A targeted real-time early warning score (TREWScore) for septic shock

4 authors, including:

David N Hager
Johns Hopkins University
145 PUBLICATIONS   **7,266** CITATIONS

Peter J Pronovost
Johns Hopkins Medicine
648 PUBLICATIONS   **58,117** CITATIONS

Suchi Saria
Johns Hopkins University
109 PUBLICATIONS   **5,155** CITATIONS

SEPSIS

A targeted real-time early warning score (TREWScore) for septic shock

Katharine E. Henry,[1] David N. Hager,[2] Peter J. Pronovost,[3,4,5] Suchi Saria[1,3,5,6]*

Sepsis is a leading cause of death in the United States, with mortality highest among patients who develop septic shock. Early aggressive treatment decreases morbidity and mortality. Although automated screening tools can detect patients currently experiencing severe sepsis and septic shock, none predict those at greatest risk of developing shock. We analyzed routinely available physiological and laboratory data from intensive care unit patients and developed "TREWScore," a targeted real-time early warning score that predicts which patients will develop septic shock. TREWScore identified patients before the onset of septic shock with an area under the ROC (receiver operating characteristic) curve (AUC) of 0.83 [95% confidence interval (CI), 0.81 to 0.85]. At a specificity of 0.67, TREWScore achieved a sensitivity of 0.85 and identified patients a median of 28.2 [interquartile range (IQR), 10.6 to 94.2] hours before onset. Of those identified, two-thirds were identified before any sepsis-related organ dysfunction. In comparison, the Modified Early Warning Score, which has been used clinically for septic shock prediction, achieved a lower AUC of 0.73 (95% CI, 0.71 to 0.76). A routine screening protocol based on the presence of two of the systemic inflammatory response syndrome criteria, suspicion of infection, and either hypotension or hyperlactatemia achieved a lower sensitivity of 0.74 at a comparable specificity of 0.64. Continuous sampling of data from the electronic health records and calculation of TREWScore may allow clinicians to identify patients at risk for septic shock and provide earlier interventions that would prevent or mitigate the associated morbidity and mortality.

INTRODUCTION

Seven hundred fifty thousand patients develop severe sepsis and septic shock in the United States each year. More than half of them are admitted to an intensive care unit (ICU), accounting for 10% of all ICU admissions, 20 to 30% of hospital deaths, and $15.4 billion in annual health care costs (1–3). Several studies have demonstrated that morbidity, mortality, and length of stay are decreased when severe sepsis and septic shock are identified and treated early (4–8). In particular, one study showed that mortality from septic shock increased by 7.6% with every hour that treatment was delayed after the onset of hypotension (9).

More recent studies comparing protocolized care, usual care, and early goal-directed therapy (EGDT) for patients with septic shock suggest that usual care is as effective as EGDT (10–12). Some have interpreted this to mean that usual care has improved over time and reflects important aspects of EGDT, such as early antibiotics and early aggressive fluid resuscitation (13). It is likely that continued early identification and treatment will further improve outcomes. However, the best approach to managing patients at high risk of developing septic shock before the onset of severe sepsis or shock has not been studied. Methods that can identify ahead of time which patients will later experience septic shock are needed to further understand, study, and improve outcomes in this population.

General-purpose illness severity scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE II), Simplified

Acute Physiology Score (SAPS II), Sequential Organ Failure Assessment (SOFA) scores, Modified Early Warning Score (MEWS), and Simple Clinical Score (SCS) have been validated to assess illness severity and risk of death among septic patients (14–17). Although these scores are useful for predicting general deterioration or mortality, they typically cannot distinguish with high sensitivity and specificity which patients are at highest risk of developing a specific acute condition.

The increased use of electronic health records (EHRs), which can be queried in real time, has generated interest in automating tools that identify patients at risk for septic shock (18–20). A number of "early warning systems," "track and trigger" initiatives, "listening applications," and "sniffers" have been implemented to improve detection and timeliness of therapy for patients with severe sepsis and septic shock (18, 20–23). Although these tools have been successful at detecting patients currently experiencing severe sepsis or septic shock, none predict which patients are at highest risk of developing septic shock.

The adoption of the Affordable Care Act has added to the growing excitement around predictive models derived from electronic health data in a variety of applications (24), including discharge planning (25), risk stratification (26, 27), and identification of acute adverse events (28, 29). For septic shock in particular, promising work includes that of predicting septic shock using high-fidelity physiological signals collected directly from bedside monitors (30, 31), inferring relationships between predictors of septic shock using Bayesian networks (32), and using routine measurements for septic shock prediction (33–35). No current prediction models that use only data routinely stored in the EHR predict septic shock with high sensitivity and specificity many hours before onset. Moreover, when learning predictive risk scores, current methods (34, 36, 37) often have not accounted for the censoring effects of clinical interventions on patient outcomes (38). For instance, a patient with severe sepsis who received fluids and never developed septic shock would be treated as a negative case, despite the possibility that he or she might have developed septic shock in the absence of such treatment and therefore could be considered a positive case up until the

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA. [2]Division of Pulmonary and Critical Care Medicine, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA. [3]Armstrong Institute for Patient Safety and Quality, Johns Hopkins University, Baltimore, MD 21202, USA. [4]Department of Anesthesiology and Critical Care Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD 21202, USA. [5]Department of Health Policy and Management, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA. [6]Department of Applied Math and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA.
*Corresponding author. E-mail: ssaria@cs.jhu.edu

time of treatment (*38*). Methods that assume that these patients are negative and do not account for the uncertainty in the outcome due to censoring can yield scores that underestimate the probability of a positive outcome (*38*, *39*).

Using supervised learning, a machine learning methodology, and the MIMIC (Multiparameter Intelligent Monitoring in Intensive Care)–II Clinical Database (*40*), we trained a model that accounts for the censoring effects of clinical interventions on patient outcomes. We used this model to develop and validate a targeted real-time early warning score (TREWScore) that identifies those patients at high risk of developing septic shock in the future. The ability of this score to identify patients before the onset of septic shock and sepsis-related organ failure (severe sepsis) was then compared to two recently used approaches: first, MEWS, a severity score, originally developed for ICU triage in surgical patients, that has been used for sepsis screening (*41*, *42*) and second, a routine septic shock screening protocol (*18*, *20*) that identifies patients who have at least two of the systemic inflammatory response syndrome (SIRS) criteria, suspicion of infection, and either hypotension or hyperlactatemia.

RESULTS

Characterization of TREWScore
TREWScore is obtained by training a model that estimates the time to an adverse event using supervised learning. The model considered 54 potential features that were derived from routinely available measurements in the EHR. The complete list of features is provided in the Supplementary Materials and Methods. The learning algorithm automatically selected a subset of the features that were most indicative of septic shock and learned a set of weights for them (table S1). The features at each time point were labeled by the time to onset, the number of hours until the onset of septic shock (Fig. 1A), and used to generate TREWScore risk predictions over time (Fig. 1B).

In the following results, we refer to a patient as positive if the patient developed septic shock during his or her stay and negative if the patient never developed septic shock and did not receive characteristic treatment for septic shock, namely, a fluid
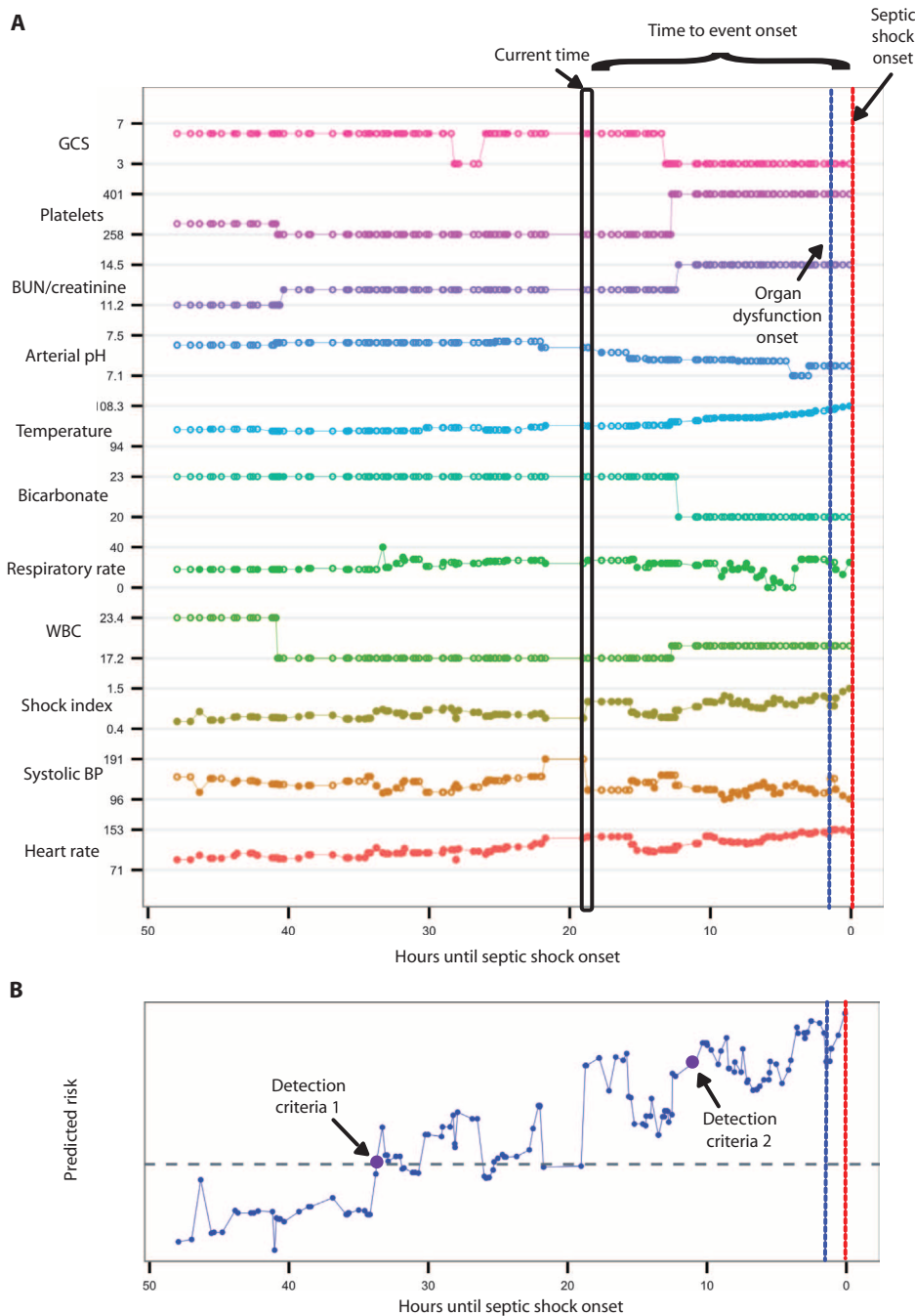
Fig. 1. Example patient features and risk trajectory. (**A**) Example features over time are shown for a patient developing septic shock (time of shock onset indicated by the red line). Point in time data used to calculate TREWScore are displayed in the black box, along with the associated time to onset and the onset of sepsis-related organ dysfunction (indicated by the blue line). Feature measurements are indicated by circles that are filled for new observations or hollow otherwise. Features displayed are Glasgow Coma Scale (GCS), platelets, ratio of blood urea nitrogen to creatinine (BUN/creatinine), arterial pH, temperature, bicarbonate, respiratory rate (RR), white blood cell count (WBC), heart rate/systolic blood pressure (SBP) (shock index), SBP, and heart rate. (**B**) The TREWScore over time for the patient in (A) is shown in blue. Risk predictions are made as new measurements are added to the EHR, as if in real time. The horizontal dashed gray line indicates the detection threshold corresponding to a sensitivity of 0.85. The figure portrays two sets of potential detection criteria: (i) Identify the patient as at high risk of septic shock the first time the risk score crosses the detection threshold. (ii) Identify the patient only after the risk score remains above the detection threshold for at least 8 hours or some other desired length of time.
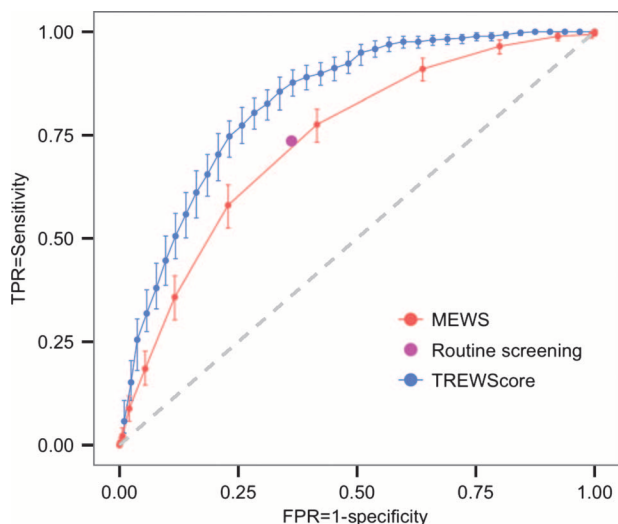
Fig. 2. ROC for detection of septic shock before onset in the validation set. The ROC curve for TREWScore is shown in blue, with the ROC curve for MEWS in red. The sensitivity and specificity performance of the routine screening criteria is indicated by the purple dot. Normal 95% CIs are shown for TREWScore and MEWS. TPR, true-positive rate; FPR, false-positive rate.

bolus of at least 500 ml. We call the patient censored due to clinical intervention if one of two situations masking the true time of septic shock onset occurred. First, if a patient with severe sepsis received characteristic treatment for septic shock but never developed shock, it is unknown whether he or she would have developed septic shock in the absence of such treatment (38). We refer to these patients as right-censored after treatment. Second, if a patient both experienced septic shock and received treatment for septic shock before shock onset, it is unknown whether the treatment delayed the onset of septic shock and to what extent. Thus, in these patients, the exact time of septic shock onset, had the patient not received treatment, could have been any point between the time of treatment and the time of shock onset. We refer to these patients as interval-censored (see Materials and Methods for more details).

Identification of patients at risk for septic shock
We first considered the performance of TREWScore at identifying patients at risk for septic shock before the onset of shock. The data set consisted of adult patients from the MIMIC-II Clinical Database (40). The patients were randomly assigned to either the development or the validation set. See table S2 for population characteristics. The Materials and Methods section provides more details on the development and validation sets.

Among the 13,014 patients (1836 positive, 11,178 negative) in the development set where the final outcome was known, TREWScore identified patients before the onset of septic shock with an area under the receiver operating characteristic (ROC) curve (AUC) of 0.82 [95% confidence interval (CI), 0.81 to 0.83]. As described in the Materials and Methods, patients who were right-censored after treatment were excluded when computing the evaluation metrics because their final outcome was unknown, but their data were used in estimating model coefficients. Model coefficients obtained from this set were fixed and then applied to data from the 3011 patients (455 positive, 2556 negative) in the validation set as though they were observed prospectively. Specifically, for each patient in the validation set, the TREWScore was recom-

puted as new data became available. A patient was identified as at risk when his or her score crossed the specified risk threshold. In the validation set, the AUC obtained for the TREWScore was 0.83 (95% CI, 0.81 to 0.85) (Fig. 2). At a specificity of 0.67 [false-positive rate (FPR) of 0.33], TREWScore achieved a sensitivity of 0.85. Patients were identified a median of 28.2 hours (IQR, 10.6 to 94.2) before shock onset (Fig. 3A).

Identification of patients before organ dysfunction
A critical event in the development of septic shock is the onset of sepsis-related organ dysfunction (severe sepsis), because mortality rates have been shown to increase after this point (1, 9). At a sensitivity of 0.85, more than two-thirds (68.8%) of the patients identified by TREWScore were identified before any sepsis-related organ dysfunction. These patients were identified a median of 7.43 hours (IQR, 2.53 to 25.4) prior (Fig. 3B).

Comparison of TREWScore to other identification methods
We evaluated the performance of two recently advocated sepsis-screening methods for the purpose of providing a comparative analysis of the clinical use of TREWScore. We first compared the performance of TREWScore to MEWS, a general metric used to identify medical patients at high risk of catastrophic deterioration (17). Although it was not specifically developed for tracking sepsis, MEWS has been used to facilitate the identification of patients at risk for severe sepsis and septic shock (41, 42). MEWS assigns points based on RR, heart rate, SBP, temperature, and the Alert, Voice, Pain, Unresponsive (AVPU) score (43). Because the AVPU scores are not routinely computed in the ICU, we mapped the GCS (another neurologic score that is routinely computed in the ICU) to AVPU values (44). Compared to the AUC of 0.83 (95% CI, 0.81 to 0.85) achieved by TREWScore, MEWS identified septic shock patients before onset with a lower AUC of 0.73 (95% CI, 0.71 to 0.76) (Fig. 2).

We next compared the performance of TREWScore against a routine screening protocol for septic shock where a patient was identified as at risk for septic shock if he or she met at least two of the SIRS criteria, had suspicion of infection, and had either hypotension or hyperlactatemia (18, 20). In the validation set, the screening protocol identified patients before the onset of septic shock with a specificity of 0.64 (FPR, 0.36) and a sensitivity of 0.74. In comparison, at a similar specificity of 0.67, TREWScore achieved a much higher sensitivity of 0.85 (Fig. 2).

Because the routine screening protocol is more commonly implemented to identify patients at risk for septic shock, we elaborate on its performance compared to that of TREWScore at a comparable specificity of 0.67 (sensitivity, 0.85). Of all the patients with septic shock in the validation set, most of them (60.9%) were first identified by TREWScore, and fewer than a quarter of them (21.8%) were first identified by the routine screening protocol (Fig. 3A). The remaining patients were either identified by both systems at the same time (9.9%) or not identified by either system before the onset of septic shock (7.5%). As explained in Fig. 3C, positive patients in Fig. 3A are shown in gray from the time of ICU admission until they are identified by one or both of the systems, at which point the color changes to indicate the identifying system(s). This allows a qualitative comparison of the relative detection times.

We further compared the two approaches on the basis of the number of detections that occurred before organ dysfunction (Fig. 3B). At comparable specificities, TREWScore identified 99 more patients before the onset of sepsis-related organ dysfunction (Fig. 3B). Notably, this is a 58.6% increase in the total number of patients identified before organ dysfunction.
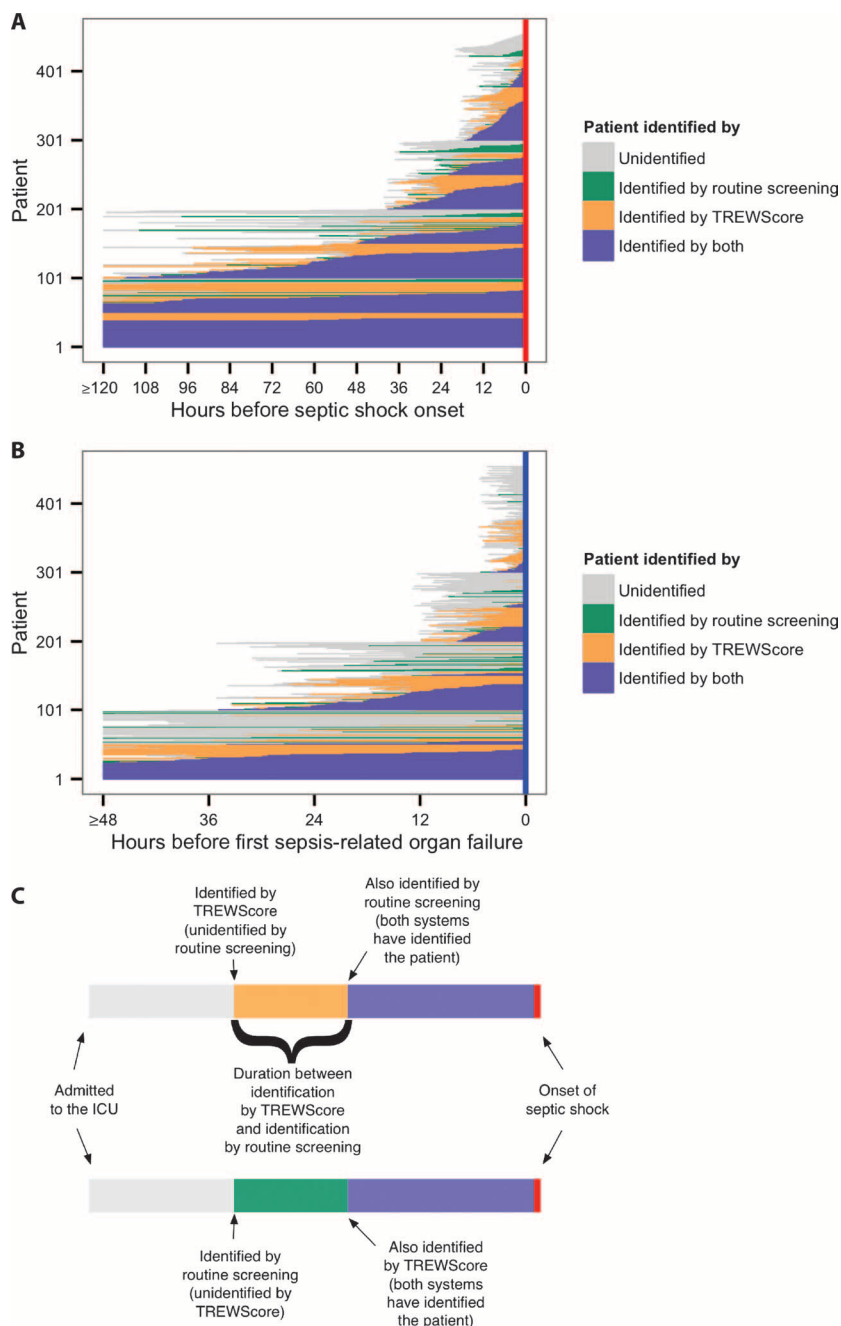
Fig. 3. Comparison of prediction performance. (**A**) Each row represents a septic shock patient in the validation set from the time of ICU admission to the onset of septic shock. The graph was truncated to only show time points within 120 hours of septic shock onset. Individuals were aligned on the basis of time to septic shock after ICU admission. For visual clarity, we further subsorted individuals with similar time-to-shock by time of TREWScore identification. (**B**) Identification times for patients are shown from up to 48 hours before organ dysfunction until the onset of sepsis-related organ dysfunction (blue line). Patients were sorted by time to organ dysfunction and then for visual clarity, patients with similar times until the onset of organ dysfunction were subsorted by the time of identification by TREWScore. (**C**) Each row depicts a patient from the time of ICU admission (left edge) until the time of septic shock (red line). The individual's data are shown in gray from the time of admission until first identification by either system. The bar then becomes orange if the patient was first identified by TREWScore or green if the patient was first identified by the routine screening protocol. This color continues unless the patient is later identified by the second system, at which point the bar becomes purple. If a patient is simultaneously identified by both systems, then the bar transitions directly from gray to purple.

DISCUSSION

Using routinely collected measurements and laboratory results from the MIMIC-II Clinical Database, we developed and validated "TREWScore," a targeted real-time early warning score for septic shock. Surveillance in the ICU is particularly important because the ICU population is fragile and sepsis that goes unnoticed for a prolonged duration and develops into septic shock can have catastrophic consequences. At comparable specificities, TREWScore achieved a higher sensitivity than did either a routine screening protocol frequently used to initiate treatment for severe sepsis and septic shock (*18*, *20*) or MEWS, a general severity score that has been used to identify sepsis (*41*, *42*). Diagnostic tools with low sensitivity are problematic because they will fail to identify many patients likely to have the outcome of interest, limiting the number of patients benefiting from early therapy.

Other investigators have implemented screening tools within the EHR based on criteria established in the Surviving Sepsis Campaign (SSC) guidelines to detect patients at various stages of the sepsis syndrome (*18*, *20*, *21*, *45*). For example, Herasevich *et al.* implemented a set of automated screening criteria (sniffer) to detect severe sepsis and septic shock (*18*). Nguyen *et al.* developed an alert based on the presence of two of the four SIRS criteria and hypotension or hypoperfusion (*20*). Recently, Umscheid *et al.* implemented a similar tool based on a six-point scoring system that included the SIRS criteria, hypotension, and hypoperfusion (*21*). Although these tools have successfully identified patients with severe sepsis and septic shock, their reliance on measures of organ dysfunction as key features limits their ability to identify patients before the onset of sepsis-related organ dysfunction. This is potentially problematic because organ dysfunction may occur suddenly and only briefly precede the onset of septic shock. The window of opportunity to intervene in such cases and mitigate or prevent shock is small or absent (*6*, *9*).

When compared to an example routine screening tool (similar to those discussed above), TREWScore showed a 58.6% increase in the number of patients identified before any sepsis-related organ failure. One factor contributing to TREWScore's ability to identify patients at risk for septic shock early is that unlike the above

approaches, it does not rely solely on SSC guideline–based features. Risk factors and associated weights are dependent on data patterns and features that are automatically chosen (learned) by the algorithm. TREWScore is calculated using 27 features automatically computed from routinely collected measurements in the EHR. Although most are direct patient measurements (SBP, RR, BUN, etc.), some reflect the input of expert opinion (SOFA score and its components and SIRS criteria). Still others, such as the shock index and the BUN/creatinine ratio, are indices that have previously been informative in characterizing severity of illness and predicting outcomes (46–48). These features were retained by the algorithm because they strengthened the model. Other features that did not improve the model were rejected. In this way, the algorithm had the opportunity to capitalize on expert opinion and strengthen the model with additional data-driven features. The retained features have face validity in that aberrations would seemingly occur in a patient progressing toward septic shock.

Previous studies have used learning algorithms and data from the EHR to develop tools related to early identification of septic shock. Hug developed a model to predict hypotension in patients with SIRS (49). Thiel et al. demonstrated that routine measurements could identify septic shock patients a few hours before septic shock but were only able to achieve a sensitivity of 55% (34). Shavdia developed a model that identified patients who later developed septic shock, but the study sample size was small, containing only 26 patients who had hypotension despite fluid resuscitation (36). Ho et al. showed improvement over the techniques presented by Shavdia by using a missing data imputation method; however, their study only evaluated performance on predictions made a few hours before septic shock onset (37). In each of these studies, the time until the onset of septic shock was used to define whether a patient was at risk for septic shock at a particular time point. However, when developing a predictive model for septic shock, none of the above methods accounted for the potential censoring effects of clinical treatments on the time of septic shock. For example, the outcome in patients who did not experience septic shock but received a fluid bolus, a characteristic treatment for septic shock, is ambiguous. These patients can be regarded as positive cases, where the fluid bolus prevented the development of septic shock. However, they can also be regarded as negative cases, where the fluid bolus was an instance of overtreatment. It is not possible to know for certain what the patient's outcome would have been in the absence of treatment. Even among patients who experience septic shock, treatment may delay the onset of septic shock and affect the estimation of time until the event. As further described in the Materials and Methods section, TREWScore addresses these challenges by incorporating uncertainty about the true label when learning the predictive model.

A key clinical implication of identifying patients at high risk for progressing to septic shock early is the potential to reliably improve sepsis treatment and patient outcomes. Earlier efforts to identify and manage patients with sepsis and sepsis-related complications have focused on achieving compliance with sepsis bundle components within 3 to 6 hours of severe sepsis or septic shock onset (6–8). However, timely bundle compliance remains a recurring challenge in early intervention studies (7, 8). A tool like TREWScore that identifies at-risk patients early should provide caregivers greater opportunity to intervene before or at the time of clinical deterioration.

An important related issue to consider is the extent to which clinicians might respond to a warning triggered by a TREWScore indicating that a patient is at high risk of progressing to septic shock. Currently,

this is not known but could be related to factors independent of the predictive value of the score. These factors include the frequency with which care providers are notified about at-risk patients and the mechanism by which the warning would be conveyed. Before deployment of a decision support tool like TREWScore, investigation of these issues, the intended work environment, and other human factors is warranted so that alarm fatigue is avoided and the value of the tool can be optimized in the clinical setting (50). The approach to warning care providers can be tailored to the end user. An alert could be sent each time a patient's score exceeds a certain threshold or only after the threshold is consistently exceeded for some predetermined period of time (as illustrated in Fig. 1B). Alternatively, notification could be restricted to a frequency deemed acceptable by the end users. Further, although our results are reported at a sensitivity of 0.85 (specificity, 0.67), the risk threshold can be adjusted by accepting higher or lower specificity.

It is also important to acknowledge that the benefits of initiating treatment many hours before the onset of septic shock or even the onset of sepsis-related organ failure have yet to be studied, in part because of the lack of a highly sensitive and specific tool, like TREWScore, to identify patients at risk for septic shock. Identifying such patients well before the onset of septic shock, and in most cases before the onset of sepsis-related organ dysfunction, would allow consideration of earlier clinical assessments, diagnostic tests, therapeutic interventions, and transfers to higher or lower levels of care. For example, some care providers may be motivated to obtain early culture data and imaging studies to identify sources of infection. They may even start antibiotics empirically. Others may contemplate the benefits of central line and Foley catheter removal to limit infection risk versus the value these devices may add to managing sepsis when it does occur. Still others may simply keep patients identified by TREWScore in the ICU longer than planned. The downstream effects on quality of care and resource allocation of these different approaches and actions are not known. However, they can only be systematically studied if high-risk patients can be identified accurately. This could lead to new treatment strategies and approaches to triage and bed utilization and potentially further decrease the mortality caused by septic shock.

There are several limitations to this study. First, TREWScore is currently validated only to measure detection performance. A prospective study evaluating whether and how the availability of TREWScore can affect therapeutic judgments is needed. Second, the MIMIC-II database, albeit large, reflects patients admitted to the Beth Israel Deaconess Medical Center between 2001 and 2007 (40); additional validation with data from other hospitals is needed. The MIMIC-II database is developed from medical, surgical, and cardiac ICUs (see table S2 for population details), but how well it will perform in a specific type of ICU has not been characterized. Further, the management of sepsis has likely evolved since this cohort was established in 2001 (13), although recent work continues to suggest that patients with sepsis are frequently identified late (7, 8, 51). Third, some of the features in the model are defined by ICD-9 (International Classification of Diseases, Ninth Revision) codes. The sensitivity and specificity of these codes are very diagnosis-dependent (52). For example, one study found that ICD-9 codes for severe sepsis and septic shock were undercoded among patients admitted to the hospital with a confirmed diagnosis of severe sepsis or septic shock (53). Moreover, coding practices were biased to more frequently code more severe cases (53). This limitation can often be overcome by extracting diagnosis-related information from the discharge notes using automated techniques (54). Last, our definition for septic shock

includes hypotension refractory to volume resuscitation with ≥20 ml/kg over the preceding 24 hours. Most studies have required the use of similar volumes to define resuscitation, but over a shorter interval, to define septic shock. Despite this limitation, the mortality among those with septic shock in this study was 39.3%, which is consistent with previously reported mortality rates (55).

In summary, our study showed that a TREWScore could predict, many hours before standard screening protocols, patients at high risk of developing septic shock. TREWScore used only measurements routinely collected in the EHR and accounts for the effects of censoring due to treatment in estimating the model. In most of these cases, TREWScore identified those at highest risk of developing septic shock long before any evidence of organ dysfunction. Although further studies are needed at other institutions to establish generalizability of the proposed tool, the high performance of TREWScore using a large and heterogeneous cohort (single center but multiple different ICUs) indicates that data-driven early warning scores can be powerful tools for adverse-event prediction. When they are coupled with evidence-based therapies and performance improvement initiatives, there is substantial potential to improve patient outcomes and help make real the vision of learning health care systems.

MATERIALS AND METHODS

Study design
We applied our method to the MIMIC-II Clinical Database (40), a publicly available data set of deidentified EHRs collected at Beth Israel Deaconess Medical Center in Boston, MA. The data set contains all patients admitted to ICUs, including medical, surgical, and cardiac units, between 2001 and 2007 (table S2). We identified 16,234 distinct patients age 15 years or greater at ICU admission with at least one assessment each of GCS, BUN, hematocrit, and heart rate recorded in the EHR.

On the basis of the SSC guidelines, a patient was identified as having SIRS if any two SIRS criteria were present simultaneously. Suspicion of infection was defined using ICD-9 codes indicating infection as in Angus et al. (1) or by the presence of a clinical note that mentioned sepsis or septic shock. Patients with SIRS and suspicion of infection were considered to have sepsis. Patients with sepsis who also had sepsis-related organ dysfunction were defined to have severe sepsis. We defined organ dysfunction due to sepsis using the criteria specified in the SSC guidelines (45) (see Supplementary Materials and Methods for a list of the criteria used).

Patients with septic shock were defined as those who met the criteria for severe sepsis, had hypotension, defined by SBP less than 90 mmHg for at least 30 min, and received adequate fluid resuscitation, defined as the total fluid replacement per kilogram over the past 24 hours ≥20 ml or total fluid replacement ≥1200 ml (48). Of the 16,232 patients in the data set, 2291 patients (14.1%) met the criteria for septic shock. We refer to patients who developed septic shock during their ICU stay as positive cases.

Patients with severe sepsis who received a fluid bolus of at least 500 ml, a characteristic treatment to prevent septic shock, were considered to have a censored outcome. The administration of treatment can affect a patient's outcome by either delaying the onset of septic shock or by preventing the patient from ever developing septic shock. In the case where the patient developed septic shock despite treatment, we considered the patient to be interval-censored, which is to say that it is unknown whether treatment delayed the onset of septic shock and to what extent

(38, 39). Therefore, the exact time of shock onset in the absence of treatment could have occurred at any point between the time of treatment and the observed time of septic shock. Although interval censoring needs to be accounted for differently in model development, for validation, these patients are still considered positive because they are ultimately known to develop septic shock (38).

Alternatively, some patients received treatment and never developed septic shock. However, it is unknown whether these patients would have developed septic shock without treatment (38). In the absence of treatment, the patient may have developed septic shock at any time point after the time of treatment or may never have developed septic shock. We call these patients right-censored after treatment. Although these patients can add information to the development of the model, we cannot report accurate model performance on them because of the uncertainty of whether they would have developed septic shock in the absence of treatment (38). Therefore, our reported validation performance does not include these patients. Patients who never developed septic shock and did not receive treatment characteristic for shock are referred to as negative cases.

We separated the data set into development and validation sets using random sampling. The development set consisted of 13,181 patients (1836 positive, 11,178 negative, and 167 patients with right censoring after treatment). The validation set consisted of 3053 patients (455 positive, 2556 negative, and 42 patients with right-censored after treatment). The development set was used to develop the TREWScore. The validation set was put aside for evaluating performance.

Model development
To develop TREWScore, the following steps were taken using the development set: (i) patient-specific measurement streams were processed to compute features (candidate risk factors); and (ii) the coefficients used in the targeted early warning score were estimated using a supervised learning algorithm. The learning algorithm automatically selected the features that were predictive of septic shock, and the resulting output was a model containing the list of predictive features and their coefficients. Below, we describe model development and evaluation in more detail. Further details on feature processing and computation are given in the Supplementary Materials.

Model development: Estimating model coefficients
To develop a model for predicting an individual's risk of developing septic shock, we fit a Cox proportional hazards model using the time until the onset of septic shock as the supervisory signal. Intuitively, this approach assumes that at times approaching the onset of shock, the sepsis severity level is worse than at times well before the onset. The risk of shock at a time t given the features X at that time, denoted by $\lambda(t|X)$, is computed from two parts: a time-varying baseline hazard function, λ_0, that computes the instantaneous probability that the onset of septic shock occurs at time t and a second term that weights an individual's feature values at time t by learned regression coefficients β (see equation below) (56).

$$\lambda(t|X) = \lambda_0(t)\, exp^{\{X\beta^T\}}$$

A key challenge with training this model is the presence of unknown or censored event times (38, 39). Censoring occurs in two ways. Clinical interventions may influence the observed time of septic shock by delaying the onset of septic shock (interval censoring) or by preventing the development of septic shock entirely (right censoring after treatment).

Whereas right censoring after treatment is naturally accounted for by the Cox proportional hazards model, the model does not a priori account for interval censoring. Model parameter estimation for the Cox proportional hazards model in the presence of interval censoring has been addressed using the expectation-maximization algorithm (57) and multiple imputation–based approaches (56). Here, we used the latter because this approach was much less computationally intensive and simpler to implement using available software.

The multiple imputation approach handles censoring by imputing multiple copies of the development data set. On the subjects for whom the time to event is interval-censored, this approach imputes the exact event time within each copy by sampling from the estimated baseline hazard function (56). Each copy is analyzed separately, and then the results are combined using Rubin's equations (58).

To impute the exact event times for each copy, the baseline hazard function was fit using a multiple imputation method (MIICD R package, version 2.0). For computational efficiency, the baseline hazard function was estimated from a subset of 400,000 time-to-event and feature pairs from the development set (59). The resulting baseline hazard function was then used to repeatedly sample the event time for each interval-censored sample and generate N complete copies of the development data set. Individual copies differ only in the imputed event times. For our experiments, we set $N = 100$.

A separate model was trained from each of the N copies of the development data set. Individual time-to-event models were learned as a Cox proportional hazards model with lasso regularization (glmnet R package, version 1.9-8) (60, 61). Using lasso regularization causes the model to automatically select a sparse subset of features that are most predictive of the labeled outcome (62). The regularization parameter, which controls the degree of parsimony in the learned model, was determined to be 0.01 using 10-fold cross-validation on the first sampled data set and was fixed to this value for training the subsequent models.

To predict on data from a new subject, predicted risk values were obtained from each of the N models. The resulting predictions were then combined using Rubin's equations, which compute the final risk value as the average of risk values outputted from each of the N models (56, 58). Combining the Cox proportional hazards model with the multiple imputation approach allows the model to incorporate information from both interval- and right-censored patients.

Model evaluation

Model coefficients obtained from the development set were fixed and applied to patients in the validation set as though they were observed prospectively. Specifically, for each patient in the validation set, as new data became available, the TREWScore was recomputed. This resulted in a point in time risk for septic shock for each individual. An example of the estimated risk trajectory for a sample patient in the 48 hours preceding the onset of septic shock is given (Fig. 1B). For a fixed risk threshold, an individual was identified as being at high risk of septic shock if his or her risk trajectory ever rose above the detection threshold before the onset of septic shock. For this threshold, we calculated sensitivity, the probability of the risk score being above the detection threshold given that the patient has septic shock, and specificity, the probability that the risk score is always below the threshold given that the patient does not have septic shock and did not experience right censoring after treatment.

We computed the sensitivity as the fraction of patients who developed septic shock and were identified as at high risk by the model. The specificity was computed as the fraction of patients who never developed septic shock and were never identified by the model. The computation of specificity excluded all patients who were right-censored after treatment. The ROC curve and the AUC were obtained by varying the threshold that determined which patients were identified by the model as at risk for septic shock. When the data are right-censored and the mechanism of censoring is not independent of time until the event, an adjusted AUC that accounts for this bias may be additionally desirable to measure. The conditional inverse probability of censoring weighting (CIPCW) is a standard approach to do so (63, 64). However, implementing CIPCW requires assuming a distribution for the conditional probability of being uncensored given patient-specific covariates. Because the unadjusted AUC is more commonly reported, we used it for this study.

At a given sensitivity and specificity, for the true-positive cases, we also computed the median time before septic shock onset and the fraction of detections that occurred before any evidence of sepsis-related organ dysfunction.

SUPPLEMENTARY MATERIALS

REFERENCES AND NOTES

1. D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, M. R. Pinsky, Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care. *Crit. Care Med.* **29**, 1303–1310 (2001).
2. *HCUP Facts and Figures: Statistics on Hospital-based Care in the United States, 2009* (Agency for Healthcare Research and Quality, Rockville, MD, 2011); www.hcup-us.ahrq.gov/reports.jsp.
3. G. Kumar, N. Kumar, A. Taneja, T. Kaleekal, S. Tarima, E. McGinley, E. Jimenez, A. Mohan, R. A. Khan, J. Whittle, E. Jacobs, R. Nanchal; Milwaukee Initiative in Critical Care Outcomes Research Group of Investigators, Nationwide trends of severe sepsis in the 21st century (2000–2007). *Chest* **140**, 1223–1231 (2011).
4. E. Rivers, B. Nguyen, S. Havstad, J. Ressler, A. Muzzin, B. Knoblich, E. Peterson, M. Tomlanovich; Early Goal-Directed Therapy Collaborative Group, Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N. Engl. J. Med.* **345**, 1368–1377 (2001).
5. H. B. Nguyen, S. W. Corbett, R. Steele, J. Banta, R. T. Clark, S. R. Hayes, J. Edwards, T. W. Cho, W. A. Wittlake, 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.* **35**, 1105–1112 (2007).
6. F. Sebat, A. A. Musthafa, D. Johnson, A. A. Kramer, D. Shoffner, M. Eliason, K. Henry, B. Spurlock, Effect of a rapid response system for patients in shock on time to treatment and mortality during 5 years. *Crit. Care Med.* **35**, 2568–2575 (2007).
7. A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg, D. Gurka, A. Kumar, M. Cheang, Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit. Care Med.* **34**, 1589–1596 (2006).
8. V. Coba, M. Whitmill, R. Mooney, H. M. Horst, M. M. Brandt, B. Digiovine, M. Mlynarek, B. McLellan, G. Boleski, J. Yang, W. Conway, J. Jordan; The Henry Ford Hospital Sepsis Collaborative Group, Resuscitation Bundle Compliance in severe sepsis and septic shock: Improves survival, is better late than never. *J. Intensive Care Med.* **26**, 304–313 (2011).
9. A. Castellanos-Ortega, B. Suberviola, L. A. García-Astudillo, M. S. Holanda, F. Ortiz, J. Llorca, M. Delgado-Rodríguez, Impact of the Surviving Sepsis Campaign protocols on hospital length of stay and mortality in septic shock patients: Results of a three-year follow-up quasi-experimental study. *Crit. Care Med.* **38**, 1036–1043 (2010).
10. ProCESS Investigators; D. M. Yealy, J. A. Kellum, D. T. Huang, A. E. Barnato, L. A. Weissfeld, F. Pike, T. Terndrup, H. E. Wang, P. C. Hou, F. LoVecchio, M. R. Filbin, N. I. Shapiro, D. C. Angus, A

randomized trial of protocol-based care for early septic shock. *N. Engl. J. Med.* **370**, 1683–1693 (2014).

11. ARISE Investigators; ANZICS Clinical Trials Group, S. L. Peake, A. Delaney, M. Bailey, R. Bellomo, P. A. Cameron, D. J. Cooper, A. M. Higgins, A. Holdgate, B. D. Howe, S. A. Webb, P. Williams, Goal directed resuscitation for patients with early septic shock. *N. Engl. J. Med.* **371**, 1496–1506 (2014).

12. P. R. Mouncey, T. M. Osborn, G. S. Power, D. A. Harrison, M. Z. Sadique, R. D. Grieve, R. Jahan, S. E. Harvey, D. Bell, J. F. Bion, T. J. Coats, M. Singer, J. D. Young, K. M. Rowan; ProMISe Trial Investigators, Trial of early, goal-directed resuscitation for septic shock. *N. Engl. J. Med.* **372**, 1301–1311 (2015).

13. M. M. Levy, Early goal directed therapy: What do we do now? *Crit. Care* **18**, 705 (2014).

14. N. O. Ghanem-Zoubi, M. Vardi, A. Laor, G. Weber, H. Bitterman, Assessment of disease-severity scoring systems for patients with sepsis in general internal medicine departments. *Crit. Care* **15**, R95 (2011).

15. S. L. Barriere, S. F. Lowry, An overview of mortality risk prediction in sepsis. *Crit. Care Med.* **23**, 376–393 (1995).

16. A. E. Jones, S. Trzeciak, J. A. Kline, The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Crit. Care Med.* **37**, 1649–1654 (2009).

17. C. Vorwerk, B. Loryman, T. J. Coats, J. A. Stephenson, L. D. Gray, G. Reddy, L. Florence, N. Butler, Prediction of mortality in adult emergency department patients with sepsis. *Emerg. Med. J.* **26**, 254–258 (2009).

18. V. Herasevich, M. S. Pieper, J. Pulido, O. Gajic, Enrollment into a time sensitive clinical study in the critical care setting: Results from computerized septic shock sniffer implementation. *J. Am. Med. Inform. Assoc.* **18**, 639–644 (2011).

19. *Acutely Ill Patients in Hospital: Recognition of and Response to Acute Illness in Adults in Hospital* (NICE clinical guideline no. 50) (National Institute for Health and Clinical Excellence, London, UK, 2007); www.nice.org.uk/CG050.

20. S. Q. Nguyen, E. Mwakalindile, J. S. Booth, V. Hogan, J. Morgan, C. T. Prickett, J. P. Donnelly, H. E. Wang, Automated electronic medical record sepsis detection in the emergency department. *PeerJ* **2**, e343 (2014).

21. C. A. Umscheid, J. Betesh, C. VanZandbergen, A. Hanish, G. Tait, M. E. Mikkelsen, B. French, B. D. Fuchs, Development, implementation, and impact of an automated early warning and response system for sepsis. *J. Hosp. Med.* **10**, 26–31 (2015).

22. J. Nelson, B. Smith, J. Jared, J. Younger, Prospective trial of real-time electronic surveillance to expedite early care of severe sepsis. *Ann. Emerg. Med.* **57**, 500–504 (2011).

23. R. J. M. Morgan, F. Williams, M. N. Wright, An early warning scoring system for detecting developing critical illness. *Clin. Intens. Care Nurs.* **8**, 100 (1997).

24. D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff.* **33**, 1123–1131 (2014).

25. J. Donzé, D. Aujesky, D. Williams, J. L. Schnipper, Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. *JAMA Intern. Med.* **173**, 632–638 (2013).

26. T. M. Osborn, G. Phillips, S. Lemeshow, S. Townsend, C. A. Schorr, M. M. Levy, R. P. Dellinger, Sepsis Severity Score: An internationally derived scoring system from the Surviving Sepsis Campaign database. *Crit. Care Med.* **42**, 1969–1976 (2014).

27. S. Saria, A. Rajani, J. Gould, D. Koller, A. Penn, Integration of early physiological responses predicts later illness severity in preterm infants. *Sci. Transl. Med.* **2**, 48ra65 (2010).

28. J. Wiens, E. Horvitz, and J. V. Guttag, Patient risk stratification for hospital-associated C. diff as a time-series classification task. *Adv. in Neural Inf. Process. Syst.* **25**, 476–484 (2012).

29. O. Gajic, O. Dabbagh, P. K. Park, A. Adesanya, S. Y. Chang, P. Hou, H. Anderson III, J. J. Hoth, M. E. Mikkelsen, N. T. Gentile, M. N. Gong, D. Talmor, E. Bajwa, T. R. Watkins, E. Festic, M. Yilmaz, R. Iscimen, D. A. Kaufman, A. M. Esper, R. Sadikot, I. Douglas, J. Sevransky, J. Malinchoc; U.S. Critical Illness and Injury Trials Group: Lung Injury Prevention Study Investigators (USCIITG-LIPS), Early identification of patients at risk of acute lung injury: Evaluation of Lung Injury Prediction Score in a multicenter cohort study. *Am. J. Respir. Crit. Care Med.* **183**, 462–470 (2011).

30. I. Stanculescu, C. K. I. Williams, Y. Freer, Autoregressive hidden Markov models for the early detection of neonatal sepsis. *IEEE J. Biomed. Health Inform.* **18**, 1560–1570 (2014).

31. M. P. Griffin, D. E. Lake, T. M. O'Shea, J. R. Moorman, Heart rate characteristics and clinical signs in neonatal sepsis. *Pediatr. Res.* **61**, 222–227 (2007).

32. E. Gultepe, J. P. Green, H. Nguyen, J. Adams, T. Albertson, I. Tagkopoulos, 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.* **21**, 315–325 (2014).

33. K. K. Giuliano, Physiological monitoring for critically ill patients: Testing a predictive model for the early detection of sepsis. *Am. J. Crit. Care* **16**, 122–130 (2007).

34. S. W. Thiel, J. M. Rosini, W. Shannon, J. A. Doherty, S. T. Micek, M. H. Kollef, Early prediction of septic shock in hospitalized patients. *J. Hosp. Med.* **5**, 19–25 (2010).

35. K. Henry, C. Paxton, K. S. Kim, J. Pham, S. Saria, REWS: Real-time early warning score for septic shock. *Crit. Care Med.* **42**, abstract 63 (2014).

36. D. Shavdia, thesis, Septic shock: Providing early warnings through multivariate logistic regression models, Harvard-MIT Division of Health Sciences and Technology (2007).

37. J. C. Ho, C. H. Lee, J. Ghosh, Septic shock prediction for patients with missing data. *ACM Trans. Manag. Inf. Syst.* **5**, Article No. 1 (2014).

38. C. Paxton, A. Niculescu-Mizil, S. Saria, Developing predictive models using electronic medical records: Challenges and pitfalls. *AMIA Annu. Symp. Proc.* **2013**, 1109–1115 (2013).

39. K. Dyagilev, S. Saria, Learning (predictive) risk scores in the presence of censoring due to interventions. arXiv:1507.07295 (2015).

40. A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220 (2000).

41. A. Hendrich, J. Byrd, V. Naylor, T. Daniels, R. Weakland, C. Wagner, *Leading Edge Advanced Practice Topic (LEAPT): Severe Sepsis and Septic Shock* (Health Research & Educational Trust, Chicago, IL, 2014); www.hret-hen.org/index.php?option=com_phocadownload&view=category&id=370&Itemid=369).

42. S. Townsend, T. Clemmer, D. Jacobsen, *IHI Expedition: Treating Sepsis in the Emergency Department and Beyond, Session 6* (Institute for Healthcare Improvement, Cambridge, MA, 2013); www.ihi.org/Engage/Memberships/Passport/Pages/Treating-Sepsis.aspx).

43. C. P. Subbe, M. Kruger, P. Rutherford, L. Gemmel, Validation of a modified Early Warning Score in medical admissions. *QJM* **94**, 521–526 (2001).

44. G. B. Smith, D. R. Prytherch, P. E. Schmidt, P. I. Featherstone, Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation* **77**, 170–179 (2008).

45. R. P. Dellinger, M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, J. E. Sevransky, C. L. Sprung, I. S. Douglas, R. Jaeschke, T. M. Osborn, M. E. Nunnally, S. R. Townsend, K. Reinhart, R. M. Kleinpell, D. C. Angus, C. S. Deutschman, F. R. Machado, G. D. Rubenfeld, S. A. Webb, R. J. Beale, J. L. Vincent, R. Moreno; Surviving Sepsis Campaign Guidelines Committee including the Pediatric Subgroup, Surviving Sepsis Campaign: International guidelines for management of severe sepsis and septic shock: 2012. *Crit. Care Med.* **41**, 580–637 (2013).

46. T. Berger, J. Green, T. Horeczko, Y. Hagar, N. Garg, A. Suarez, E. Panacek, N. Shapiro, Shock index and early recognition of sepsis in the emergency department: Pilot study. *West. J. Emerg. Med.* **14**, 168–174 (2013).

47. H. D. Humes, in *Kelley's Essentials of Internal Medicine* (Lippincott Williams & Wilkins, Philadelphia, ed. 2, 2001), chap. 18.

48. J. S. Rachoin, R. Daher, C. Moussallem, B. Milcarek, K. Hunter, C. Schorr, M. Abboud, P. Henry, L. S. Weisberg, The fallacy of the BUN:creatinine ratio in critically ill patients. *Nephrol. Dial. Transplant.* **27**, 2248–2254 (2012).

49. C. Hug, "Detecting hazardous intensive care patient episodes using real-time mortality models" (MIT Technical Report, 2009).

50. H. van der Sijs, J. Aarts, A. Vulto, M. Berg, Overriding of drug safety alerts in computerized physician order entry. *J. Am. Med. Inform. Assoc.* **13**, 138–147 (2006).

51. W. J. Gu, F. Wang, J. Bakker, L. Tang, J. C. Liu, The effect of goal-directed therapy on mortality in patients with sepsis - earlier is better: A meta-analysis of randomized controlled trials. *Crit. Care* **18**, 570 (2014).

52. J. Campbell, T. Payne, A comparison of four schemes for codification of problem lists. *Proc. Annu. Symp. Comput. Appl. Med. Care* 201–205 (1994).

53. S. A. Whittaker, M. E. Mikkelsen, D. F. Gaieski, S. Koshy, C. Kean, B. D. Fuchs, Severe sepsis cohorts derived from claims-based strategies appear to be biased toward a more severely ill patient population. *Crit. Care Med.* **41**, 945–953 (2013).

54. S. Saria, G. McElvain, A. K. Rajani, A. A. Penn, D. L. Koller, Combining structured and free-text data for automatic coding of patient outcomes. *AMIA Annu. Symp. Proc.* **2010**, 712–716 (2010).

55. M. M. Levy, R. P. Dellinger, S. R. Townsend, W. T. Linde-Zwirble, J. C. Marshall, J. Bion, C. Schorr, A. Artigas, G. Ramsay, R. Beale, M. M. Parker, H. Gerlach, K. Reinhart, E. Silva, M. Harvey, S. Regan, D. C. Angus; Surviving Sepsis Campaign, The Surviving Sepsis Campaign: Results of an international guideline-based performance improvement program targeting severe sepsis. *Crit. Care Med.* **38**, 367–374 (2010).

56. W. Pan, A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56**, 199–203 (2000).

57. Z. Zhang, L. Sun, J. Sun, D. M. Finkelstein, Regression analysis of failure time data with informative interval censoring. *Stat. Med.* **26**, 2533–2546 (2007).

58. R. Little, D. Rubin, *Statistical Analysis with Missing Data* (Wiley Series in Probability and Mathematical Statistics, Wiley, New York, NY, 2002).

59. T. M. Therneau, T. Lumley, The Survival Package (The Comprehensive R Archive Network, 2014).

60. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

61. N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).

62. R. Tibshirani, The lasso method for variable selection in the cox model. *Stat. Med.* **16**, 385–395 (1997).
63. P. J. Heagerty, T. Lumley, M. S. Pepe, Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344 (2000).
64. P. Blanche, J.-F. Dartigues, H. Jacqmin-Gadda, Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biom. J.* **55**, 687–704 (2013).

Editor's Summary

Evening the score against sepsis
Sepsis is a major cause of death, which remains difficult to treat despite modern antibiotics. Early aggressive treatment of this disease improves patient mortality, but the tools currently available in the clinic do not predict who will develop sepsis and its late manifestation, septic shock, until the patients are already in advanced stages of the disease. Henry *et al*. used readily available data from patient monitors and medical records to develop TREWScore, a targeted real-time early warning score that predicts in advance which patients are at risk for septic shock. With a median lead time of over 24 hours, this scoring algorithm may allow clinicians enough time to intervene before the patients suffer the most damaging effects of sepsis.