kerem, Yaniv; Stanford University School of Medicine Zhou, Yifan; University of California Berkeley, Statistics Das, Ritankar; Dascena, Research
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Diagnostics, Infectious diseases, Intensive care, Emergency medicine
Keywords:	sepsis, septic shock, clinical decision support, prediction, machine learning, electronic health records

SCHOLARONE™
Manuscripts

Multicenter validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU

Qingqing Mao¹, Melissa Jay¹, Jana L. Hoffman^{1*}, Jacob Calvert¹, Christopher Barton², David Shimabukuro³, Lisa Shieh⁴, Uli Chettipally^{2,5}, Grant Fletcher⁶, Yaniv Kerem^{7,8}, Yifan Zhou^{1,9}
Ritankar Das¹

¹ Dascena Inc., Hayward, CA

² Department of Emergency Medicine, University of California San Francisco, San Francisco, CA, United States

³ Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, CA, United States

⁴ Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

⁵ Kaiser Permanente South San Francisco Medical Center, South San Francisco, CA, United States

⁶ Division of Internal Medicine, University of Washington School of Medicine, Seattle, WA, United States

⁷ Department of Clinical Informatics, Stanford University School of Medicine, Stanford, CA, United States

⁸ Department of Emergency Medicine, Kaiser Permanente Redwood City Medical Center, Redwood City, CA, United States

⁹ Department of Statistics, University of California Berkeley, Berkeley, CA, United States

* Corresponding author

Email: Jana@Dascena.com

22710 Foothill Blvd., Suite #2

Hayward, CA 94541

Keywords: sepsis, septic shock, clinical decision support, prediction, machine learning, electronic health records

Word count: 5,340

Abstract

Objectives: We validate a machine learning-based sepsis prediction algorithm (*InSight*) for detection and prediction of three sepsis-related gold standards, using only six vital signs. We evaluate robustness to missing data, customization to site-specific data using transfer learning, and generalizability to new settings.

Design: A machine learning algorithm with gradient tree boosting. Features for prediction were created from combinations of only six vital sign measurements and their changes over time.

Setting: A mixed-ward retrospective data set from the University of California, San Francisco (UCSF) Medical Center (San Francisco, CA) as the primary source, an intensive care unit data set from the Beth Israel Deaconess Medical Center (Boston, MA) as a transfer learning source, and four additional institutions' datasets to evaluate generalizability.

Participants: 684,443 total encounters, with 90,353 encounters from June 2011 to March 2016 at UCSF.

Interventions: none

Primary and secondary outcome measures: Area under the receiver operating characteristic curve (AUROC) for detection and prediction of sepsis, severe sepsis, and septic shock.

Results: For detection of sepsis and severe sepsis, *InSight* achieves an area under the receiver operating characteristic (AUROC) curve of 0.92 (95% CI 0.90 - 0.93) and 0.87 (95% CI 0.86 - 0.88), respectively. Four hours before onset, *InSight* predicts septic shock with an AUROC of 0.96 (95% CI 0.94 - 0.98), and severe sepsis with an AUROC of 0.85 (95% CI 0.79 - 0.91).

Conclusions: *InSight* outperforms existing sepsis scoring systems in identifying and predicting sepsis, severe sepsis, and septic shock. This is the first sepsis screening system to exceed an AUROC of 0.90 using only vital sign inputs. *InSight* is robust to missing data, can be customized to novel hospital data using a small fraction of site data, and retained strong discrimination across all institutions.

Strengths and limitations of this study

- Machine learning is applied to the detection and prediction of three separate sepsis standards in the emergency department, general ward and intensive care settings.
- Only six commonly measured vital signs are used as input for the algorithm.

- The algorithm is robust to randomly missing data.
- Transfer learning successfully leverages large dataset information to a target dataset.
- Retrospective nature of the study does not predict clinician reaction to information.

Introduction

Sepsis is a major health crisis and one of the leading causes of death in the United States [1]. Approximately 750,000 hospitalized patients are diagnosed with severe sepsis in the United States annually, with an estimated mortality rate of up to one-third [2,3]. The cost burden of sepsis is disproportionately high, with estimated costs of \$20.3 billion dollars annually, or \$55.6 million per day in US hospitals [4]. Additionally, the average hospital stay for sepsis is twice as expensive as other conditions [5], and the average incidence of severe sepsis is increasing by approximately 13% per year [6]. Early diagnosis and treatment have been shown to reduce mortality and associated costs [7-9]. Despite clear benefits, early and accurate sepsis detection remains a difficult clinical problem.

Sepsis has been defined as a dysregulated host response to infection. In practice, sepsis can be challenging to recognize because of the heterogeneity of the host response to infection, and the diversity of possible infectious insult. Sepsis has been traditionally recognized as two or more Systemic Inflammatory Response Syndrome (SIRS) [10] criteria together with a known or suspected infection; progressing to severe sepsis, in the event of organ dysfunction; and finally to septic shock, which additionally includes refractory hypotension [10]. However, ongoing debates over sepsis definitions and clinical criteria, as evidenced by the recent proposed redefinitions of sepsis [11], underscore a fundamental difficulty in the identification and accurate diagnosis of sepsis.

Various rule-based disease severity scoring systems are widely used in hospitals in an attempt to identify septic patients. These scores, such as the Modified Early Warning Score (MEWS) [12], the Systemic Inflammatory Response Syndrome (SIRS) criteria [13], and the Sequential Organ Failure Assessment (SOFA) [14], are manually tabulated at the bedside and lack accuracy in sepsis diagnosis. However, the increasing prevalence of Electronic Health Records (EHR) in clinical settings provides an opportunity for enhanced patient monitoring and increased early detection of sepsis.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

This study validates a machine learning algorithm *InSight*, which uses only six vital signs taken directly from the EHR, in the detection and prediction of sepsis, severe sepsis, and septic shock in a mixed-ward population at the University of California, San Francisco (UCSF). We investigate the effects of induced data sparsity on *InSight* performance, and compare all results with other scores that are commonly used in the clinical setting for the detection and prediction of sepsis. We additionally train and test the algorithm for severe sepsis detection on data from Stanford Medical Center and three community hospitals in order to better estimate its expected clinical performance. Furthermore, we apply a transfer learning scheme to customize a Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III-trained algorithm to the UCSF patient population using a minimal amount of UCSF-specific data.

Methods

Data sets

We used a data set provided by the UCSF Medical Center representing patient stays from June 2011 to March 2016 in all experiments. The UCSF data set contains 17,467,987 hospital encounters, including inpatient and outpatient visits to all units within the UCSF medical system. The data were de-identified to comply with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. For transfer learning, we used the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III v1.3 data set, compiled from the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA between 2001 and 2012, composed of 61,532 ICU stays [15]. This database is a publicly available database constructed by researchers at MIT's Laboratory for Computational Physiology, and the data were also de-identified in compliance with HIPAA. Additionally, we trained and tested the algorithm for severe sepsis detection on data from Stanford Medical Center (Stanford, CA), Oroville Hospital (Oroville, CA), Bakersfield Heart Hospital (BHH; Bakersfield, CA), and Cape Regional Medical Center (CRMC; Cape May Courthouse, NJ). Details on these datasets are included in the Supplementary Materials (Supplementary Tables 1 and 2). Data collection for all datasets did not impact patient safety. Therefore, this study constitutes non-human subjects research, which does not require Institutional Review Board approval.

Data Extraction and Imputation

The data were provided in the form of comma separated value (CSV) files and stored in a PostgreSQL [16] database. Custom SQL queries were written to extract measurements and patient outcomes of interest. The measurement files were then binned by hour for each patient. To be included, patients were required to have at least one of each type of measurement recorded during the encounter. If a patient did not have a measurement in a given hour, the missing measurement was filled in using carry-forward imputation. This imputation method applied the patient's last measured value to the following hour (a causal procedure). In the case of multiple measurements within an hour, the mean was calculated and used in place of an individual measurement. After the data were processed and imputed in Python [17], they were used to train the *InSight* classifier and test its predictions at sepsis onset and at fixed time points prior to onset.

Gold Standards

In this study, we tested *InSight*'s performance according to various gold standards (clinical indications). We investigated *InSight*'s ability to predict and detect sepsis, severe sepsis, and septic shock. Further, we compared *InSight*'s performance to SIRS, MEWS, and SOFA, for each of the following gold standards. For training and testing the algorithm, we conservatively identified each septic condition by requiring that the ICD-9 code corresponding to the diagnosis was coded for each positive case, in addition to meeting the clinical requirements for the definition of each septic standard as defined below.

Sepsis

The sepsis gold standard was determined using the 2001 consensus sepsis definition [10]: “the presence of two or more SIRS criteria paired with a suspicion of infection.” To identify a case as positive for sepsis, we required ICD-9 code 995.91. The onset time was defined as the first time two or more SIRS criteria were met within the same hour. SIRS criteria are defined as:

- heart rate > 90 beats/ min,
- body temperature > 38 °C or < 36 °C,
- respiratory rate >20 breaths/min or PaCO₂ < 32 mmHg, and

- white blood cell count $> 12,000$ cells/ μL or $< 4,000$ cells/ μL . [10]

Severe Sepsis

The severe sepsis gold standard used the definition of severe sepsis as “organ dysfunction caused by sepsis” which can be represented by one or more of the criteria below, and identified for patients with the severe sepsis ICD-9 code 995.92. We assigned the severe sepsis onset time to be the first instance during which two SIRS criteria as described above and one of the following organ dysfunction criteria were met within the same hour.

- Lactate > 2 mmol/L
- Systolic blood pressure < 90 mmHg
- Urine output < 0.5 mL/kg, over two hours, prior to organ dysfunction after fluid resuscitation
- Creatinine > 2 mg/dL without renal insufficiency or chronic dialysis
- Bilirubin > 2 mg/dL without having liver disease or cirrhosis
- Platelet count $< 100,000$ μL
- International normalized ratio > 1.5
- PaO₂/FiO₂ < 200 in addition to pneumonia
 < 250 with acute kidney injury but without pneumonia

Septic Shock

We identified as positive cases for septic shock those patients who received the septic shock ICD-9 code 785.52 and additionally demonstrated the following conditions:

- systolic blood pressure of < 90 mmHg, defined as hypotension, for at least 30 minutes, and
- who were resuscitated with ≥ 20 ml/kg over a 24 hour period, or
- who received ≥ 1200 ml in total fluids. [18]

The onset time was defined as the first hour when either the hypotension or fluid resuscitation criterion was met.

Calculating Comparators

1
2
3 We compared *InSight* predictions for each gold standard to three common patient deterioration
4 scoring systems: SIRS, SOFA, and MEWS. Area under the receiver operating characteristic
5 (AUROC) curve, sensitivity, and specificity were compared across all prediction models. The
6 SIRS criteria, as explained in the sepsis definition, were evaluated independently of the suspicion
7 of infection. To calculate the SOFA score, we collected each patient's PaO₂/FiO₂, Glasgow
8 Coma Score, mean arterial blood pressure or administration of vasopressors, bilirubin level,
9 platelet counts, and creatinine level. Each of the listed measurements is associated with a SOFA
10 score of 1-4, based on severity level, as described by Vincent et al. [14]. After receiving a score
11 for each of the six organ dysfunction categories, the overall SOFA score was computed as the
12 sum of the category scores and used as a comparator to *InSight*. Finally, the MEWS score, which
13 ranges from 0 (normal) to 14 (high risk of deterioration), was determined by tabulating subscores
14 for heart rate, systolic blood pressure, respiratory rate, temperature, and Glasgow Coma Score.
15 We used the subscore system presented in Fullerton et al. [19] to compute each patient's
16 MEWS score.
17
18
19
20
21
22
23
24
25
26
27
28
29

30 **Measurements and Patient Inclusion**

31
32 In order to generate *InSight* scores, patient data were analyzed from each of the following
33 six clinical vital sign measurements: systolic blood pressure, diastolic blood pressure, heart rate,
34 respiratory rate, peripheral capillary oxygen saturation (SpO₂), and temperature. We used only
35 vital signs, which are frequently available and routinely taken in the ICU, ED, and floor units.
36 Patient data were used from the course of a patient's hospital encounter, regardless of the unit the
37 patient was in when the data were collected.
38
39
40
41
42

43 All patients over the age of 18 were considered for this study. For a given encounter, if
44 the patient was admitted to the hospital from the ED, the start of the ED visit is where the
45 analysis began. Patients in our final data sets were required to have at least one measurement for
46 each of the six vital signs. In order to ensure enough data to accurately characterize sepsis
47 predictions at four hours pre-onset, we further limited the study group to exclude patients whose
48 septic condition onset time was within seven hours after the start of their record, which was
49 either the time of admission to the hospital or the start of their ED visit; the latter was applicable
50 only if the patient was admitted through the ED. A smaller window to sepsis onset time would
51 have resulted in insufficient testing data to make 4-hour prediction possible in some cases, which
52
53
54
55
56
57
58
59
60

would inappropriately affect performance metrics such as sensitivity and specificity. Patients with sepsis onset after 2,000 hours post-admission were also excluded, to limit the data analysis matrix size. The final UCSF data set included 90,353 patients (Fig.1) and the MIMIC-III data set contained 21,604 patients, following the same inclusion criteria. Inclusion criteria and final inclusion numbers for the Stanford, Oroville, BHH, and CRMC datasets are included in Supplementary Table 1.

After patient exclusion, our final group of UCSF patients was composed of 55% women and 45% men with a median age of 55. The median hospital length of stay was 4 days, IQR = (2,6). Of the 90,353 patients, 1,179 were found to have sepsis (1.30%), 349 were identified as having severe sepsis without shock (0.39%), and 614 were determined to have septic shock (0.68%). The in-hospital mortality rate was 1.42%. Patient encounters spanned a variety of wards. The most common units represented in our study were perioperative care, the emergency department, the neurosciences department, and cardiovascular and thoracic transitional care. In the MIMIC-III data set, approximately 44% of patients were women and 56% were men. Stays were typically shorter in this data set, since each encounter included only an ICU stay. The median length of stay was 2 days. Furthermore, due to the nature of intensive care, there was a higher prevalence of sepsis (1.91%), severe sepsis (2.82%), and septic shock (4.36%). A full summary of baseline characteristics for both data sets is presented in Table 1. Full demographic information for the Stanford, Oroville, BHH, and CRMC datasets is provided in Supplementary Table 2.

Table 1: Demographic and clinical characteristics for UCSF patient population analyzed (N=90,353) and MIMIC-III patient population analyzed (N=21,604).

Demographic Overview	Characteristic	UCSF		MIMIC-III	
		Count	Percentage	Count	Percentage
Gender	Female	49,763	55.08%	9,499	43.97%
	Male	40,590	44.92%	12,105	56.03%
	18-29	10,652	11.79%	978	4.53%

Age UCSF: median 55, IQR (38-67) MIMIC-III: median 65, IQR (53-77)	30-39	14,202	15.72%	1,114	5.16%
	40-49	11,888	13.16%	2,112	9.78%
	50-59	16,856	18.66%	3,880	17.96%
	60-69	19,056	21.09%	4,906	22.71%
	70+	17,699	19.59%	8,614	39.87%
Length of Stay (days) UCSF: median 4, IQR (2-6) MIMIC-III: median 2, IQR (2-4)	0-2	28,258	31.26%	11,054	51.17%
	3-5	35,128	38.88%	7,004	32.42%
	6-8	12,664	14.02%	1,673	7.74%
	9-11	4,934	5.46%	734	3.40%
	12+	9,369	10.37%	1,139	5.27%
Death During Hospital Stay	Yes	1,279	1.42%	1,328	6.15%
	No	89,074	98.58%	20,276	93.85%
ICD-9 Code	Sepsis	1,179	1.30%	413	1.91%
	Severe Sepsis	349	0.39%	609	2.82%
	Septic Shock	614	0.68%	943	4.36%

Feature Construction

We minimally processed raw vital sign data to generate features. Following EHR data extraction and imputation as described above, we obtained three hourly values for each of the six vital sign measurement channels from that hour, the hour prior, and two hours prior. We also calculated two difference values between the current hour and the prior hour, and between the prior hour and the hour before that. We concatenated these five values from each vital sign into a causal feature vector x with 30 elements (five values from each of six measurement channels).

Machine Learning

We used gradient tree boosting to construct our classifier. Gradient tree boosting is an ensemble technique which combines the results from multiple weak decision trees in an iterative fashion. Each decision tree, was built by discretizing features into two categories. For example, one node of the decision tree might have stratified a patient based on whether their respiratory rate was greater than 20 breaths per minute, or not. Depending on the answer for a given patient, a second, third, etc., vital sign may be checked. A risk score was generated for the patient based on their path along the decision tree. We limited each tree to split no more than six times; no more than 1000 trees were aggregated in the iteration through gradient boosting to generate a robust risk score. Training was performed separately for each distinct task and prediction window, and observations were accordingly labeled positive for model fitting for each specific prediction task. Patient measurements were not used after the onset of a positive clinical indication.

We performed ten-fold cross validation to validate *InSight's* performance and minimize potential model overfit. We randomly split the UCSF data set into a training set, comprised of 80% of UCSF's encounters, and an independent test set with the remaining 20% of encounters. Of the training set, data were divided into ten groups, nine of which were used to train *InSight*, and one of which was used to test. After cycling through all combinations of train and test set, we then tested each of the ten models on the independent test set. Mean performance metrics were calculated based on these ten models. For severe sepsis detection at time of onset on each of Stanford, Oroville, BHH, and CRMC datasets, we performed four-fold cross validation of the model.

Additionally, we trained and validated *InSight's* performance in identifying sepsis, severe sepsis, and septic shock after removing all features which were used in our gold standard definitions for each condition. This resulted in the removal of vital sign SIRS criteria measurements for sepsis and severe sepsis predictions, and the removal of systolic and diastolic blood pressure measurements for septic shock. We also trained and validated the algorithm for each of the three gold standards for randomly selected, up- and down-sampled subpopulations with positive class prevalence between zero and one hundred percent.

Missing Data

After assessing *InSight's* performance on complete data sets, we used a random deletion process to simulate the algorithm's robustness to missing measurements. Individual measurements from the test set were deleted according to a probability of deletion, P . We set $P = \{0, 0.1, 0.2, 0.4, \text{ and } 0.6\}$ for each of our missing data experiments and tested the *InSight* algorithm on the sparse data sets.

Transfer Learning

To evaluate *InSight's* performance on a minimal amount of UCSF data, we used a transfer learning approach [20]. There are clear dissimilarities in patient demographics, clinical characteristics, and average measurement frequencies between the UCSF and MIMIC-III data sets (see Table 1). Partially this is because the UCSF data involves a variety of hospital wards, whereas the MIMIC-III data set provides only measurements taken in the ICU. We sought to determine improved performance metrics on the UCSF target data set, when the algorithm is primarily trained on MIMIC-III. Using MIMIC-III data as the source, and UCSF as the target, we trained the *InSight* classifier according to the severe sepsis gold standard. Variable amounts of UCSF training data were incrementally added to the MIMIC-III training data set, and the resulting model was then validated on the separate UCSF test data set. Specifically, we left 50% of the UCSF patients as test data, and we randomly selected different fractions of the remaining UCSF data and combined them with the entire MIMIC-III data set as the training data. For each fraction used, we trained 100 models with different random relative weights on the UCSF and MIMIC-III training data. Then, the mean and standard deviation of AUROC values for each of these models were calculated on 20 randomly sampled sets, and the model with highest mean AUROC value among these 100 was used.

Results

InSight's performance with respect to MEWS, SOFA, and SIRS is summarized in Figures 2A-C. Figures 2A, 2B, and 2C demonstrate *InSight's* ability to accurately detect the onset of sepsis and severe sepsis, and to accurately predict septic shock four hours prior to onset, compared to the performance of common sepsis scoring systems. Each figure presents *InSight's*

1
2
3 receiver operating characteristic (ROC) curve together with the ROC curves for MEWS, SOFA,
4 and SIRS. *InSight* achieves an area under the receiver operating characteristic (AUROC) curve
5 for sepsis onset of 0.92 (95% confidence interval (CI), 0.90 - 0.93), for severe sepsis onset of
6 0.87 (95% CI 0.86 - 0.88), and for septic shock of 0.99 (95% CI 0.9991 - 0.9994); compared to
7 SIRS, which demonstrates an AUROC of 0.75, 0.72, and 0.84, respectively. Even when all gold
8 standard involved measurements were removed from model training, *InSight* continued to
9 demonstrate improved accuracy over SIRS, MEWS, and SOFA, with AUROC values of 0.84
10 (95% CI 0.83-0.85) for sepsis onset, 0.80 (95% CI 0.79-0.81) for severe sepsis onset, and 0.96
11 (95% CI 0.96-0.97) for septic shock onset.

12
13 Comparing *InSight*'s performance across the three sepsis-related gold standards, it is
14 clear that the septic shock criteria are relatively less challenging to anticipate, as its four hour
15 prediction metrics are stronger than those for the detection of both sepsis and severe sepsis.
16 Accordingly, we display the four hour prior to onset prediction case for septic shock (Fig 2C),
17 where existing tools fail to adequately meet prediction standards relevant for sound clinical use.
18 Four hours in advance of septic shock onset, *InSight* achieved an AUROC of 0.96 (95% CI 0.94 -
19 0.98). The resulting confusion matrix from the ten-fold cross validation of *InSight* can be found
20 in Supplementary Tables 3 and 4.

21
22 Additional comparison metrics at time of detection for each gold standard are available in
23 Table 2. In order to compare the specificities from each gold standard, we fixed sensitivities near
24 0.80; that is, we fixed a point on the ROC curve (i.e. set a specific threshold) after model
25 development and tested algorithm performance under the chosen conditions in order to present
26 data as consistently as possible. We similarly fixed specificities near 0.80 in order to compare
27 sensitivities. Across all gold standards, a sensitivity of 0.80 results in a high specificity for
28 *InSight*; however, the sensitivities for MEWS, SOFA, and SIRS are significantly lower. Notably,
29 at 0.80 sensitivity, *InSight* achieves a specificity of 0.95 for sepsis, 0.84 for severe sepsis, and
30 0.99 for septic shock detection.

31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51 **Table 2:** Performance metrics for three sepsis gold standards at time of onset (zero hour), with
52 sensitivities fixed at or near 0.80 in the first instance, and specificities fixed at or near 0.80 in the
53 second instance.
54
55
56
57
58
59
60

	Gold Standard	<i>InSight</i> (95% CI)	<i>InSight</i> , label definitions removed (95% CI)	MEWS	SOFA	SIRS
AUROC	Sepsis	0.92 (0.90, 0.93)	0.84 (0.83, 0.85)	0.76	0.63	0.75
	Severe Sepsis	0.87 (0.86, 0.88)	0.80 (0.79, 0.81)	0.77	0.65	0.72
	Septic Shock	0.9992 (0.9991, 0.9994)	0.963 (0.959, 0.968)	0.94	0.86	0.82
Sensitivity <i>(Specificity fixed near 0.80)</i>	Sepsis	0.98 (0.96, 1.00)	0.99 (0.97, 1.00)	0.98	0.82	0.82
	Severe Sepsis	0.996 (0.989, 1.000)	1.00 (1.00, 1.00)	0.98	0.90	0.81
	Septic Shock	1.00 (1.00, 1.00)	0.994 (0.992, 0.997)	1.00	0.99	0.91
Specificity <i>(Sensitivity fixed near 0.80)</i>	Sepsis	0.95 (0.93, 0.97)	0.75 (0.73, 0.77)	0.72	0.32	0.51
	Severe Sepsis	0.85 (0.84, 0.86)	0.68 (0.62, 0.75)	0.72	0.37	0.50
	Septic Shock	0.9990 (0.9987, 0.9993)	0.95 (0.94, 0.96)	0.91	0.58	0.49

In addition to *InSight*'s ability to detect sepsis, severe sepsis, and septic shock, Figure 3A illustrates the ROC of severe sepsis detection and prediction four hours prior to severe sepsis onset. Even four hours in advance, the *InSight* severe sepsis AUROC is 0.85 (95% CI 0.79 - 0.91), which is significantly higher than the onset time SIRS result of 0.75 AUROC. Figure 3B summarizes *InSight*'s predictive advantage, using the severe sepsis gold standard, over MEWS, SOFA, and SIRS at the same time points in the hours leading up to onset. *InSight* maintains a

high AUROC in the continuum up to four hours preceding severe sepsis onset. *InSight's* predictions four hours in advance produce a sensitivity and specificity that are greater than the at-onset time sensitivity and specificity of each MEWS, SOFA, and SIRS (Table 2, Fig. 3B). In order to determine the generalizability of the algorithm to different settings, we tested *InSight* on additional patient data sets from four distinct hospitals. For severe sepsis detection at time of onset, *InSight* achieved AUROC over 0.92 on patients from Stanford, Oroville Hospital, Bakersfield Heart Health, and Cape Regional Medical Center (Table 3). ROC curves and comparisons to alternate sepsis classification systems on these datasets are presented in the data supplement (Supplementary Tables 5-8, Supplementary Figures 1 and 2). *InSight* AUROC values exceed those of the MEWS, SIRS, qSOFA and SOFA scores on the same datasets for severe sepsis detection at time of onset.

Table 3. Algorithm performance for severe sepsis detection at time of onset. LR= Likelihood ratio.

	Stanford	Oroville	BHH
AUROC (95% CI)	0.924 (0.9202,0.9278)	0.983 (0.9804, 0.9856)	0.945 (0.921,0.969)
Sensitivity	0.798	0.806	0.875
Specitivity	0.901	0.989	0.940
Accuracy	0.900	0.971	0.963
LR+	8.253	77.92	58.94
LR-	0.224	0.197	0.129

We ranked feature importance for the classifiers developed in this experiment, and determined that systolic blood pressure at the time of prediction was consistently the most important feature in making accurate model predictions. The relative importance of other features varied significantly based on the specific prediction task.

In our second set of experiments, we validated *InSight*'s performance in the presence of missing data. We tested *InSight*'s ability to detect severe sepsis at time of onset with various rates of data dropout. Table 4 presents the results of these experiments. After randomly deleting data from the test set with a probability of 0.10, *InSight*'s AUROC for severe sepsis detection is 0.82. Dropping approximately 60% of the test set measurements results in an AUROC of 0.75, demonstrating *InSight*'s robustness to missing data. Of note, the AUROC of *InSight* at 60% data dropout achieves slightly better performance than SIRS with no missing data. Further, our experiments on applying *InSight* to up- and down-sampled sets showed that AUROC was largest when the set was chosen such that around half the patients met the gold standard. Moving lower on prevalence from 50% down to 0%, the AUROC values were only slightly lower while they dropped steeply when moving higher on prevalence from 50% up to 100% (a clinically unrealistic range).

Table 4: *InSight*'s severe sepsis screening performance at time of onset in the presence of data sparsity, compared to SIRS with a full data complement.

	<i>InSight</i>					SIRS
% Data Missing	0%	10%	20%	40%	60%	0%
AUROC	0.90	0.82	0.79	0.76	0.75	0.72
Sensitivity	0.80	0.80	0.80	0.80	0.80	0.80
Specificity	0.84	0.66	0.57	0.50	0.49	0.51

Transfer Learning

InSight is flexible by design, and can be easily trained on an appropriate retrospective data set before being applied to a new patient population. However, sufficient historical patient data is not always available for training on the target population. We evaluated *InSight*'s performance when trained on a mixture of the MIMIC-III data together with increasing amounts of UCSF training data, and then tested on a separate hold-out UCSF patient population using

1
2
3 transfer learning. In Figure 4, we show that the performance of the algorithm improves as the
4 fraction of UCSF target population data used in training increases.
5
6

7 Feature importance was quite stable across transfer learning experiments, with systolic
8 blood pressure measurements consistently playing an important role. Systolic blood pressure at
9 two hours before onset, at time of onset, and at one hour before onset, in that order, were the
10 most important features for accurate prediction in all tasks. Heart rate and diastolic blood
11 pressure at time of onset were consistently the fourth and fifth most important features, though
12 order of importance of the two features varied between tasks.
13
14
15
16
17

18 19 Discussion

20 We have validated the machine learning algorithm, *InSight*, on the mixed-ward data of
21 UCSF, which includes patients from the ED and floor units as well as the ICU, with varying
22 types and frequencies of patient measurements. *InSight* outperformed commonly-used disease
23 severity scores such as SIRS, MEWS, and SOFA for the screening of sepsis, severe sepsis, and
24 septic shock (Figure 2). These results, shown in Table 2, confirm *InSight*'s strength in predicting
25 these sepsis-related gold standard outcomes. The algorithm's strong performance across the
26 academic and community hospital data used in this study suggests potential strong performance
27 in a variety of future clinical settings.
28
29
30
31
32
33
34
35

36 To the authors' knowledge, *InSight* is first sepsis screening system to meet or exceed an
37 AUROC of 0.90 using only vital sign inputs, on each of the sepsis gold standards evaluated in
38 this study. Additionally, *InSight* provides predictive capabilities in advance of sepsis onset, aided
39 by the analysis of trends and correlations between vital sign measurements. This advantage is
40 apparent in the comparison with SIRS made in Figure 3A. Up to four hours prior to severe sepsis
41 onset, *InSight* maintains a high AUROC above 0.85 (Figure 3). This advance warning of patients
42 trending toward severe sepsis could extend the window for meaningful clinical intervention.
43
44
45
46
47

48 *InSight* uses only six common vital signs derived from a patient's EHR to detect sepsis
49 onset, as well as to predict those patients most at risk for developing sepsis. The decreased
50 performance of *InSight* for recognition of severe sepsis relative to sepsis onset may be in part
51 because the organ failure characteristic of severe sepsis is more easily recognizable through
52 laboratory tests for organ function. Because we have not incorporated metabolic function panels
53 in this validation of *InSight*, the detection of organ failure using only six common vital signs may
54
55
56
57
58
59
60

1
2
3 be more difficult. In practice, *InSight* is adaptable to different inputs and is able to incorporate
4 laboratory results as they become available. Inclusion of these results may well increase the
5 performance of *InSight* for the detection and prediction of severe sepsis. However, in this work
6 we have chosen to benchmark the performance of *InSight* using only six commonly measured
7 vital signs. The ordering of metabolic panel laboratory tests are often predicated on clinician
8 suspicion of severe sepsis, and therefore, early or developing cases may be missed. Additionally,
9 because these vital sign inputs do not require time-dependent laboratory results or additional
10 manual data entry, surveillance by *InSight* is frequent, and as a result, sepsis conditions are
11 detected in a more timely manner. Minimal data requirements also lighten the burden of
12 implementation in a clinical setting and broaden the potential clinical applications of *InSight*.

13
14
15
16
17
18
19
20
21 Although *InSight* uses only a handful of clinical variables, it maintains a high level of
22 performance in experiments with randomly missing data. We demonstrate in Table 4 that for the
23 detection of severe sepsis, even with up to 60% of randomized test patient data missing, *InSight*
24 still achieves slightly better performance to SIRS calculated with complete data availability.

25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44 Additionally, we have investigated the customizability of *InSight* to local hospital
45 demographics and measurements. The incorporation of site-specific data into the training set
46 using transfer learning improves performance on test sets, over that of a training set comprised
47 entirely of an independent population. This indicates that it may be possible to adequately train
48 *InSight* for use in a new clinical setting, while still predominantly using existing retrospective
49 data from other institutions. Further, the results of our up- and down-sampling experiments
50 indicate that *InSight* is likely to only be slightly less effective (in AUROC terms) in settings with
51 lower prevalence of sepsis, severe sepsis or septic shock, than UCSF or slightly more effective if
52 the prevalence is higher than UCSF.

53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

Our previous studies, performed on earlier versions of the model, have investigated *InSight* applied to individual sepsis standards such as the SIRS standard for sepsis [21], severe sepsis [22], and septic shock [23], on the MIMIC retrospective datasets. We have also developed a related algorithm to detect patient stability [24] and predict mortality [25, 26]. However, this study, which evaluates a significantly improved algorithm, is the first to apply *InSight* to all three standard sepsis definitions simultaneously, and to validate the algorithm on a mixed ward population, including ED, ICU and floor wards from UCSF. This study is also the first to use

1
2
3 only six minimal vital signs, without utilizing a mental status evaluation such as Glasgow Coma
4 Score, or even age, in the detection and prediction of those sepsis standards.
5
6

7 The separate models trained for each gold standard and prediction window in this study
8 further demonstrate the potential clinical utility of machine learning methods. In addition to
9 training on a specific patient population, machine learning methods can allow for the
10 development of prediction models which are tailored to a hospital's unique needs, data
11 availability, and existing workflow practices. Any one of the models developed in this study
12 could be independently deployed in a clinical setting; choice of model deployment would be
13 contingent upon the needs of a particular hospital, and the expected tradeoff in performance for
14 different model choices. Additionally, this study demonstrates the adaptability of the machine
15 learning algorithm to an entirely new patient data set with markedly different demographics and
16 outcomes through both site-specific retraining and transfer learning techniques.
17
18
19
20
21
22
23
24
25
26
27

28 **Limitations**

29
30 While we incorporated data from multiple institutions, we cannot claim generalizability
31 of our results to other populations on the basis of this study alone. However, we are aided by the
32 minimality of data used to make predictions; because *InSight* requires only six of the most basic
33 and widely-available clinical measurements, it is likely that it will perform similarly in other
34 settings if vital sign data is available. The gold standard references we use to determine sepsis,
35 severe sepsis and septic shock rely on ICD-9 codes from the hospital database; this standard
36 potentially limits our ability to capture all septic patients in the dataset, should any have been
37 undiagnosed or improperly recorded. The administrative coding procedures may vary by
38 hospital and do not always precisely reproduce results from manual chart review for sepsis
39 diagnosis, although ICD-9 codes have been previously validated for accuracy in the detection of
40 severe sepsis [27]. The vital sign measurements abstracted from the EHR are basic
41 measurements routinely collected from all patients regardless of diagnosis and independent of
42 physician judgement, and therefore this input to *InSight* is not dependent on the time of clinical
43 diagnosis. However, the ordering of laboratory tests is contingent on physician suspicion, and
44 the timing of these inputs may reflect clinician judgement rather than true onset time, potentially
45 limiting the accuracy of our analysis.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 While the imputation and averaging performed before feature construction eliminated
4 some information about sampling frequency, these methods do not remove all non-physiological
5 information inherent to our system. Further, imputation of the most recently available past
6 measurement may artificially alter the rate of the temporal changes in patient vital signs that we
7 incorporate into feature vectors, which may in turn affect risk predictions. Averaging multiple
8 patient measurements may similarly remove informative variation in vital signs.
9

10
11 It is important to note that we designed the study as a classification task rather than a
12 time-to-event modeling experiment, because the former is significantly more common in the
13 literature [28-31]. The alternative would not allow for the use of an established, standard set of
14 performance metrics such as AUROC and specificity without custom modification, and would
15 make it more difficult to compare the present study to prior work in the field. This study was
16 conducted retrospectively, and so we are unable to make claims regarding performance in a
17 prospective setting, which involves the interpretation and use of *InSight's* predictions by
18 clinicians. Additionally, our inclusion criteria requiring at least seven hours of patient data
19 preceding sepsis onset also limits generalizability to a clinical setting where the predictor would
20 receive data in real time. Algorithm performance in a clinical setting may reasonably be expected
21 to be lower than its retrospective performance in this study. Finally, our random deletion of data
22 is not necessarily representative of data scarcity as it would occur in clinical settings where the
23 rate of missing measurements would depend on the standard rate of data collection, which can
24 vary widely, especially between the emergency department, general ward, and intensive care
25 units. We intend to evaluate these algorithms in prospective clinical studies in future work.
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42 **Conclusions**

43
44 We have validated the machine learning algorithm, *InSight*, in a multicenter study in a
45 mixed-ward population from UCSF and an ICU population from BIDMC. *InSight* provides high
46 sensitivity and specificity for the detection and prediction of sepsis, severe sepsis, and septic
47 shock using the analysis of only six common vital signs taken from the electronic health record.
48 *InSight* outperforms scoring systems in current use for the detection of sepsis, is robust to a
49 significant amount of missing patient data, and can be customized to novel sites using a limited
50 amount of site-specific data. Our results indicate that *InSight* outperforms tools currently used
51
52
53
54
55
56
57
58
59
60

1
2
3 for sepsis detection and prediction, which may lead to improvements in sepsis-related patient
4 outcomes.
5
6
7
8
9

10
11
12 **Acknowledgments:** We acknowledge the assistance of Siddharth Gampa, Anna Lynn-Palevsky
13 and Emily Huynh for editing contributions. We thank Dr. Hamid Mohamadlou and Dr. Thomas
14 Desautels for contributions to the development of the machine learning algorithm *InSight*. We
15 acknowledge Zirui Jiang for valuable computational assistance. We gratefully thank Matthew N.
16 Fine, MD, Dr. Andrea McCoy, and Chris Maupin, RN for access to patient data sets.
17
18
19

20
21
22 **Author Statement:** QM, JC, and RD conceived the described experiments. DS acquired the
23 UCSF data. QM and YZ executed the experiments. QM, RD, JC, and MJ interpreted the results.
24 QM, MJ, and JH wrote the manuscript. QM, RD, MJ, JH, JC, CB, DS, LS, UC, GF, and YK
25 revised the manuscript, with assistance from Emily Huynh and Siddharth Gampa. All authors
26 approved the version to be published and agree to be accountable for all aspects of the work in
27 ensuring that questions related to the accuracy or integrity of any part of the work are
28 appropriately investigated and resolved.
29
30
31
32
33
34
35
36

37 **Competing Interests:** All authors who have affiliations listed with Dascena (Hayward, CA,
38 USA) are employees or contractors of Dascena. Dr. Barton reports receiving consulting fees
39 from Dascena. Dr. Barton, Dr. Shieh, Dr. Shimabukuro and Dr. Fletcher report receiving grant
40 funding from Dascena.
41
42
43
44
45

46 **Funding:** Research reported in this publication was supported by the National Science
47 Foundation under Grant No. 1549867. The content is solely the responsibility of the authors and
48 does not necessarily represent the official views of the National Science Foundation. The funder
49 had no role in the conduct of the study; collection, management, analysis, and interpretation of
50 data; preparation, review, and approval of the manuscript; and decision to submit the manuscript
51 for publication.
52
53
54
55
56
57
58
59
60

Data Sharing: No data obtained from UCSF, Stanford, Oroville Hospital, Cape Regional Medical Center or Bakersfield Heart Hospital in this study can be shared or made available for open access. MIMIC-III is a publicly available database. Please visit <https://mimic.physionet.org/> for information on using the MIMIC-III database.

References

1. Murphy SL, Xu J, Kochanek KD. Deaths: final data for 2010. National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System. 2013;61: 1-17.
2. Angus, Derek C et al. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Crit Care Med.* 2001;29: 1303-1310.
3. Stevenson, Elizabeth K et al. Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis. *Crit Care Med.* 2014;42: 625.
4. Pfunter A, Wier LM, Steiner C. Costs for Hospital Stays in the United States, 2010: Statistical Brief #146. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2006 Feb-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK121966/>
5. O'Brien, J. The Cost of Sepsis. CDC Safe Healthcare Blog. 2015. Available from: <https://blogs.cdc.gov/safehealthcare/the-cost-of-sepsis/#ref>
6. Gaieski DF, Edwards JM, Kallan MJ, Carr BG. Benchmarking the incidence and mortality of severe sepsis in the United States. *Crit Care Med.* 2013;41: 1167-1174.
7. Rivers, Emanuel et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New Engl J Med.* 2001;345: 1368-1377.
8. Nguyen, H Bryant et al. Implementation of a bundle of quality indicators for the early management of severe sepsis and septic shock is associated with decreased mortality. *Crit Care Med.* 2007;35: 1105-1112.
9. Kumar, Anand et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med.* 2006;34: 1589-1596.

10. Levy, Mitchell M et al. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Intensive Care Med.* 2003;29: 530-538.
11. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315: 801-10.
12. Subbe CP, Slater A, Menon D, Gemmell L. Validation of physiological scoring systems in the accident and emergency department. *Emerg Med J.* 2006;23:841-845.
13. Rangel-Frausto MS, Pittet D, Costigan M, Hwang T, Davis CS, Wenzel RP. The natural history of the systemic inflammatory response syndrome (SIRS): a prospective study. *JAMA.* 1995;273:117-123.
14. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med.* 1996;22: 707-710.
15. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data.* 2016. doi: 10.1038/sdata.2016.35.
16. PostgreSQL Global Development Group, <https://www.postgresql.org/>
17. G. Van Rossum. *The Python Language Reference Manual.* Network Theory Ltd. Python Software Foundation. 2003. Available from: <https://www.python.org/>
18. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med.* 2015;7:299ra122. doi: 10.1126/scitranslmed.aab3719.
19. Fullerton JN, Price CL, Silvey NE, Brace SJ, Perkins GD. Is the Modified Early Warning Score (MEWS) superior to clinician judgement in detecting critical illness in the pre-hospital environment? *Resuscitation.* 2012;83: 557-562.
20. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. *Mach Learn.* 2010;79: 151-175.
21. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. *Comp Biol Med.* 2016;74: 69-73.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
22. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform.* 2016;4: e28.
23. Calvert J, Desautels T, Chettipally U, Barton C, Hoffman J, Jay M, Mao Q, Mohamadlou H, Das R. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Ann Med Surg.* 2016;8: 50-55.
24. Calvert JS, Price DA, Barton CW, Chettipally UK, Das R. Discharge recommendation based on a novel technique of homeostatic analysis. *J Am Med Inform Assoc.* 2016;24: 24-29.
25. Calvert J, Mao Q, Rogers AJ, Barton C, Jay M, Desautels T, Mohamadlou H, Jan J, Das R. A computational approach to mortality prediction of alcohol use disorder inpatients. *Computers in biology and medicine.* 2016 Aug 1;75:74-9.
26. Calvert J, Mao Q, Hoffman JL, Jay M, Desautels T, Mohamadlou H, et al. Using electronic health record collected clinical variables to predict medical intensive care unit mortality. *Ann Med Surg.* 2016;11: 52-57.
27. Iwashyna TJ, Odden A, Rohde J, Bonham C, Kuhn L, Malani P, et al. Identifying patients with severe sepsis using administrative claims: patient-level validation of the angus implementation of the international consensus conference definition of severe sepsis. *Med Care.* 2014;52:e39.
28. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *J Am Med Inform Assoc.* 2013; 315-325. doi: 10.1136/amiajnl-2013-001815
29. Brause R, Hamker F, Paetz J. Septic shock diagnosis by neural networks and rule based systems. In: Schmitt M, Teodorescu HN, Jain A, et al., editors. *Computational intelligence techniques in medical diagnosis and prognosis.* New York: Springer, 2002; 323-356.
30. Thiel SW, Rosini JM, Shannon W, Doherty JA, Micek ST, Kollef MH, Early Prediction of Septic Shock. *J. Hosp. Med* 2010;1;19-25. doi:10.1002/jhm.530
31. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage

1
2
3 using machine learning. PLOS ONE 2017; 12(4): e0174708 Doi:
4
5 10.1371/journal.pone.0174708
6
7
8
9

10 **Figure 1:** Patient inclusion flow diagram for the UCSF data set.

11
12
13
14 **Figure 2:** ROC curves for *InSight* and common scoring systems at time of (A) sepsis onset, (B)
15 severe sepsis onset, and (C) four hours before septic shock onset.
16
17

18
19 **Figure 3:** A) ROC detection (zero hour, blue) and prediction (four hour prior to onset, red)
20 curves using *InSight* and ROC detection (zero hour, green) curve for SIRS, with the severe sepsis
21 gold standard. B) Predictive performance of *InSight* and comparators, using the severe sepsis
22 gold standard, as a function of time prior to onset.
23
24
25
26
27

28 **Figure 4.** Learning curves (mean AUROC on the UCSF target data set) with increasing number
29 of target training examples. Error bars represent the standard deviation. When data availability of
30 the target set is low, target-only training exhibits lower AUROC values and high variability.
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

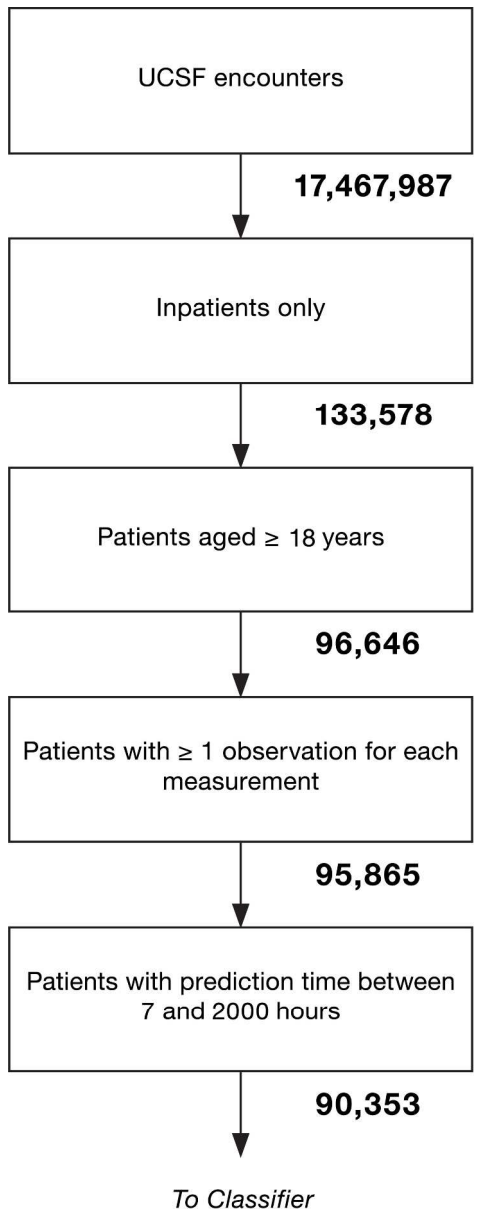


Figure 1. Patient inclusion flow diagram for the UCSF data set.

109x282mm (300 x 300 DPI)

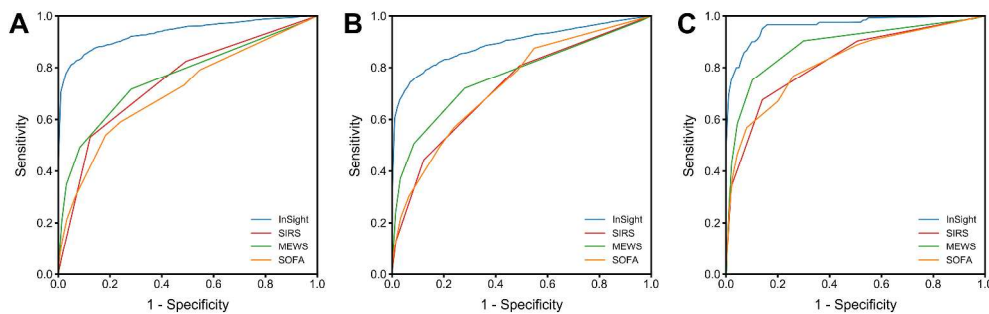


Figure 2. ROC curves for InSight and common scoring systems at time of (A) sepsis onset, (B) severe sepsis onset, and (C) four hours before septic shock onset.

609x203mm (300 x 300 DPI)

Peer review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

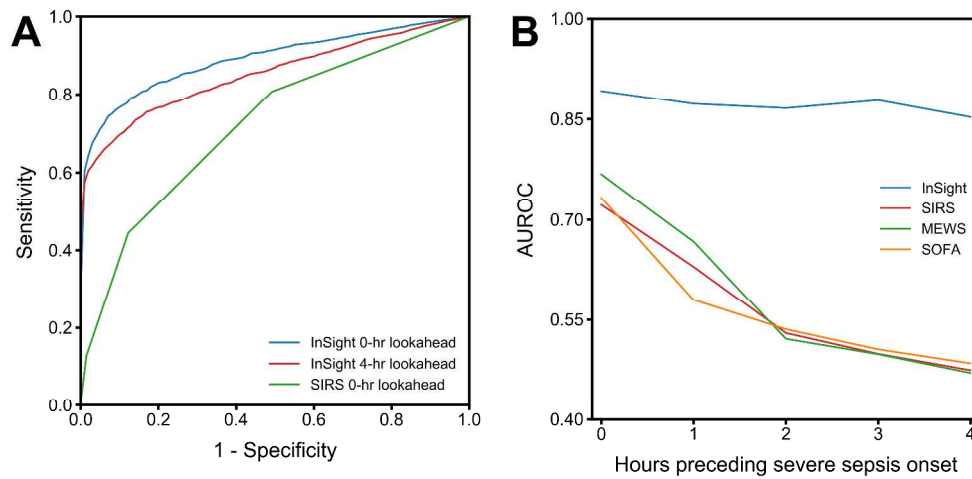


Figure 3. A) ROC detection (zero hour, blue) and prediction (four hour prior to onset, red) curves using InSight and ROC detection (zero hour, green) curve for SIRS, with the severe sepsis gold standard. B) Predictive performance of InSight and comparators, using the severe sepsis gold standard, as a function of time prior to onset.

406x203mm (300 x 300 DPI)

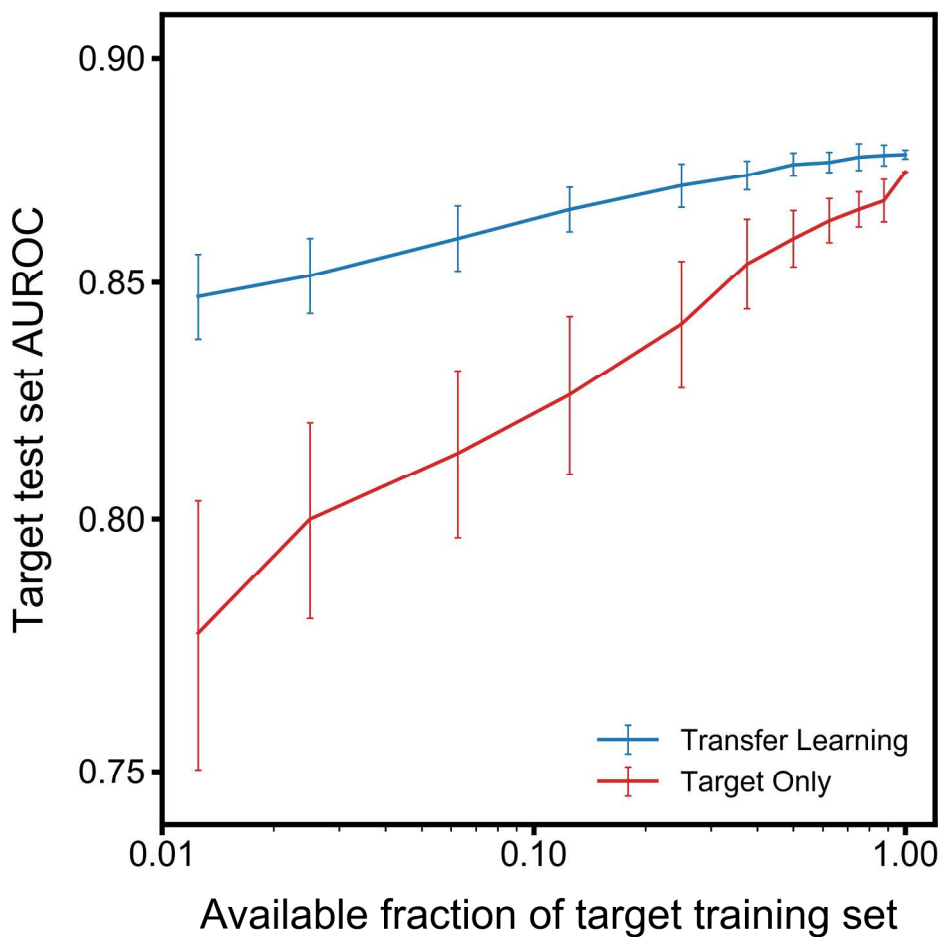


Figure 4. Learning curves (mean AUROC on the UCSF target data set) with increasing number of target training examples. Error bars represent the standard deviation. When data availability of the target set is low, target-only training exhibits lower AUROC values and high variability.!! †

203x203mm (300 x 300 DPI)



Supplementary Table 1. Inclusion flowcharts for Stanford Medical Center, Oroville Hospital, Bakersfield Heart Hospital, and Cape Regional Medical Center.

	Stanford	Oroville	BHH	CRMC
Total encounters	521,040	1,200	5,681	4,637
Inpatients Only	441,208	1,200	5,305	4,631
Patients Aged ≥ 18	358,017	1,150	5,272	4,510
Patients with ≥ 1 observation of each required measurement*	239,767	1,140	2,231	4,295
Patients with prediction time between 7 and 2000 hours	239,767	1,140	2,231	4,295

*required measurements include heart rate, respiratory rate, peripheral oxygen saturation (SpO₂), temperature, systolic blood pressure, and diastolic blood pressure.

Supplementary Table 2. Demographic information for Stanford Medical Center, Oroville Hospital, Bakersfield Heart Hospital, and Cape Regional Medical Center.

Demographic Overview	Characteristic	Stanford (%)	Oroville (%)	BHH (%)	CRMC (%)
Gender	Female	53.91	54.87	45.94	52.30
	Male	46.09	45.13	54.06	47.70
Age Median Ages Stanford: 53 UCSF: 55 BIDMC: 65 Oroville: 61 BHH: 60 CRMC: 68	18-29	16.75	7.74	8.54	3.93
	30-39	13.28	9.22	10.52	4.90
	40-49	14.50	10.43	12.79	8.28
	50-59	18.20	19.39	17.53	15.75
	60-69	17.71	20.96	18.47	20.55
	70+	19.56	32.26	32.15	46.60
Length of Stay (days)	0-2	71.51	97.33	63.42	21.42
	3-5	15.53	34.17	11.07	7.17
	6-8	5.53	6.667	4.80	2.37
	9+	7.43	8.50	20.71	69.03
Death During Hospital Stay	Yes	1.91	N/A	N/A	1.21
	No	98.09	N/A	N/A	98.79

Supplementary Table 3. Confusion matrix of ten-fold cross validation results, using all features. Values in the table represent averages \pm standard deviations.

Sepsis:

	Predicted Positive	Predicted Negative
Actual Positive	17109 \pm 424	774 \pm 424
Actual Negative	30 \pm 2	110 \pm 2

Severe Sepsis:

	Predicted Positive	Predicted Negative
Actual Positive	15322 \pm 369	2531 \pm 369
Actual Negative	48 \pm 4	171 \pm 4

Septic Shock:

	Predicted Positive	Predicted Negative
Actual Positive	18801 \pm 9	19 \pm 9
Actual Negative	67 \pm 3	265 \pm 3

Supplementary Table 4. Confusion matrix of ten-fold cross validation results, with gold standard definition associated inputs removed. Values in the table represent averages \pm standard deviations.

Sepsis:

	Predicted Positive	Predicted Negative
Actual Positive	13655 \pm 488	4231 \pm 488
Actual Negative	34 \pm 3	120 \pm 3

Severe Sepsis:

	Predicted Positive	Predicted Negative
Actual Positive	13855 \pm 2381	4015 \pm 2381
Actual Negative	72 \pm 25	147 \pm 25

Septic Shock:

	Predicted Positive	Predicted Negative
Actual Positive	17972 \pm 320	878 \pm 320
Actual Negative	66 \pm 0	260 \pm 0

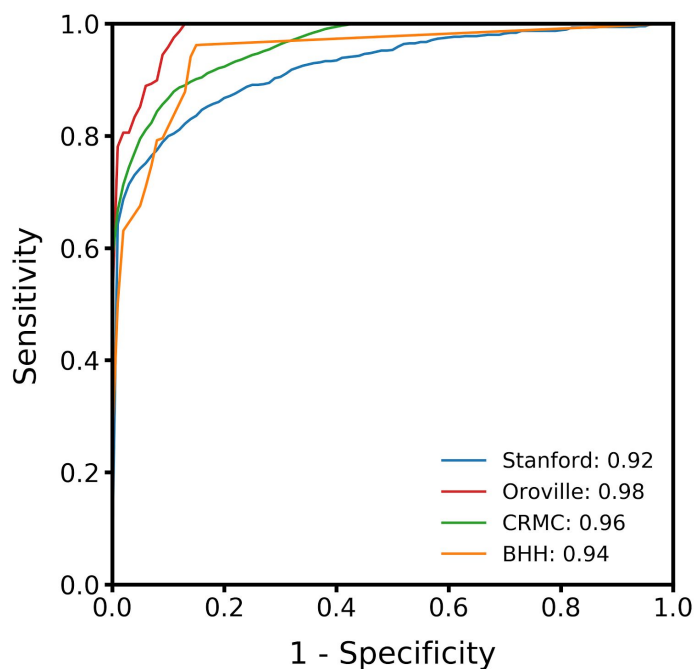
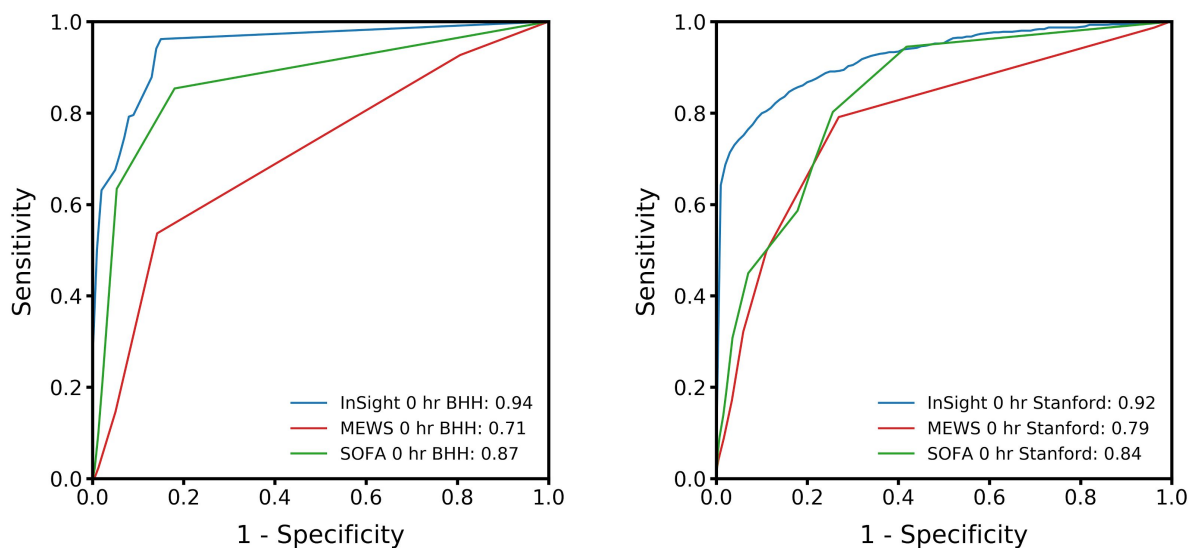


Figure 1. ROC curves for *InSight* at time of severe sepsis onset on data from Stanford, Oroville, Cape Regional Medical Center (CRMC), and Bakersfield Heart Hospital (BHH).



Supplementary Figure 2. ROC curves for *InSight*, MEWS, and SOFA applied to severe sepsis detection at time of onset, on data from Bakersfield Heart Hospital (BHH, left) and Stanford Medical Center (right).

Supplementary Table 5. Comparison of performance metrics for *InSight* and rules-based methods for severe sepsis detection at time of onset on data from Stanford Medical Center. LR: Likelihood ratio; SIRS: Systemic Inflammatory Response Syndrome; MEWS: Modified Early Warning Score; SOFA: Sequential (Sepsis-Related) Organ Failure Assessment; qSOFA: Quick SOFA.

Stanford	<i>InSight</i>	SIRS	MEWS	SOFA	qSOFA
AUROC	0.924	0.710	0.786	0.836	0.836
Sensitivity	0.798	0.798	0.791	0.802	0.802
Specitivity	0.901	0.901	0.731	0.744	0.744
Accuracy	0.900	0.900	0.885	0.789	0.789
LR+	8.253	8.253	2.940	3.133	3.133
LR-	0.224	0.224	0.286	0.266	0.266

Supplementary Table 6. Comparison of performance metrics for *InSight* and rules-based methods for severe sepsis detection at time of onset on data from Bakersfield Heart Hospital (BHH). LR: Likelihood ratio; SIRS: Systemic Inflammatory Response Syndrome; MEWS: Modified Early Warning Score; SOFA: Sequential (Sepsis-Related) Organ Failure Assessment; qSOFA: Quick SOFA.

BHH	<i>InSight</i>	SIRS	MEWS	SOFA	qSOFA
AUROC	0.945	0.678	0.707	0.869	0.665
Sensitivity	0.875	0.561	0.927	0.854	0.366
Specitivity	0.940	0.764	0.194	0.820	0.964
Accuracy	0.963	0.957	0.851	0.940	0.977
LR+	58.94	2.373	1.150	4.736	10.27
LR-	0.129	0.574	0.378	0.179	0.658

Supplementary Table 7. Comparison of performance metrics for *InSight* and rules-based methods for severe sepsis detection at time of onset on data from Oroville Hospital. LR: Likelihood ratio; SIRS: Systemic Inflammatory Response Syndrome; MEWS: Modified Early Warning Score; SOFA: Sequential (Sepsis-Related) Organ Failure Assessment; qSOFA: Quick SOFA.

Oroville	<i>InSight</i>	SIRS	MEWS	SOFA	qSOFA
AUROC	0.983	0.708	0.792	0.938	0.731
Sensitivity	0.806	0.602	0.685	0.778	0.537
Specitivity	0.989	0.757	0.811	0.926	0.921
Accuracy	0.971	0.909	0.883	0.917	0.914
LR+	77.92	2.476	3.616	10.52	6.836
LR-	0.197	0.526	0.388	0.240	0.502

Supplementary Table 8. Comparison of performance metrics for *InSight* and rules-based methods for severe sepsis detection at time of onset on data from Cape Regional Medical Center (CRMC). LR: Likelihood ratio; SIRS: Systemic Inflammatory Response Syndrome; MEWS: Modified Early Warning Score; SOFA: Sequential (Sepsis-Related) Organ Failure Assessment; qSOFA: Quick SOFA.

CRMC	<i>InSight</i>	SIRS	MEWS	SOFA	qSOFA
AUROC	0.960	0.732	0.554	0.749	0.560
Sensitivity	0.802	0.591	0.478	0.631	0.155
Specitivity	0.946	0.864	0.567	0.831	0.965
Accuracy	0.931	0.860	0.826	0.866	0.855
LR+	16.85	4.346	1.104	3.739	4.470
LR-	0.210	0.473	0.921	0.444	0.875

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	D;V Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	Title
Abstract	2	D;V Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	1
Introduction			
Background and objectives	3a	D;V Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	2
	3b	D;V Specify the objectives, including whether the study describes the development or validation of the model or both.	3
Methods			
Source of data	4a	D;V Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	3
	4b	D;V Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	3
Participants	5a	D;V Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	3
	5b	D;V Describe eligibility criteria for participants.	6
	5c	D;V Give details of treatments received, if relevant.	n/a
Outcome	6a	D;V Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	4-5
	6b	D;V Report any actions to blind assessment of the outcome to be predicted.	n/a
Predictors	7a	D;V Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	6
	7b	D;V Report any actions to blind assessment of predictors for the outcome and other predictors.	n/a
Sample size	8	D;V Explain how the study size was arrived at.	6
Missing data	9	D;V Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	4
Statistical analysis methods	10a	D Describe how predictors were handled in the analyses.	8
	10b	D Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	8
	10c	V For validation, describe how the predictions were calculated.	
	10d	D;V Specify all measures used to assess model performance and, if relevant, to compare multiple models.	6
	10e	V Describe any model updating (e.g., recalibration) arising from the validation, if done.	n/a
Risk groups	11	D;V Provide details on how risk groups were created, if done.	n/a
Development vs. validation	12	V For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	
Results			
Participants	13a	D;V Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	8
	13b	D;V Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	8
	13c	V For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	
Model development	14a	D Specify the number of participants and outcome events in each analysis.	8
	14b	D If done, report the unadjusted association between each candidate predictor and outcome.	n/a
Model specification	15a	D Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	n/a
	15b	D Explain how to use the prediction model.	n/a
Model performance	16	D;V Report performance measures (with CIs) for the prediction model.	12-13
Model-updating	17	V If done, report the results from any model updating (i.e., model specification, model performance).	
Discussion			
Limitations	18	D;V Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	16
Interpretation	19a	V For validation, discuss the results with reference to performance in the development data, and any other validation data.	
	19b	D;V Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	15
Implications	20	D;V Discuss the potential clinical use of the model and implications for future research.	15
Other information			
Supplementary information	21	D;V Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	3
Funding	22	D;V Give the source of funding and the role of the funders for the present study.	18

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.