BMJ Open Respiratory Research

Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial

David W Shimabukuro,[1] Christopher W Barton,[2] Mitchell D Feldman,[3] Samson J Mataraso,[4,5] Ritankar Das[6]

To cite: Shimabukuro DW, Barton CW, Feldman MD, *et al*. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Resp Res* 2017;**4**:e000234. doi:10.1136/bmjresp-2017-000234

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjresp-2017-000234)

Received 24 July 2017
Revised 18 October 2017

For numbered affiliations see end of article.

Correspondence to
Ritankar Das;
ritankar@dascena.com

ABSTRACT

Introduction Several methods have been developed to electronically monitor patients for severe sepsis, but few provide predictive capabilities to enable early intervention; furthermore, no severe sepsis prediction systems have been previously validated in a randomised study. We tested the use of a machine learning-based severe sepsis prediction system for reductions in average length of stay and in-hospital mortality rate.

Methods We conducted a randomised controlled clinical trial at two medical-surgical intensive care units at the University of California, San Francisco Medical Center, evaluating the primary outcome of average length of stay, and secondary outcome of in-hospital mortality rate from December 2016 to February 2017. Adult patients (18+) admitted to participating units were eligible for this factorial, open-label study. Enrolled patients were assigned to a trial arm by a random allocation sequence. In the control group, only the current severe sepsis detector was used; in the experimental group, the machine learning algorithm (MLA) was also used. On receiving an alert, the care team evaluated the patient and initiated the severe sepsis bundle, if appropriate. Although participants were randomly assigned to a trial arm, group assignments were automatically revealed for any patients who received MLA alerts.

Results Outcomes from 75 patients in the control and 67 patients in the experimental group were analysed. Average length of stay decreased from 13.0 days in the control to 10.3 days in the experimental group (p=0.042). In-hospital mortality decreased by 12.4 percentage points when using the MLA (p=0.018), a relative reduction of 58.0%. No adverse events were reported during this trial.

Conclusion The MLA was associated with improved patient outcomes. This is the first randomised controlled trial of a sepsis surveillance system to demonstrate statistically significant differences in length of stay and in-hospital mortality.

Trial registration NCT03015454.

Key messages

- This study represents the first randomised controlled trial for a machine learning-based sepsis prediction algorithm to demonstrate statistically significant differences in length of stay and in-hospital mortality.
- The algorithm uses only six vital signs to provide higher sensitivity and specificity than commonly used sepsis scoring systems.
- This is a single-centre study in the intensive care unit only. We plan to address limited generalisability in further studies and algorithm development.

INTRODUCTION

Severe sepsis affects more than 700 000 individuals in the USA each year[1] at a cost of more than 20 billion dollars.[2] While sepsis definitions vary, including the recent third international consensus definitions,[3] here we define severe sepsis as 'organ dysfunction caused by sepsis',[4] and sepsis as a dysregulated host response to infection.[5] Severe sepsis has an estimated annual mortality of 250 000,[1, 6] but early diagnosis has been shown to reduce delays in treatment, increase appropriate care and reduce mortality.[7, 8]

In prior studies, we have demonstrated the efficacy of a machine learning algorithm (MLA) developed by Dascena (Hayward, California, USA) for the early prediction of sepsis, severe sepsis and septic shock.[9–11] Requiring inputs of only the most commonly recorded measurements in the electronic health record (EHR), primarily vital signs and age, the MLA predicted sepsis with accuracy which was superior to disease severity scoring systems in current use, such as the Sequential Organ Failure Assessment (SOFA),[12] the Systemic Inflammatory Response Syndrome (SIRS) criteria[13] and the Modified Early Warning Score (MEWS).[14] At the time of severe sepsis onset, the MLA achieved an area under the receiver operating characteristic (AUROC) curve of 0.880 (SD=0.006) compared with 0.725, 0.609 and 0.803 for SOFA, SIRS and MEWS,

respectively.[9] Although these disease severity scoring systems were designed to predict patient risk, rather than specifically to identify sepsis, they are commonly used in severe sepsis diagnostic criteria due to their designed purposes of identifying systemic inflammation as a sign of possible infection and detecting possible organ dysfunction. Because of their close relation to sepsis diagnostic criteria, the clinical utility of such scoring systems for identifying patients with sepsis has been closely studied in the literature.[15 16] These scoring systems therefore serve as important comparators for any newly developed severe sepsis prediction system. Though relatively new additions to the field of sepsis care, MLAs have the potential to greatly improve patient outcomes through their accuracy and advanced warning of impending sepsis onset, making studies of such tools of great importance. The MLA used in this study has been described at length in previous peer-reviewed publications.[9–11]

MLAs for sepsis prediction[17] have primarily been tested retrospectively or investigated non-interventionally.[18–22] Here, we report a prospective, randomised controlled study, in which an algorithm was applied to EHR data for the prediction of severe sepsis (in a manner akin to a biomarker) and if warranted, generated real-time telephonic notifications at the University of California, San Francisco (UCSF) Medical Center (San Francisco, California, USA). We tested the hypothesis that the use of an MLA would result in reductions in the average length of stay (LOS) and the in-hospital mortality rate. To the best of the authors' knowledge, this present work represented the first time a machine learning-based sepsis prediction system has been investigated in a randomised, interventional design.

The design of this study involved little or no risk of harm but conferred a large potential benefit. Specifically, the prediction algorithm's ability to identify patients with severe sepsis before onset provided the opportunity for early intervention, which has been widely shown to decrease patient mortality.[23 24] Kumar *et al* [7] found that survival decreased by 7.6% for every hour in which antimicrobial therapy is not administered to patients with septic shock following the first hour after onset. Although there is some controversy about the conclusions drawn by Kumar *et al* about the linear relationship between antibiotic timing and survival, as well as concerns that the researchers did not properly consider confounding factors in reporting their outcomes, conflicting evidence has largely shown that waiting to administer pathogen-specific antibiotics until after confirmation of a positive microbiology is associated with improved patient outcomes.[25] Therefore, early identification of patients with severe sepsis still provides a large potential benefit by providing an opportunity for earlier confirmation of infection. If the MLA produced an alert when a patient was not trending towards severe sepsis (false positive), there was no direct harm to the patient, but clinicians would incur additional burden to assess the patient and dismiss the false alert. However, with the algorithm's high specificity (as demonstrated by its high AUROC value),[9–11] this risk was minimised. In the case that the algorithm did not identify a patient trending towards severe sepsis (false negative), there was no risk of additional harm, since UCSF's current rules-based severe sepsis detection system was still active.

METHODS

Enrolment and study design

From December 2016 to February 2017, we conducted a randomised clinical trial (Trial Registration: ClinicalTrials.gov NCT03015454) in two mixed medical-surgical intensive care units (ICUs) at the UCSF Medical Center at Parnassus Heights. Across both units, the MLA monitored a total of 32 patient beds. This study was approved by the UCSF Institutional Review Board with a waiver of informed consent for all patients.

During the study period, all patients over the age of 18 admitted to the participating units were automatically enrolled in the trial. A patient admitted with a sepsis diagnosis was still monitored by the prediction algorithm for potential further septic episodes; thus, these patients were not excluded from the trial. Enrolment entailed that a patient's vital signs and selected lab results were abstracted from UCSF's EHR software, APeX, into the prediction algorithm. APeX was developed by Epic Systems (Verona, Wisconsin, USA), and the prediction algorithm was developed by Dascena (Hayward, California, USA).

Patients were assigned to the experimental group or control group based on a random allocation sequence, generated by a computer program using simple randomisation before the start of the trial. This allocation sequence was concealed within a vector in the backend of the prediction algorithm software. Healthcare providers, patients and investigators were thus unaware of patient assignment, although group assignments were naturally revealed for patients who generated MLA alerts. The programme drew 10 000 samples from a probability distribution with $P(x=0)=0.50$ and $P(x=1)=0.50$ for each sample, x. Participants assigned '0' were placed in the control group by the application, and patients assigned a '1' were placed in the experimental group. This method was designed to achieve a 1:1 allocation ratio.

Patients were enrolled in accordance with the trial period, during which we estimated that approximately 150 participants would be enrolled. The trial had a factorial, open-label design, and healthcare providers, patients and investigators were not made aware of group assignment but could not be fully blinded as some group assignments became naturally revealed upon receipt of MLA alerts.

Patients in the control group received the normal standard of care and were monitored by the existing EHR-based severe sepsis detector, which uses threshold-based cut-offs of SIRS criteria and end-organ dysfunction haemodynamic or lab results.[26] UCSF recognised

severe sepsis as organ dysfunction caused by infection as defined by the Surviving Sepsis Campaign.[27] Clinician suspicion of infection was required for diagnosis, though identification of the infectious agent through a positive culture was not required to diagnose severe sepsis. The standard of care consisted of a nurse evaluation of the patient at the bedside for suspicion of infection. Nurse evaluation included assessment of patient vital signs, EHR notes, laboratory results such as white blood cell (WBC) count and results of any additional testing ordered. If severe sepsis was suspected, a physician subsequently assessed the patient and placed an order for administration of the standard UCSF sepsis bundle (see online supplementary materials).

In the experimental group, patients were monitored by the MLA, in addition to the existing severe sepsis detector. If the algorithm predicted severe sepsis for a given patient, a phone call was placed to the charge nurse on duty; however, no recommendations for treatment were provided with the notification. The charge nurse then followed UCSF's standard severe sepsis evaluation and intervention process. Alternatively, if the MLA has failed to forecast severe sepsis for a given patient, the existing UCSF severe sepsis detector may have identified the patient at a later time. Participants received the same severe sepsis assessment and treatment, regardless of which system predicted or detected their severe sepsis. Thus, the trial was designed to demonstrate the superiority of using an algorithmic predictor relative to the hospital's current EHR-native rules-based severe sepsis surveillance system.

Data collection and analysis

Demographic information and clinical measurements were collected from each enrolled patient's medical record. All patient data were collected from admission to discharge in the participating units, although participants were followed until hospital discharge in order to determine in-hospital mortality and overall LOS.

Using the required measurements (systolic/diastolic blood pressure, heart rate, temperature, respiratory rate, peripheral capillary oxygen (SpO_2) and age), the MLA monitored patients hourly. The MLA also incorporated additional optional inputs as they were available (selected labs, eg, pH, WBC count, glucose). A full list of variables collected by the MLA is available as online supplemental table S1. The machine learning-based classifiers in the algorithm were used to generate a risk score predictive of severe sepsis. These classifiers analysed multidimensional patterns and time-series trends to improve severe sepsis forecasting.[11] The resulting risk score ranged between 0 and 100 for each patient, and if it exceeded the preset threshold of 80, the charge nurse was called.

At the conclusion of the clinical trial, we evaluated the primary outcome (average hospital LOS) and secondary outcome (in-hospital mortality rate) as well as ICU LOS, using outcome-related measurements (online supplemental table S1) which were collected for all enrolled patients. Given the high historical sepsis prevalence in the study units,[26] sepsis-related LOS and mortality were expected to be sufficiently represented in these more general outcome metrics. Additionally, we retrospectively compared algorithm performance on patient data from the study with the performance of MEWS, SIRS, SOFA and the quick SOFA (qSOFA) score on the same prospectively collected data. No interim analyses were performed before the conclusion of the trial.

Statistical analysis

The desired sample size for this study was calculated to detect a reduction of 1.5 days in hospital LOS at a power of 0.80 and a type I error rate of 0.05. Two-sample t-tests were used to determine if there was a statistically significant difference in means between the experimental and control groups for hospital LOS and for ICU LOS. We used the two-proportion (risk difference) and relative risk (risk ratio) z-tests to determine if there was a statistically significant decrease in the in-hospital mortality rate with the use of the predictive algorithm. All tests were single tailed with an alpha level of 0.05, and were performed using the MATLAB software (V.R2016a) developed by MathWorks (Natick, Massachusetts, USA).

RESULTS

Patient inclusion and baseline characteristics

During the course of the trial period, 142 patients from the participating units were assessed for eligibility. Patient enrolment was stopped before the projected enrolment size of 150 due to the conclusion of the study period. Of those patients, all met the inclusion criteria, as none were younger than 18 years of age. This resulted in the randomisation of 142 patients, 75 of which were assigned to the control group and the other 67 assigned to the experimental group (figure 1). No patients were lost to follow-up since participants were tracked throughout their hospital stay, and the study did not require patients to return for additional intervention or questioning. The outcomes of all 142 participants were analysed, and the trial concluded at that point.

Demographic and clinical characteristics were collected for all patients (table 1). There were 53.5% women and 46.5% men in this study, and the average age of participating patients was 59 years old (SD=16.5 years). No differences in demographic distributions between the two arms of the study were statistically significant. The top three routes of admission for patients in both arms were, in ranked order, the emergency department, the transfer centre and UCSF's medical-surgical high acuity ward.

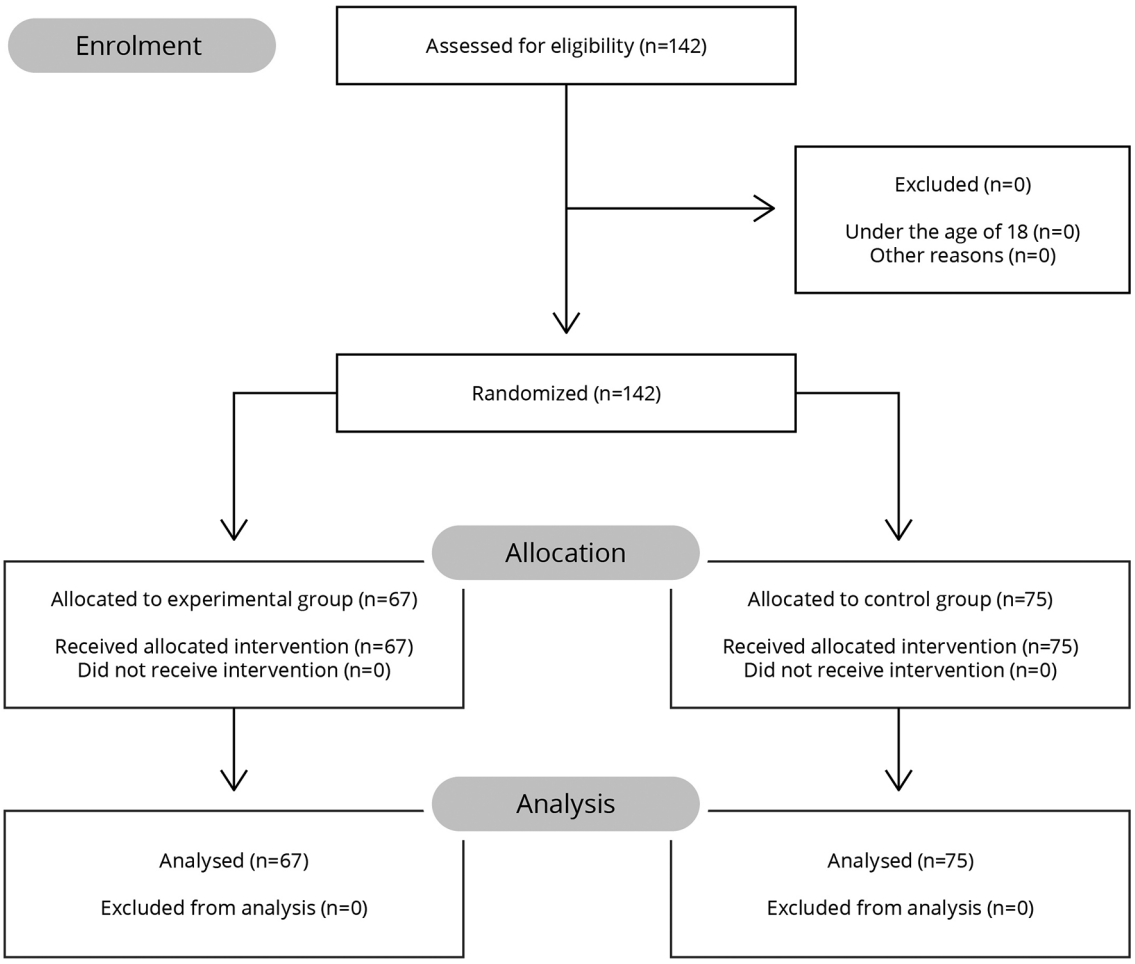

Figure 1 Patients assessed, enrolled, randomised and analysed in each arm of the randomised controlled trial.

Outcomes

The primary outcome, average hospital LOS, was 13.0 days in the control group and 10.3 days in the experimental group, representing a 20.6% reduction (figure 2, table 2). This decrease was statistically significant with a 95% one-sided confidence interval (CI) with an upper bound of −3.08 hours. Similarly, average ICU LOS was 8.40 days in the control group and 6.31 days in the experimental group (table 2). Using a two-sample t-test, we found a statistically significant decrease in ICU LOS with a 95% one-sided CI with an upper bound of −6.30 hours.

To assess the secondary outcome, for each of the 142 patients we tabulated the number of in-hospital deaths within each group and divided by the number of total group members. The control group contained 16 out of 75 patients with in-hospital mortality, while the experimental group had 6 in-hospital fatalities out of 67 patients (figure 3, table 2). There was a statistically significant decrease in the difference of in-hospital mortality rate, with a one-sided 95% CI with an upper bound of −0.0271. There was also a statistically significant decrease in the risk ratio of in-hospital mortality rate (p=0.026), with a 95% CI with an upper bound of 0.880. Therefore, in-hospital mortality decreased by 12.4 percentage points in the experimental group, a 58.0% relative decrease. One patient included in the study was discharged to hospice care. This patient was randomised to the control group, and was treated as 'alive' for the purposes of calculating in-hospital mortality.

Patient outcomes also improved in the experimental group for the subpopulation of patients who received International Classification of Diseases 10 codes for sepsis, severe sepsis or septic shock (online supplemental table S2). For these patients, the average hospital LOS was 16.8 days in the control group and 9.83 days in the experimental group (p=0.042). For the same subpopulation, the mortality rate was 40.0% in the control group and 13.6% in the experimental group (p=0.023).

Patients in the experimental group additionally received antibiotics an average of 2.76 hours earlier than patients in the control group, and had blood cultures drawn an average of 2.79 hours earlier than patients in the control group. Of the 75 patients in the control group, 39 were administered antibiotics and 30 had blood cultures drawn. Of the 67 patients in the experimental group, 31 received antibiotics and 22 had blood cultures drawn.

On physiological data collected during the study from the enrolled participants, the algorithm more accurately detected severe sepsis than MEWS, SIRS criteria, the SOFA score or the qSOFA score in a retrospective analysis

Table 1 Patient demographics and comorbidities in the experimental and control groups

	Control (n=75)	Experimental (n=67)	P values
Male, count (%)	31 (41)	35 (52)	0.09
Age, mean (SD)	59.3 (16.3)	58.9 (16.8)	0.49
Race and ethnicity, count (%)			
White	36 (48)	30 (45)	0.35
African American	10 (13)	6 (9.0)	0.21
Asian American	13 (17)	9 (13)	0.26
Hispanic	13 (17)	17 (25)	0.12
Other	3 (4.4)	5 (7.5)	0.18
Comorbidities, count (%)			
Sepsis	9 (12)	16 (24)	0.03
Severe sepsis with septic shock	7 (9.3) 4 (5.3)	5 (7.5) 1 (1.5)	0.34 0.11
Cardiovascular	17 (23)	14 (21)	0.39
Renal	10 (13)	8 (12)	0.40
Liver	4 (5.3)	3 (4.5)	0.41
Organ transplant	10 (13)	11 (16)	0.30
HIV positive	2 (2.7)	2 (3.0)	0.45
Mental health disorder	2 (2.7)	1 (1.5)	0.31
Diabetes	9 (12)	9 (13)	0.40
COPD	3 (4)	1 (1.5)	0.18
Cancer	26 (35)	32 (48)	0.06
Alcohol abuse	4 (5.3)	1 (1.5)	0.11
Pneumonia	7 (9.3)	6 (9)	0.47

Comorbidities are based on International Classification of Diseases 10 codes (see online supplemental table S2). P value for statistically significant differences in the distribution of demographics were calculated with a two-proportion z-test for all categorical variables, and a two-sample t-test for the continuous variable (age). Significance was set at 0.05.
COPD, Chronic Obstructive Pulmonary Disease.

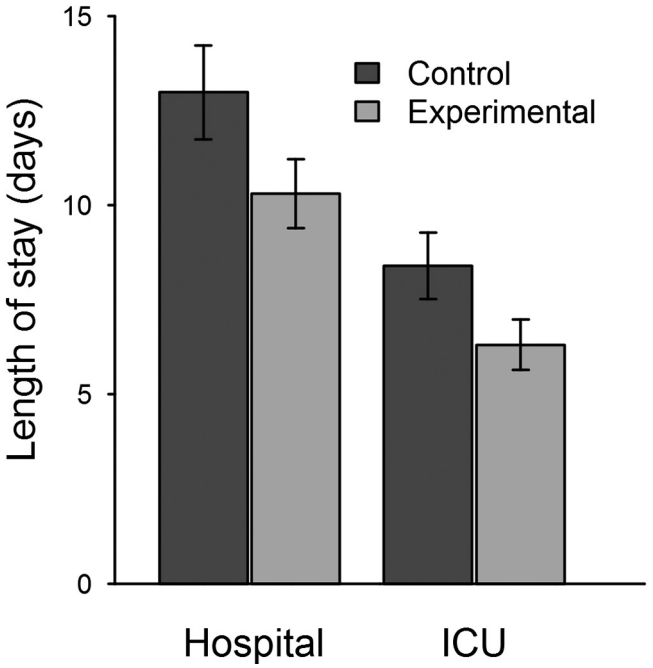

Figure 2 Decrease in average hospital and ICU length of stay with the use of the machine learning algorithm. The error bars represent one standard error above and below the mean length of stay. ICU, intensive care unit.

(table 3). Our gold standard for severe sepsis was defined as patients meeting two SIRS criteria and two organ dysfunction criteria within the same hour (see online supplementary materials).

No harms with respect to either intervention were reported throughout the duration of the trial. Similarly, no adverse events were observed in the experimental group. The outcomes data indicated clear benefits in LOS and mortality reduction when using the predictive algorithm over the EHR-based sepsis detector.

DISCUSSION

The use of the machine learning-based predictor resulted in significant decreases in LOS and in-hospital mortality rate during this randomised controlled trial. Specifically, we found a 20.6% decrease in average hospital LOS from 13.0 to 10.3 days (p=0.042) and a 12.4% decrease in in-hospital mortality rate from 21.3% to 8.96% (p=0.018) when using the MLA. While one patient from the control group of the study was discharged to hospice care, changing the mortality classification of this patient would not have impacted the results of the mortality analysis in this study. Decreases in average hospital LOS and in-hospital mortality were also found in the experimental group among the subpopulation of patients diagnosed with sepsis, severe sepsis or septic shock. Additionally, improvement was found in time to blood culture draws and antibiotic administration in the experimental group. We have demonstrated past success with retrospective applications of sepsis detection methods,[9–11] as well as mortality prediction methods[28 29]; this study suggested that these retrospective successes will translate into improved clinical outcomes.

Past randomised controlled trials investigating electronic sepsis monitoring tools have failed to achieve statistically significant outcomes for reduction of in-hospital mortality rate and LOS, and improvement in bundle compliance[30 31] (online supplemental table S3). Semler *et al* implemented an electronic sepsis alerting tool based on a series of logic rules in two ICUs. While they aimed to improve bundle compliance and clinical outcomes, they did not find a statistically significant difference in time to intervention, and their tool was underused by clinicians.[31] Similarly, Hooper *et al* installed a sepsis surveillance system that alerted clinicians when two or more modified SIRS criteria were met. They were

Table 2 Differences in hospital LOS, ICU LOS, and in-hospital mortality between the experimental and control groups

Outcome	Control (n=75)	Experimental (n=67)	Amount of reduction	P value
Hospital LOS (days)	13.0 (1.23)	10.3 (0.912)	2.30 days	0.042
ICU LOS (days)	8.40 (0.881)	6.31 (0.666)	2.09 days	0.030
In-hospital mortality rate	21.3% (4.76%)	8.96% (3.51%)	12.3%	0.018

The mean and the standard error (in parentheses) for each outcome are noted in the table. All outcomes demonstrate statistically significant reductions when using the machine learning algorithm ($p<0.05$).
ICU, intensive care unit; LOS, length of stay.

unsuccessful in showing reduced LOS and time to bundle compliance in the ICU.[30]

Unlike previous randomised clinical trials, several prospective observational studies have shown improvements in severe sepsis-related clinical or patient outcomes. In a recent before-and-after study, Manaktala and Claypool[32] reported a 53% relative decrease in the sepsis-related in-hospital mortality rate (p=0.03) through the implementation of a complex rules-based sepsis alerting system. Although not based on machine-learning models, the study by Manaktala and Claypool demonstrates the potential for large improvements in patient outcomes with the implementation of advanced sepsis surveillance, and the present work provides further evidence in that direction. Similarly, the prospective studies of both Sawyer *et al*[22] and Berger *et al*[33] achieved statistically significant decreases in time to clinical intervention for patients who triggered a sepsis alert. Though they did not provide evidence of improved patient outcomes, the results of these studies indicate the potential for reductions in key clinical metrics using automated sepsis surveillance systems. See online supplemental table S3 for a full comparison of recent studies and associated outcomes.

The broader sepsis screening literature primarily consists of rule-based thresholds for alerts, triggered when predetermined criteria are met. These systems, unlike the present work, generally detect, not predict, a sepsis syndrome.[30 33] There is likely a relationship between this algorithm's significant improvements in patient outcomes and its design to predict severe sepsis up to 4 hours in advance.[11] Predictions made by the MLA in this study likely also influenced the septic shock counts observed in table 1. With extra time to intervene in the experimental group, patients may not have ultimately progressed to septic shock, thus producing different prevalences in the experimental (1.5%) and control (5.3%) groups. This interpretation was supported by the higher proportion of patients diagnosed with sepsis in the experimental arm of the trial; this higher proportion implied that sepsis was detected and treated earlier in the experimental group, thereby deterring progression to septic shock.

In addition to the algorithm's predictive nature, improvements in patient outcomes likely reflected the MLA's combination of high sensitivity and high specificity.[10 11] This reasoning is in close alignment with Manaktala and Claypool's study.[32] Sepsis scoring systems such as SIRS and SOFA often correctly identify patients with sepsis, maintaining a high sensitivity, but have low specificities, incorrectly classifying non-septic patients.

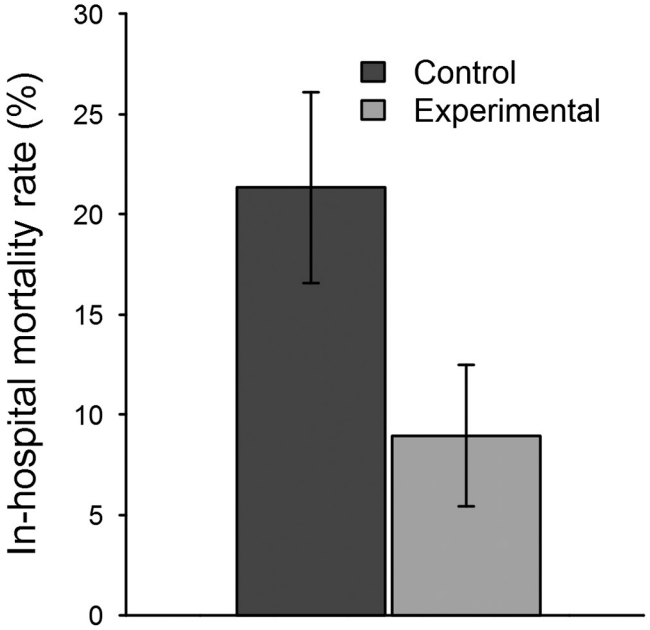

Figure 3 Reduction of in-hospital mortality rate when using the machine learning algorithm. The error bars represent one standard error above and below the average in-hospital mortality rate.

Table 3 Comparison of AUROC, sensitivity and specificity for the MLA applied to severe sepsis detection and SIRS criteria, MEWS, the SOFA score and the qSOFA score on patient physiological data collected during the study

	MLA	SIRS	MEWS	SOFA	qSOFA
AUROC	0.952 (0.946 to 0.958)	0.681	0.524	0.756	0.518
Sensitivity	0.900 (0.870 to 0.930)	0.590	0.365	0.910	0.288
Specificity	0.900 (0.878 to 0.922)	0.764	0.667	0.181	0.750

A 95% CI for the MLA is also included in parentheses.
AUROC, area under the receiver operator characteristic curve; MEWS, Modified Early Warning Score; MLA, machine learning algorithm; SIRS, Systemic Inflammatory Response Syndrome; qSOFA, quick Sequential Organ Failure Assessment.

Systems that use these scores deliver many false alarms, which could impact a clinician's willingness to use the sepsis classification tool.[34] The prediction algorithm may have been more readily used compared with other sepsis classification systems with lower specificities.

Moreover, several of the models described in the literature require multiple laboratory test results, which may not be ordered if clinicians do not suspect sepsis. Some also require specialist annotation and interpretation of clinical notes. In contrast, this MLA required only vital signs and age to forecast sepsis onset several hours early,[11] and was additionally able to incorporate optional laboratory results when they were available for accurate forecasting using a range of possible clinical measurements. This aspect of the algorithm may have also contributed to improved outcomes in this randomised controlled trial.

Limitations

The present work was a single centre randomised controlled trial conducted at the UCSF Medical Center, which treats a fairly heterogeneous patient population, relative to US national averages. Thus, this finding may be less generalisable to institutions whose patient populations are more homogeneous with similar demographics and comorbidities among patients. Further, because this trial was conducted in two ICUs, we cannot generalise the algorithm's performance in this study to other wards such as the emergency department or floor units, where data collection frequency and admission diagnoses differ. The small sample size used in this study, the short study period and the single season over which the study was conducted additionally limit generalisability of these results.

In previous studies, this MLA has been shown to predict sepsis, severe sepsis and septic shock several hours early with a higher sensitivity and specificity than rules-based sepsis screening approaches.[11] These metrics were not monitored prospectively during the study due to the likely misrepresentation of such results. With advanced notice from the predictive algorithm, clinicians may have initiated treatment before severe sepsis onset, thus averting the diagnosis. These cases could be documented as false positives, skewing the prediction algorithm's sensitivity.

We are unable to rule out the possibility that the predictive algorithm improved clinical outcomes by improving clinician awareness of high-risk patients rather than by predicting sepsis early. Further, we did not implement any precautionary measures which prevented clinicians from more closely monitoring patients in the experimental arm of the trial after MLA alerts were generated. However, this is likely reflective of the type of care known high-risk patients typically receive, and may therefore illustrate the algorithm's expected performance in future care settings. Additionally, the use of overall metrics, LOS and in-hospital mortality for all comers, rather than sepsis-specific LOS and mortality as primary and secondary outcomes, may underestimate the impact of the intervention on outcomes for patients with sepsis.

An additional limitation is the potential for competing risks in the selected endpoints. Because mortality may shorten a patient's LOS, there is some inevitable censoring when measuring changes in the two outcomes. The actual reduction in LOS for patients who were monitored by the MLA could have been larger than reported in this study, given the lower mortality rate in the experimental group. However, we did not have the data required to calculate additional metrics such as hospital free days out of 30 and therefore could not use these metrics in order to manage the competing risks of the selected endpoints. Further, this study was patient outcome-oriented; in future work, we plan to more thoroughly study endpoints related to clinical workflow and interventional actions such as time to fluid bolus administration.

CONCLUSION

In this clinical trial, we demonstrated improvements in patient outcomes when using a machine learning-based sepsis prediction algorithm. We found a statistically significant decrease in the hospital LOS and in-hospital mortality when using this algorithm compared with the current rules-based sepsis detector. From these results, we deduced that machine learning-based sepsis prediction algorithms may lead to earlier clinical intervention and improved patient outcomes. Some limitations of this study include its containment to a single centre and to ICUs only. In future studies, we intend to validate the MLA's performance in hospitals with varying demographic and clinical characteristics as well as in non-critical care units.

Author affiliations
[1]Division of Critical Care Medicine, Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, California, USA
[2]Department of Emergency Medicine, University of California San Francisco, San Francisco, California, USA
[3]Division of General Internal Medicine, Department of Medicine, University of California San Francisco, San Francisco, California, USA
[4]Department of Bioengineering, University of California Berkeley, Berkeley, California, USA
[5]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California, USA
[6]Dascena, Inc, Hayward, California, USA

Acknowledgements The authors gratefully acknowledge Jana Hoffman, Melissa Jay, Qingqing Mao, Megan Handley, Anna Lynn-Palevsky, Emily Huynh, and Jacob Calvert for their suggestions and help with editing.

Contributors DWS, CWB, MF, SJM and RD contributed to the conception and design of this study, and to acquisition, analysis or interpretation of data. SJM and RD were responsible for statistical analysis. SJM, DWS and RD drafted the manuscript. DWS, CWB, MDF, SJM and RD revised the manuscript and have approved it in this final form.

Funding This material is based on work supported by the National Science Foundation under Grant No. 1549867. The funder had no role in the conduct of the study; collection, management, analysis and interpretation of data; preparation, review and approval of the manuscript; and decision to submit the manuscript for publication. Research reported in this publication was supported by the National Institute of Nursing Research, of the National Institutes of Health, under award number R43NR015945. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests SM and RD are employees of Dascena. CB reports receiving consulting fees from Dascena. CB, DS and MF report receiving research grant funding from Dascena.

Ethics approval This study was approved by the University of California, San Francisco Institutional Review Board with a waiver of informed consent for all patients.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Any inquiries regarding the dataset can be addressed to the corresponding author.

REFERENCES

1. Angus DC, Linde-Zwirble WT, Lidicker J, *et al*. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Crit Care Med* 2001;29:1303–10.
2. Chalupka AN, Talmor D. The economics of sepsis. *Crit Care Clin* 2012;28:57–76.
3. Singer M, Deutschman CS, Seymour CW, *et al*. The third international consensus definitions for Sepsis and Septic Shock (sepsis-3). *JAMA* 2016;315:801–10.
4. Henry KE, Hager DN, Pronovost PJ, *et al*. A Targeted Real-time Early Warning Score (TREWScore) for septic shock. *Sci Transl Med* 2015;7:299ra122.
5. Lever A, Mackenzie I. Sepsis: definition, epidemiology, and diagnosis. *BMJ* 2007;335:879–83.
6. Stevenson EK, Rubenstein AR, Radin GT, *et al*. Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis*. *Crit Care Med* 2014;42:625.
7. Kumar A, Roberts D, Wood KE, *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;34:1589–96.
8. Nguyen HB, Corbett SW, Steele R, *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;35:1105–12.
9. Calvert JS, Price DA, Chettipally UK, *et al*. A computational approach to early sepsis detection. *Comput Biol Med* 2016;74:69–73.
10. Calvert J, Desautels T, Chettipally U, *et al*. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Ann Med Surg* 2016;8:50–5.
11. Desautels T, Calvert J, Hoffman J, *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.
12. Vincent JL, Moreno R, Takala J, *et al*. 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;22:707–10.
13. Balk RA. Severe sepsis and septic shock. Definitions, epidemiology, and clinical manifestations. *Crit Care Clin* 2000;16:179–92.
14. Subbe CP, Slater A, Menon D, *et al*. Validation of physiological scoring systems in the accident and emergency department. *Emerg Med J* 2006;23:841–5.
15. McLymont N, Glover GW. Scoring systems for the characterization of sepsis and associated outcomes. *Ann Transl Med* 2016;4:527.
16. Churpek MM, Snyder A, Han X, *et al*. Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside the intensive care unit. *Am J Respir Crit Care Med* 2017;195:906–11.
17. Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research - commentary. *Biomed Eng Online* 2014;13:94.
18. Nachimuthu SK, Haug PJ. Early detection of sepsis in the emergency department using dynamic bayesian networks. *AMIA Annu Symp Proc* 2012;2012:653–62.
19. Stanculescu I, Williams CK, Freer Y. A Hierarchical Switching Linear Dynamical System Applied to the Detection of Sepsis in Neonatal Condition Monitoring. *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI)*. 2014;752–61.
20. Stanculescu I, Williams CK, Freer Y. Autoregressive hidden Markov models for the early detection of neonatal sepsis. *IEEE J Biomed Health Inform* 2014;18:1560–70.
21. Dyagilev K, Saria S. Learning (predictive) risk scores in the presence of censoring due to interventions. *Mach Learn* 2016;102:323–48.
22. Sawyer AM, Deal EN, Labelle AJ, *et al*. Implementation of a real-time computerized sepsis alert in non-intensive care unit patients. *Critical Care Medicine* 2011;39.3:469–73.
23. Rivers E, Nguyen B, Havstad S, *et al*. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med* 2001;345:1368–77.
24. Yealy DM, Kellum JA, Huang DT, *et al*. A randomized trial of protocol-based care for early septic shock. *N Engl J Med* 2014;370:1683–93.
25. Singer M. Antibiotics for sepsis: does each hour really count, or is it incestuous amplification? *Am J Respir Crit Care Med* 2017;196:800–2.
26. Chang J, Sullivan M, Shea E, *et al*. *The effectiveness of a real-time electronic alert to detect severe sepsis in an intensive care unit. Poster 13 presented at: SOCCA 28th Annual Meeting and Critical Care Update*. 2015. Honolulu HI: URL:https://socca.org/2015-SOCCA-syllabus.pdf (accessed 22 Feb 2017).
27. Dellinger RP, Levy MM, Rhodes A, *et al*. Surviving sepsis campaign: international guidelines for management of severe Sepsis and Septic Shock, 2012. *Intensive Care Med* 2013;39:165–228.
28. Calvert J, Mao Q, Rogers AJ, *et al*. A computational approach to mortality prediction of alcohol use disorder inpatients. *Comput Biol Med* 2016;75:74–9.
29. Calvert J, Mao Q, Hoffman JL, *et al*. Using electronic health record collected clinical variables to predict medical intensive care unit mortality. *Ann Med Surg* 2016;11:52–7.
30 Hooper MH, Weavind L, Wheeler AP, *et al*. Randomized trial of automated, electronic monitoring to facilitate early detection of sepsis in the intensive care unit. *Crit Care Med* 2012;40:2096–101.
31. Semler MW, Weavind L, Hooper MH, *et al*. An electronic tool for the evaluation and treatment of Sepsis in the ICU: a randomized controlled trial. *Crit Care Med* 2015;43:1595.
32. Manaktala S, Claypool SR. Evaluating the impact of a computerized surveillance algorithm and decision support system on sepsis mortality. *J Am Med Inform Assoc* 2017;24:88–95.
33. Berger T, Birnbaum A, Bijur P, *et al*. A computerized alert screening for severe sepsis in emergency department patients increases lactate testing but does not improve inpatient mortality. *Appl Clin Inform* 2010;1:394–407.
34. Cvach M. Monitor alarm fatigue: an integrative review. *Biomed Instrum Technol* 2012;46:268–77.