nature medicine

ARTICLES

https://doi.org/10.1038/s41591-022-01894-0

Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis

Roy Adams[1,2], Katharine E. Henry[2,3], Anirudh Sridharan[4], Hossein Soleimani[5], Andong Zhan[2,3], Nishi Rawat[6], Lauren Johnson[7], David N. Hager[8], Sara E. Cosgrove[8], Andrew Markowski[9], Eili Y. Klein[10], Edward S. Chen[8], Mustapha O. Saheed[10], Maureen Henley[7], Sheila Miranda[11], Katrina Houston[7], Robert C. Linton[4], Anushree R. Ahluwalia[7], Albert W. Wu[6,8,12,13,14] ✉ and Suchi Saria[1,3,8,12,15] ✉

Early recognition and treatment of sepsis are linked to improved patient outcomes. Machine learning-based early warning systems may reduce the time to recognition, but few systems have undergone clinical evaluation. In this prospective, multi-site cohort study, we examined the association between patient outcomes and provider interaction with a deployed sepsis alert system called the Targeted Real-time Early Warning System (TREWS). During the study, 590,736 patients were monitored by TREWS across five hospitals. We focused our analysis on 6,877 patients with sepsis who were identified by the alert before initiation of antibiotic therapy. Adjusting for patient presentation and severity, patients in this group whose alert was confirmed by a provider within 3 h of the alert had a reduced in-hospital mortality rate (3.3%, confidence interval (CI) 1.7, 5.1%, adjusted absolute reduction, and 18.7%, CI 9.4, 27.0%, adjusted relative reduction), organ failure and length of stay compared with patients whose alert was not confirmed by a provider within 3 h. Improvements in mortality rate (4.5%, CI 0.8, 8.3%, adjusted absolute reduction) and organ failure were larger among those patients who were additionally flagged as high risk. Our findings indicate that early warning systems have the potential to identify sepsis patients early and improve patient outcomes and that sepsis patients who would benefit the most from early treatment can be identified and prioritized at the time of the alert

Sepsis is a leading cause of in-hospital death in the United States, with a recent study finding sepsis as the immediate cause of nearly 35% of in-hospital deaths[1]. Effective intervention has been elusive; the current state has been referred to as a 'treatment graveyard'[2] because few effective sepsis treatments have been developed and there is persistent debate about best treatment practices[2,3]. There is agreement that early recognition and treatment are critical to successful outcomes[4]. Early administration of broad-spectrum intravenous antibiotics, in particular, is associated with decreased mortality and morbidity[5–10]. However, heterogeneity in the presentation of sepsis often makes early recognition challenging, and many patients receive delayed care[1]. This has spurred interest in the development of automated sepsis early warning systems to help clinicians recognize sepsis as early as possible.

Several retrospective studies have demonstrated that machine learning (ML)-based models can detect sepsis early[11–15]. However, few studies have reported on clinical implementations of these models[11,16,17]. Although the existing implementation studies of both general[18] and sepsis-specific alert systems[11,16,17,19,20] demonstrate the feasibility of deployment, several of these studies have relied on dedicated staff to respond to alerts and the reported clinical value has been mixed, with clinical suspicion of sepsis often noted before alerts[19,20]. Additional studies are needed to understand the impact of sepsis-specific early warning systems on patient outcomes and to demonstrate the potential of decentralized bedside alert systems.

After 3 years of development, the TREWS ML-based early warning system was deployed[21], starting in 2018 as part of an electronic health record (EHR)-based sepsis alert system in two academic and three community hospitals in the Maryland and DC areas. Details of the model, deployment and workflow are described in a companion paper[22]. Of note, when a TREWS alert occurs, providers may open a dedicated TREWS page in the electronic medical record system (in this case, Epic) and, from that page, may enter an evaluation of the patient as having sepsis or not. In a companion paper, we analyzed the predictive performance of TREWS, the adoption of TREWS by providers and the association between interaction with TREWS and the time between the alert and antibiotic ordering[22]. Notably, we found that sepsis patients who had their alert evaluated and confirmed within 3 h had a 1.85-h lower median time from alert to first antibiotic order[22].

[1]Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA. [2]Department of Psychiatry and Behavioral Science, Johns Hopkins School of Medicine, Baltimore, MD, USA. [3]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [4]Howard County General Hospital, Columbia, MD, USA. [5]Health Informatics, University of California, San Francisco, CA, USA. [6]Armstrong Institute for Patient Safety and Quality, Johns Hopkins School of Medicine, Baltimore, MD, USA. [7]Department of Quality Improvement, Johns Hopkins Hospital, Baltimore, MD, USA. [8]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA. [9]Suburban Hospital, Bethesda, MD, USA. [10]Department of Emergency Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA. [11]Department of Medicine, Johns Hopkins Hospital, Baltimore, MD, USA. [12]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [13]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [14]Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [15]Bayesian Health, New York, NY, USA. ✉e-mail: awu@jhu.edu; ssaria@cs.jhu.edu

In the present study, we analyzed the association between provider interaction with TREWS and patient outcomes for a target population of sepsis patients who had an alert before they received any antibiotics. Using EHR data collected from the five deployment sites, we conducted a prospective, multicenter, two-arm cohort study to evaluate the association between timely provider evaluation and confirmation of the TREWS alert and mortality for a target population of sepsis patients who had an alert before receiving any antibiotics. Additional outcomes of interest are progression in a patient's total sequential organ failure assessment (SOFA) score, following the alert, and length of stay. In a secondary analysis, we further examined this association across patients identified as having increased risk of death in the absence of timely antibiotics, referred to henceforth as the high-risk cohort. This high-risk cohort was defined based on a secondary risk score that uses measurements available at or near the time of alert.

Results

Study population. During the study period, 590,736 patient encounters involving patients over the age of 18 years who presented to an emergency department (ED) or were admitted to an inpatient unit at one of the five deployment hospitals were monitored by TREWS. Of these encounters, 42,089 (7.1%) triggered an alert and 13,680 (2.3%) met the retrospective criteria for sepsis[23]. Among encounters that triggered an alert, 24,799 (59%) had their first alert before admission to an inpatient unit, whereas 17,290 (41%) had their first alert after inpatient admission. A total of 6,877 patient encounters met the inclusion criteria for our target population: triggered an alert at most 1 h before ED triage or inpatient admission, met the retrospective definition for sepsis, received their first antibiotic after the alert and within 24 h of the alert and were not admitted directly to an intensive care unit (ICU). Of these, 2,366 were included in the high-risk cohort. In addition, 2,458 encounters involved patients who had sepsis but did not trigger an alert, and 3,006 encounters involved sepsis patients who received antibiotics before triggering an alert. These two off-target cohorts had lower general severity at admission than the primary analysis cohort (for example, the average Acute Physiology and Chronic Health Evaluation (APACHE II) and SOFA scores were lower in these groups), as well as having a lower in-hospital mortality rate. Figure 1 shows a waterfall diagram for the primary analysis cohort, and Table 1 shows summary statistics for all patients, sepsis patients with no alert, sepsis patients who received antibiotics before triggering an alert and our primary analysis cohort. A breakdown of the primary analysis cohort by hospital can be found in Extended Data Table 1.

Association between response to TREWS and patient outcomes. Table 1 shows select summary statistics for the treatment and comparison groups, and Extended Data Table 2 shows a comparison of all variables. Among included patients, 4,220 (61%) had TREWS used as intended with their alert evaluated and confirmed within 3 h of the alert (study arm) and 2,657 (39%) did not (comparison arm). The full distribution of time from alert occurrence to alert confirmation is shown in Fig. 2. Among other differences, patients in the present study arm were older (67 versus 65 years, $P<0.001$), had lower SOFA scores at the time of the alert (4.0 versus 4.3, $P<0.001$) and were more likely to have a lactate >2 mmol l^{-1} (75% versus 61%, $P<0.001$).

Table 2 shows the unadjusted outcomes and adjusted comparisons of outcomes between study and comparison arms. Patients in the treatment group had lower unadjusted mortality rates (14.6% versus 19.2%, $P<0.001$), improved SOFA score progression (−0.8 versus −0.4, $P<0.001$) and lower median length of stay among survivors (6.6 d versus 8.1 d, $P<0.001$). After adjustment for patient demographics, medical history, laboratory measurements, vital signs, comorbidities and admitting hospital, timely alert confirmation by the provider was associated with lower mortality (adjusted risk difference (ARD) −3.34%, CI −5.10, −1.67%, and adjusted relative reduction (ARR) −18.18%, CI −26.31, −9.65%; $P<0.001$), improved SOFA progression (ARD −0.26; CI −0.42, −0.11; $P=0.001$) and lower median length of stay among survivors (ARD −11.58 h, CI −18.13, −5.03 h; $P=0.001$). In sensitivity analyses, none of the following changed the direction or significance of these associations: changing the estimation method; changing the data source for comorbidities, including only the first encounter for each patient and only encounters before the COVID-19 pandemic (that is, before 31 March 2020); or restricting the list of antibiotics used (Extended Data Table 3).

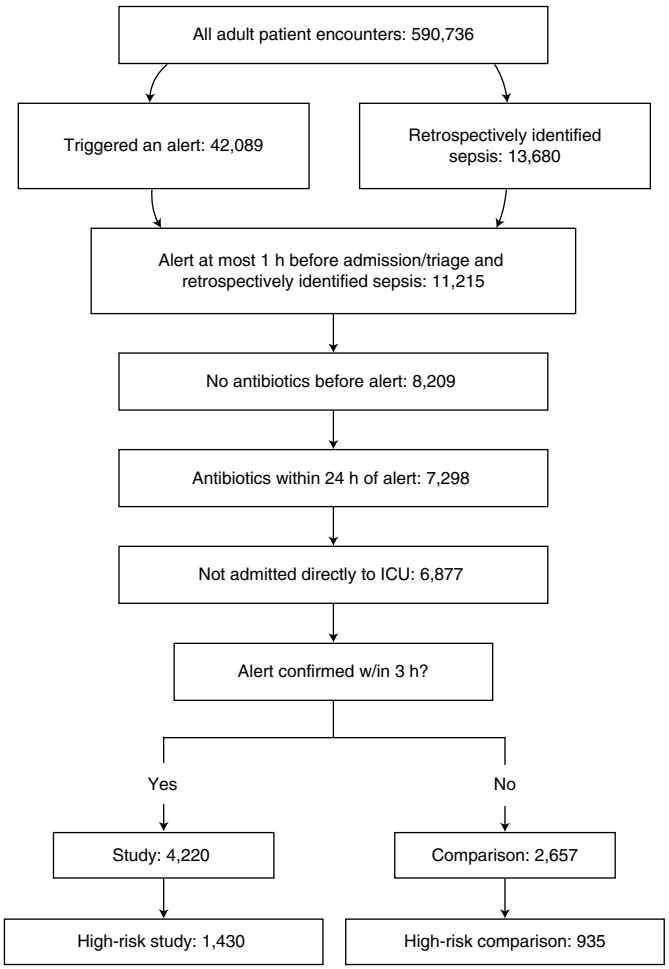

Fig. 1 | Waterfall diagram for the primary analysis cohort. Waterfall diagram describing the primary analysis and high-risk cohorts. A total of 590,736 unique adult patient encounters occurred during the study period. Of these, 6,877 patient encounters were included in our primary analysis, and 2,365 patient encounters were included in our analysis of high-risk patients.

Sepsis is highly heterogeneous[24], and there is evidence that the importance of early antibiotic therapy is similarly heterogeneous[7,25]. Considering this, we sought to estimate the associations between provider interaction with TREWS and patient outcomes among those patients who were predicted, based on baseline variables, to be most sensitive to antibiotic timing. We predicted sensitivity to antibiotic timing as the increase in predicted mortality risk when using a model trained on patients who received timely antibiotics (within 3 h of their alert) versus a model trained on those who did not (see Methods for details). This approach was based in part on existing methods for predicting which patients are likely to respond

Table 1 | Sample statistics

	All patient encounters	Sepsis patients not flagged by the alert	Sepsis patients who received an antibiotic before their alert	Primary analysis cohort ($n=6,877$)			High-risk cohort ($n=2,365$)		
				Alert confirmed within 3 h	Alert not confirmed within 3 h	*P* value[a]	Alert confirmed within 3 h	Alert not confirmed within 3 h	*P* value[a]
No. of encounters	590,736	2,458	3,006	4,220	2,657	–	1,430	935	–
No. of patients	342,650	2,268	2,826	3,925	2,577	–	1,379	923	–
Age (years)	51.7 ± 20.0	62.2 ± 17.6	63.6 ± 18.7	66.6 ± 17.5	65.1 ± 16.8	< 0.001	67.9 ± 16.0	65.1 ± 15.6	< 0.001
Documented men	333,877 (56.5%)	1,130 (46.0%)	1,458 (48.5%)	1,992 (47.2%)	1,239 (46.6%)	0.661	701 (49.0%)	437 (46.7%)	0.296
Admitted from ED	412,163 (69.8%)	1,496 (60.9%)	2,545 (84.7%)	4,148 (98.3%)	2,434 (91.6%)	< 0.001	1,414 (98.9%)	866 (92.6%)	< 0.001
ED discharge	287,264 (48.6%)	16 (0.7%)	14 (0.5%)	62 (1.5%)	24 (0.9%)	0.052	14 (1.0%)	9 (1.0%)	0.862
Ever admitted to ICU	24,254 (4.1%)	1,013 (41.2%)	1,393 (46.3%)	1,811 (42.9%)	1,273 (47.9%)	< 0.001	1,147 (80.2%)	731 (78.2%)	0.254
Mean CCI	2.7 ± 3.2	5.8 ± 3.8	6.2 ± 4.0	6.5 ± 3.9	6.5 ± 4.0	0.372	7.5 ± 4.0	6.9 ± 4.1	< 0.001
At admission[b]									
APACHE II	3.3 ± 3.8	10.3 ± 7.0	11.3 ± 6.8	13.4 ± 8.2	13.7 ± 8.8	0.972	21.3 ± 8.0	21.8 ± 8.4	0.071
SOFA	0.7 ± 1.5	3.5 ± 3.5	3.6 ± 3.2	4.6 ± 3.8	5.0 ± 3.8	< 0.001	8.5 ± 3.5	8.8 ± 3.4	0.014
GCS < 15	44,048 (7.5%)	674 (27.4%)	986 (32.8%)	1,647 (39.0%)	1,085 (40.8%)	0.143	1,053 (73.6%)	710 (75.9%)	0.227
Lactate > 2 mmol l^{-1}	23,033 (3.9%)	628 (25.5%)	1,191 (39.6%)	3,248 (77.0%)	1,711 (64.4%)	< 0.001	1,129 (79.0%)	644 (68.9%)	< 0.001
Sepsis	13,680 (2.3%)	–	–	–	–	–	–	–	–
Septic shock[c]	5,215 (0.9%)	688 (28.0%)	975 (32.4%)	1,813 (43.0%)	1,084 (40.8%)	0.081	1,120 (78.3%)	660 (70.6%)	< 0.001
Had an alert	42,089 (7.1%)	–	–	–	–	–	–	–	–
At alert									
APACHE II	–	–	–	11.7 ± 7.5	11.9 ± 8.1	0.555	19.3 ± 7.4	19.6 ± 8.1	0.567
SOFA	–	–	–	4.0 ± 3.4	4.3 ± 3.4	< 0.001	7.7 ± 3.0	7.8 ± 3.0	0.159
GCS < 15	–	–	–	1,381 (32.7%)	907 (34.1%)	0.237	949 (66.4%)	645 (69.0%)	0.199
Lactate > 2 mmol l^{-1}	–	–	–	3,160 (74.9%)	1,630 (61.3%)	< 0.001	1,099 (76.9%)	595 (63.6%)	< 0.001
Died in hospital	4,610 (0.8%)	318 (12.9%)	392 (13.0%)	617 (14.6%)	509 (19.2%)	< 0.001	422 (29.5%)	320 (34.2%)	0.018

All discrete values are reported as 'count (percentage)' and all continuous values are reported either as 'mean ± s.d.' or 'median (interquartile range (IQR))'. [a]*P* values were based on Pearson's χ^2 and two-sided Wilcoxon's rank-sum tests for categorical and continuous variables, respectively. *P* values were not adjusted for multiple comparisons. [b]'At admission' values were calculated based on measurements taken within 24 h of ED triage or admission. [c]Septic shock was defined as being both positive for sepsis and either receiving vasopressors or having a lactate >4 mmol l^{-1} within 48 h of the first signs of organ dysfunction.

most to a particular treatment[26]. We developed the two mortality risk models using retrospective, pre-deployment data and applied them prospectively to our study population to identify a high-risk cohort of patients. Among the 2,365 patients who fell into this high-risk cohort, alert evaluation and confirmation within 3 h was associated with a greater decrease in mortality (ARD −4.50%, CI −8.31, −0.78% and ARR −13.19%, CI −22.81, −2.45%; $P=0.012$) and SOFA progression (ARD −0.38, CI −0.71, −0.05; $P=0.025$). The association with length of stay was slightly greater in the high-risk cohort than in the full sample, although it was no longer statistically significant. Full results on the high-risk stratum are presented in Table 2.

Antibiotic timing relative to TREWS and patient outcomes. Previous work has established an association between antibiotic timing relative to ED admission and patient mortality[5–8,25]. In the present study, we instead measure the association between antibiotic timing relative to a TREWS alert and patient outcomes. Adjusting for patient demographics, medical history, labs, vital signs, comorbidities and admitting hospital, patients in the target population who received antibiotics within 3 h of their alert had a lower mortality rate (ARD −3.54%, CI −5.52, −1.66%, and ARR −18.69%, CI −27.03, −9.36%; $P<0.001$), SOFA score progression (ARD −0.26, CI −0.42, −0.11; $P=0.001$) and median length of stay (ARD −11.58 h, CI −18.13, −5.03 h; $P=0.001$) compared with those who received antibiotics >3 h after their alert. Associations in the high-risk cohort are presented in Extended Data Table 4. In addition, for the purposes of comparison with previous retrospective studies, we estimated the adjusted association between in-hospital mortality and each hour of delay from the time of the alert to first antibiotics for patients receiving antibiotics within 6 and 12 h of their alert. The 6- and 12-h windows were chosen for comparison to previous literature. The adjusted odds ratios for in-hospital mortality per hour delay in antibiotics were 1.08 (CI 1.02, 1.15) and 1.05 (CI 1.02, 1.09) among patients who received antibiotics within 6 and 12 h, respectively. In the high-risk cohort, the adjusted odds ratios for in-hospital mortality per hour delay in antibiotics were 1.07 (CI 0.99, 1.17) and 1.08 (CI 1.03, 1.14) among patients who received antibiotics within 6 and 12 h, respectively.

Discussion

In the present study, we found that, after adjusting for a range of patient variables, timely provider evaluation and confirmation of the TREWS alert were associated with reduced mortality among target sepsis patients, as well as improved SOFA score progression and reduced median length of stay among survivors. Absolute associations were of greater magnitude among patients prospectively identified as being at high risk for death without timely treatment. In addition, we found that delay in antibiotics relative to the time of the alert was associated with increased patient mortality in our target population. These results complement our companion paper which found that TREWS had high adoption (89% of alerts received

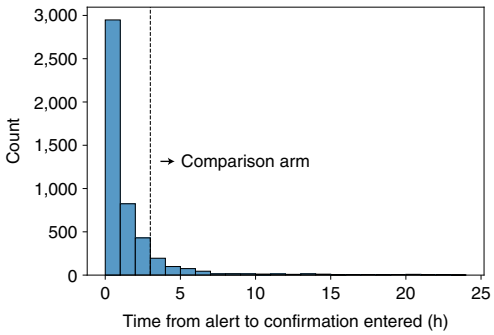

Fig. 2 | Distribution of the time from which the TREWS alert was given to the time confirmation was entered for the target population. Each bar has a width of 1 h, and the bar height represents the number of subjects who had their alert confirmed in that 1-h time bin. The dashed line is placed at 3 h, such that bars to the left and right of the line correspond to subjects in the study and comparison arms, respectively. Note that 1,840 (27%) patients in the target population either did not have an evaluation entered or had their alert dismissed by a provider and thus do not appear in this plot.

an evaluation) and timely provider confirmation of the TREWS alert was associated with a 1.85-h reduction in median time from alert to first antibiotic order[22].

Although studies of deployed sepsis early warning systems have been uncommon, a few studies have found positive benefits when using such systems[16,27–29]. Three independent pre-/post-deployment studies found that sepsis-related mortality dropped after system deployment[27–29]. One major limitation of pre-/post-studies of sepsis-related mortality is the potential for surveillance bias, which occurs when more patients are coded as septic in the post-period due to the alerting tool or concurrent changes in coding practices. Without appropriate adjustments, clinical outcomes appear to have improved because the group of patients coded as septic are less sick than before the alert was implemented. Our study avoided this issue by including only post-deployment data, making use instead of natural variability in provider practice to create a convenience control group[30,31]. The reasons behind this variability in provider practice are explored more deeply in our companion paper[22].

One randomized trial by Shimabukuro et al. found that use of a sepsis alert system led to reductions in all-cause mortality and length of stay[16]. Although limited in size (42 septic patients and 142 total patients) and restricted to ICU patients, Shimabukuro et al. is, to our knowledge, the only randomized trial of a prospective sepsis alerting system. In contrast, our study does not make use of randomization. However, our findings are consistent with those of Shimabukuro et al. and complement their findings by including a larger sample of patients from diverse units across multiple sites.

Studies of two other systems did not demonstrate impact on patient outcomes[11,19,20]. One pre-/post-deployment study found that deployment led to modest but statistically significant increases in lactate ordering and intravenous fluid administration but did not impact antibiotic ordering or patient outcomes[11]. In follow-up interviews, providers reported that the alerts were unlikely to affect their perception of a patient[20]. A similar pre-/post-deployment study found that deployment of an early warning system did not lead to changes in clinical practice or patient outcomes[19]. A notable difference between these systems and TREWS is the accuracy of the alerts. For example, the latter study revealed that only 29% of sepsis cases were flagged by the alerts and all alerts on sepsis patients occurred after the administration of antibiotics, at which point the alerts had limited clinical utility[19]. A recent evaluation of the Epic Sepsis Model, one of the most widely deployed sepsis early warning systems, found that only 12% of alerts occurred on sepsis patients and only 33% of sepsis cases were flagged by the system[32]. In contrast to these three studies, our companion paper found that 82% of sepsis cases were flagged by the TREWS alerts and 38% of alerts with an evaluation were confirmed by a provider[22].

Beyond sepsis, a recent large-scale study of the Advanced Alert Monitor—an alert system for general deterioration—found that the alert system improved patient outcomes[18]. Although an important demonstration of the potential of such systems, this study relied on dedicated staff to respond to each alert. In contrast, the system studied in the present study is a decentralized bedside system that does not trigger pop-ups or phone calls that interrupt the clinical workflow. There were no dedicated staff for reviewing and responding to alerts and providers chose if, when and how to respond to alerts. Having a decentralized bedside system may allow TREWS to better scale to hospital systems that lack the resources to allocate staff to review all alerts. On the other hand, systems without dedicated staff require buy-in and adoption from providers to be successful. In our companion paper, we used EHR data to examine various patient, provider and environmental factors associated with the adoption of TREWS[22]. In a further separate study, we analyzed qualitative impressions of TREWS and factors influencing its integration into workflow through semi-structured interviews with providers using the system[33].

Much of the work on early detection and treatment for sepsis has been motivated by a collection of retrospective studies showing that delays in antibiotics are associated with worse patient outcomes[5–8,25]. In the absence of a viable alternative, these studies generally measured the time to antibiotics using ED presentation as time zero. Although informative, these results do not suggest a way to recognize sepsis earlier and, as a result, much of the subsequent work has focused on reducing the time from recognition to treatment, for example, the extensive work on sepsis treatment bundle guidelines[9,10]. A recent position paper from the Infectious Disease Society of America highlights the need for such a time zero, noting that existing proposals for time zero are overly complex and subjective[24]. In the current study, the per-hour associations we found between time-to-first antibiotics and mortality closely matched two of the largest such retrospective studies[7,25]. Among sepsis patients who received antibiotics within 6 h and 12 h, these studies found an adjusted per-hour odds ratios for mortality of 1.09 (CI 1.05, 1.13) and 1.04 (CI 1.03, 1.06), respectively (we found odds ratios of 1.08 (CI 1.02, 1.15) and 1.05 (CI 1.02, 1.09)). The implication of this result is that the potential benefits of early recognition implied by these retrospective studies could be realized using a high-precision bedside early warning system such as TREWS. Furthermore, such high-precision alerts may be used as a more objective time zero when measuring compliance with sepsis treatment guidelines.

Previous studies have reported different magnitudes of mortality reduction associated with early antibiotic administration, with the strongest benefits observed in patients with septic shock[7,8]. However, thus far, these retrospective studies have not outlined methods to prospectively identify patients at highest risk for developing septic shock. One important finding of the present study is that patients who benefit more from timely treatment, namely, the high-risk cohort, can be identified near the time of the alert, which may be used to prioritize alerts. Adding an indicator of alert priority, alerting a rapid response team to high-risk cases or limiting alerts to such cases could reduce alert burden and help providers allocate time and resources to the patients most likely to benefit from timely intervention.

Our study has several limitations. First, the present study reflects one set of prespecified alert settings; the behavior of the alert, and thus associations between alert use and clinical outcomes, may vary under different alert settings. For example, in this deployment, the timing of the alerts was optimized to issue alerts only when patients had notable findings such as the presence of deterioration in key

Table 2 | Associations between alert confirmation and patient outcomes

	Treatment	Comparison	ARD or ARR	*P* value[a]
All included	*n* = 4,220	*n* = 2,657		
In-hospital mortality, no. (rate)	617 (14.6%)	509 (19.2%)	ARD −3.34% (−5.10, −1.67%)	<0.001
			ARR −18.18% (−26.31, −9.65%)	<0.001
SOFA progression at 72 h[b]	−0.8 ± 2.7	−0.4 ± 2.9	ARD −0.26 (−0.42, −0.11)	0.001
Median length of stay (h)[c]	156 (99–260)	190 (118–323)	ARD −11.58 (−18.13, −5.03)	0.001
High-risk cohort	*n* = 1,430	*n* = 935		
In-hospital mortality, no. (rate)	422 (29.5%)	320 (34.2%)	ARD −4.50% (−8.31, −0.78%)	0.012
			ARR −13.19% (−22.81, −2.45%)	0.012
SOFA progression at 72 h[b]	−1.5 ± 3.5	−1.2 ± 3.6	ARD −0.38 (−0.71, −0.05)	0.025
Median length of stay (h)[c]	210 (129–351)	246 (155–434)	ARD −14.21 (−32.47, −4.04)	0.127

Mortality is reported as 'count (percentage)', SOFA progression is reported as 'mean ± s.d.' and length of stay is reported as 'median (IQR)'. Associations are reported as either an ARD or ARR and are presented as 'ARD/ARR (95% CI)'. [a]*P* values for in-hospital mortality were based on nonparametric bootstrap resampling using 5,000 bootstrap samples. *P* values for SOFA progression and median length of stay were based on Student's *t*-tests. All tests were two sided and not adjusted for multiple comparisons. [b]SOFA progression at 72 h excludes patients who were discharged to hospice, left against medical advice or transferred to another acute care facility within 72 h of the alert. [c]Median length of stay was calculated only on patients who did not die in hospital.

markers. In future deployments, this restriction may be relaxed to provide earlier warning and more lead time.

Second, although we adjusted for a substantial list of potential confounding variables, we did not conduct a randomized trial, and therefore, we observed that associations may be subject to residual confounding; that is, it is possible that unobserved variables remained not included in our analysis, which cause both the provider response to TREWS and patient outcomes. Aspects of patient presentation that are not captured through clinical data in the EHR or billing codes are not represented in our patient variables. A recent paper raised the concern that observational studies on sepsis bundles do not account sufficiently for providers appropriately delaying care on medically complex patients[24]. We have attempted to address this concern by adjusting for the coding of comorbidities that may mimic sepsis, but it is possible that suspected comorbidities that are not documented as codes were not included. In addition, besides the hospital at which the encounter occurred, we have not adjusted for provider-specific variables. Thus, if alert confirmation was associated with patient care, for reasons that are not explained by patient variables, this may lead to additional confounding. Unfortunately, large-scale randomized trials of early warning systems are operationally difficult to carry out, requiring tight control over clinical response if randomizing at the patient level or deployments spanning many sites if randomizing at the unit or hospital level. The most common alternative, pre-/post-deployment studies, suffer from their own set of potential biases such as surveillance bias and changes in treatment standards over time. Although we used a prospective cohort design in the present study, it is likely that all three types of studies—randomized, pre/post and cohort—will be needed to improve our understanding of these systems.

Third, there is continued debate about how to reliably identify sepsis retrospectively. We identified sepsis cases using an EHR-based sepsis phenotype that accounts for comorbidities that mimic sepsis and has shown increased sensitivity and precision compared with the alternatives[23,34], but we cannot exclude the possibility that some identified patients had noninfectious syndromes with symptoms that mimic the presentation of sepsis. Fourth, we did not have data on whether the antibiotics given to a patient were appropriate for their infection. Assessing the appropriateness of antimicrobial therapy, potentially via positive blood culture results when available, may give more accurate assessments of when effective treatment began and sepsis-related outcomes[5,35]. Fifth, our study was limited to a single hospital system and geographical region. Although studies including more diverse populations from other geographical regions are needed, the concordance of our results with several diverse studies on antibiotics timing suggests that our results may be applicable to a broader target population. Finally, the present study considers the associations between alert use and patient outcomes on a target population of sepsis patients who received an alert but had not received antibiotics at the time of the alert. It is also important to consider the impact of TREWS on other populations. For example, it is important to ensure that alerts on nonsepsis patients do not lead to over-prescription of antibiotics and that resources are not being drawn away from sepsis patients who do not receive an alert. These populations will be the subject of future studies.

Although large-scale randomized trials are needed, our findings indicate the potential for high-precision alert systems to identify sepsis patients early and improve patient outcomes. Furthermore, our findings among high-risk patients indicate that future alert systems may reduce overall alert burden by prioritizing alerts on certain patients. Work is under way to further improve on TREWS and to generalize TREWS to other acute conditions. Furthermore, work is needed to understand how the use and benefits of alert systems evolve over time, best practices for presenting information from ML-based alerts in critical care settings, and the tradeoffs between workflows with and without dedicated staff in terms of provider burden, benefits to patients and scalability.

Online content

Any methods, additional references, Nature Research reporting summaries, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-022-01894-0.

Received: 16 September 2021; Accepted: 8 June 2022;
Published online: 21 July 2022

References

1. Rhee, C. et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. *JAMA Netw. Open* **2**, e187571–e187571 (2019).
2. Riedemann, N. C., Guo, R. F. & Ward, P. A. The enigma of sepsis. *J. Clin. Invest.* **112**, 460–467 (2003).
3. Marshall, J. C. Why have clinical trials in sepsis failed? *Trends Mol. Med.* **20**, 195–203 (2014).
4. Rhodes, A. et al. *Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. Crit. Care Med.* **43**, 304–377 (2017).

5. Kumar, A. et al. 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).
6. Ferrer, R. et al. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Crit. Care Med.* **42**, 1749–1755 (2014).
7. Liu, V. X. et al. The timing of early antibiotics and hospital mortality in sepsis. *Am. J. Respir. Crit. Care Med.* **196**, 856–863 (2017).
8. Peltan, I. D. et al. ED door-to-antibiotic time and long-term mortality in sepsis. *Chest* **155**, 938–946 (2019).
9. Chamberlain, D. J., Willis, E. M. & Bersten, A. B. The severe sepsis bundles as processes of care: a meta-analysis. *Aust. Crit. Care* **24**, 229–243 (2011).
10. Damiani, E. et al. Effect of performance improvement programs on compliance with sepsis bundles and mortality: a systematic review and meta-analysis of observational studies. *PLoS ONE* **10**, e0125827 (2015).
11. Giannini, H. M. et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit. Care Med.* **47**, 1485–1492 (2019).
12. Desautels, T. et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med. Inform.* **4**, 1–15 (2016).
13. Shashikumar, S. P., Josef, C. S., Sharma, A. & Nemati, S. DeepAISE—an interpretable and recurrent neural survival model for early prediction of sepsis. *Artif. Intell. Med.* **113**, 102036 (2021).
14. Horng, S. et al. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS ONE* **12**, e0174708 (2017).
15. Bedoya, A. D. et al. Machine learning for early detection of sepsis: an internal and temporal validation study. *JAMIA Open* **3**, 252–260 (2020).
16. Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J. & Das, R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir. Res.* **4**, e000234 (2017).
17. McCoy, A. & Das, R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual.* **6**, e000158 (2017).
18. Escobar, G. J. et al. Automated identification of adults at risk for in-hospital clinical deterioration. *N. Engl. J. Med.* **383**, 1951–1960 (2020).
19. Topiwala, R., Patel, K., Twigg, J., Rhule, J. & Meisenberg, B. Retrospective observational study of the clinical performance characteristics of a machine learning approach to early sepsis identification. *Crit. Care Explor.* **1**, e0046 (2019).
20. Ginestra, J. C. et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. *Crit. Care Med.* **47**, 1477 (2019).
21. Henry, K. E., Hager, D. N., Pronovost, P. J. & Saria, S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* **7**, 299ra122–299ra122 (2015).
22. Henry, K. E. et al. Factors driving provider adoption of the TREWS machine-learning-based early warning system and its effects on sepsis treatment timing. *Nat. Med.* https://doi.org/10.1038/s41591-022-01895-z (2022).
23. Henry, K. E., Hager, D. N., Osborn, T. M., Wu, A. W. & Saria, S. Comparison of automated sepsis identification methods and electronic health record-based sepsis phenotyping: improving case identification accuracy by accounting for confounding comorbid conditions. *Crit. Care Explor.* **1**, e0053 (2019).
24. Rhee, C. et al. Infectious diseases society of america position paper: recommended revisions to the national severe sepsis and septic shock early management bundle (SEP-1) sepsis quality measure. *Clin. Infect. Dis.* **72**, 541–552 (2021).
25. Seymour, C. W. et al. Time to treatment and mortality during mandated emergency care for sepsis. *N. Engl. J. Med.* **376**, 2235–2244 (2017).
26. Vanderweele, T. J., Luedtke, A. R., Van Der Laan, M. J. & Kessler, R. C. Selecting optimal subgroups for treatment using many covariates. *Epidemiology* **30**, 334–341 (2019).
27. Manaktala, S. & Claypool, S. R. Evaluating the impact of a computerized surveillance algorithm and decision support system on sepsis mortality. *J. Am. Med. inform. Assoc.* **24**, 88–95 (2017).
28. Burdick, H. et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. *BMJ Health Care Inform.* **27**, e100109 (2020).
29. Guy, J. S., Jackson, E. & Perlin, J. B. Accelerating the clinical workflow using the sepsis prediction and optimization of therapy (SPOT) tool for real-time clinical monitoring. *NEJM Catal. Innov. Care Deliv*. https://doi.org/10.1056/CAT.19.1036 (2020).
30. Rosenbaum, P. R. & Briskman. *Design of Observational Studies* Vol. 10 (Springer, 2010).
31. Hernán, M. A. & Robins, J. M. Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183**, 758–764 (2016).
32. Wong, A. et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* **48109**, 1–6 (2021).
33. Henry, K. E. et al. Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. NPJ Digit. Med. https://doi.org/10.1038/s41746-022-00597-7 (2022).
34. Saria, S. & Henry, K. E. Too many definitions of sepsis: Can machine learning leverage the electronic health record to increase accuracy and bring consensus? *Crit. Care Med.* **48**, 137–141 (2020). https://doi.org/10.1097/CCM.0000000000004144
35. Rhee, C. et al. Prevalence of antibiotic-resistant pathogens in culture-proven sepsis and outcomes associated with inadequate and broad-spectrum empiric antibiotic use. *JAMA Netw. Open* **3**, e202899 (2020).

Methods

The present study was approved by the Johns Hopkins University internal review board (IRB no. 00252594) and a waiver of consent was obtained.

TREWS. TREWS is an ML-based early warning system that continuously monitors patients for risk of sepsis using routinely collected EHR data. Its underlying model uses vital signs, laboratory data, clinician notes, medication history (excluding antibiotic orders), procedure history and clinical history to generate a sepsis risk score in real time as new information becomes available in the EHR. The score is based on an approach that allows the model[36] to account for different sepsis presentations and to incorporate patient context such as confounding comorbidities, procedures and level of care. For a full description of the TREWS model and a retrospective performance evaluation, as well as the interface, deployment and integration into the clinical workflow, see our companion paper[22]. Importantly, when an alert occurs, providers can enter a dedicated TREWS page in the EHR that displays various pieces of information about the alert and the patient and allows the provider to enter a patient evaluation. An evaluation consists of entering a suspected source of infection (confirmation) or clicking a button to indicate that no new or worsening infection is present and then reviewing the list of organ dysfunction indicators and removing any that the provider believes should not be attributed to sepsis (dismissal). If a provider left the TREWS page without entering an evaluation, the interaction is not counted as an evaluation in the context of the present study. We used these evaluations as our primary measure of provider interaction with TREWS and to define study and comparison arms.

Population. We conducted a prospective, multi-site, two-arm cohort study using EHR data from two academic and three community hospitals at which the TREWS alert system was deployed. Patients were eligible for inclusion if they were aged ≥18 years and presented to the ED or were admitted to an inpatient unit at one of the five deployment hospitals between deployment of TREWS and 30 September 2020. The specific hospitals and deployment dates were Howard County General Hospital (HCGH, 1 April 2018), Suburban Hospital (10 October 2018), Bayview Medical Center (BMC, 27 February 2019), Johns Hopkins Hospital (JHH, 16 April 2019) and Sibley Memorial Hospital (15 May 2019). We treated each time a patient presented to the ED or was admitted as a unique patient encounter and included each encounter separately.

For our primary analysis of patient outcomes, we included all patients who triggered an alert at most 1 h before admission or ED triage and met the criteria for sepsis based on a strategy called EHR-based sepsis phenotyping[23,34]. Briefly, this approach is a refinement of the criteria used in the third sepsis consensus definition (Sepsis-3)[37] and the CDC Adult Sepsis Event Toolkit sepsis[38], in which patients with features consistent with sepsis, but explained by other conditions (for example, hemorrhage) are excluded. As the full description of this sepsis definition is involved, we refer interested readers to the original work in which it appeared[23].

To ensure that the alerts considered were actionable and to exclude outliers, we excluded patients who did not receive antibiotics within 24 h of the alert. See Extended Data Table 5 for a list of the antibiotics considered. Patients who had antibiotics administered before the alert were excluded, because we are interested primarily in patients for whom the alert may impact antibiotic administration timing and antibiotics given before the alert could not have been influenced by the alert. Finally, patients admitted directly to the ICU who did not pass through the ED were excluded because such patients were likely to have received previous treatment not reflected in the health record. A waterfall diagram for this cohort is shown in Fig. 1.

Study outcomes. The primary outcome was all-cause in-hospital mortality which was measured as the patient's status at discharge. Secondary outcomes were the difference between a patient's SOFA score at the time of the alert and 72 h after the alert and hospital length of stay after the alert until discharge among survivors[39]. SOFA scores were calculated using the worst measurements taken between 48 h and 72 h after the alert. For patients who died or were discharged before 72 h after the alert, the SOFA score was based on the worst measurements from the 24 h preceding death or discharge.

Study design. To evaluate the association between provider response to TREWS and patient outcomes, we compared outcomes between two arms: patients who had their alert evaluated and confirmed by a provider within 3 h of the alert (study arm) and patients who did not (comparison arm). The comparison arm is made up of patients who never had an evaluation entered into the TREWS page, as well as patients whose alert was dismissed. To account for potential differences between the arms, we adjusted for a variety of patient variables described in detail below[30,31]. We made this comparison among all included patients and in the cohort of high-risk patients (described in detail below). Patients who died during their stay were excluded from the length-of-stay analysis. Patients who left against medical advice, were discharged to a hospice or were transferred to another acute care facility were excluded from the SOFA progression analysis (all these patients were included in the mortality analysis).

Adjustment variables. The relationship between antibiotics timing and patient outcomes may be confounded by several aspects of patient presentation. We adjusted for a range of patient variables to account for potential confounding. Similar to previous studies[7,8,25], we adjusted for patient age, documented sex, comorbidity history and measurements taken during the current encounter. To account for differences due to pre-existing conditions and patient history, we adjusted for the patient's Charlson comorbidity index (CCI) as well as individual indicators for a history of diabetes (with and without complications), dementia, malignant tumors or metastatic solid tumors. These conditions were identified from *International Classification of Diseases*, 10th revision (ICD-10) codes included in the patient's medical history[49]. To account for differences in treatment arising from patient presentation at the current visit, we adjusted for several sepsis-relevant labs, vital signs and treatments, including systolic blood pressure, Glasgow Coma Score (GCS) <15 (indicating any abnormal mental activity)[8], temperature, white blood cell count, lactate >2 mmol l^{-1} and indicators for vasopressors and mechanical ventilation. We additionally adjusted for composite measures of acute patient severity, including individual SOFA score components and the APACHE II score[39,40]. For labs and vital signs, we used the most recent measurement taken in the 24 h before the alert and imputed a normal value if no measurement was observed during this period. SOFA and APACHE II scores were calculated using the worst measurements from the 24 h before the alert. In cases where the alert occurred within 12 h of presentation to the ED or admission to an inpatient unit (whichever came first), we used the worst measurements recorded within 12 h of presentation/admission to allow for lab processing and recording delays.

In addition to the variables used in previous studies, we included several additional patient and hospital-related variables. In medically complex patients, clinicians may wish to obtain more information on the patients before ordering antibiotics. To account for this, we further adjusted for the presence of several comorbidities that may obscure a sepsis diagnosis, including metastatic cancer, end-stage renal disease, congestive heart failure, acute liver disease, gastrointestinal bleeding and chronic obstructive pulmonary disease. These comorbidities were identified from ICD-10 codes listed as part of the patient's problem list and marked as present on arrival. In addition, for ED patients, we adjusted for whether or not the trauma team was activated on arrival. To account for differences in clinical practice between hospitals, we included an indicator for the admitting hospital. Finally, we included a binary indicator for presentation at the hospital after 1 April 2020, to account for differences in treatment patterns arising from the COVID-19 pandemic.

High-risk cohort definition. In addition to the TREWS sepsis detection model, a separate model was developed for forecasting how much a patient's mortality risk would increase without timely antibiotics (that is, sensitivity to antibiotic timing). Identifying patients with large increases in mortality risk may allow providers to prioritize alerts on patients who will benefit the most from timely care and allows us to examine heterogeneity in the associations between alert use and patient outcomes, that is, to perform stratification. This approach was inspired by previous work by VanderWeele et al. on identifying subgroups of patients who respond most to treatment[26]. VanderWeele et al. performed effect stratification by first developing a model to predict, based on baseline variables, whether a patient would respond to treatment and then estimated treatment effects among patients predicted to have a strong response. We applied a similar approach to our study by first developing a model to predict how much a patient's mortality risk would increase without timely antibiotics (analogous to predicting treatment response) and then using this prediction to stratify the associations between interaction with TREWS and patient outcomes.

To predict sensitivity to antibiotic timing, we forecast each patient's probability of in-hospital mortality with and without administration of antibiotics within 3 h of their alert and then took the ratio between these two probabilities. A higher ratio indicates a higher predicted sensitivity to antibiotic timing. The probability of in-hospital mortality with and without 3-h antibiotics was forecast using two separate ridge logistic regression models trained on retrospective EHR data from before deployment of TREWS. These data included 4,860 adult patients from our development dataset who were admitted to three of the five deployment hospitals (HCGH, JHH and BMC) between 1 January 2016 and 31 March 2018, who would have triggered an alert during their stay had the alert been active and who met all the inclusion criteria for our primary analysis (described above). Both models included as predictors all baseline variables described in the previous section except for the admitting hospital. All-cause in-hospital mortality was the binary prediction target. To account for nonlinearities, continuous lab values and vital signs were included as piecewise linear terms according to the thresholds used in APACHE II (ref. [40]). The time from alert to first antibiotic administration was calculated on this retrospective data by running the TREWS model retrospectively to calculate the approximate time at which an alert would have occurred had TREWS been active for these patients. The two logistic regression models were trained separately on patients who did and did not receive an antibiotic within 3 h of when their alert would have occurred. To define a discrete high-risk cohort, a threshold for the ratio of mortality probabilities was chosen such that approximately one-third of patients in the retrospective data were identified as high risk.

The regularization strength for both models was tuned using grid search and 20 random 50:50 splits of the development data. For each random split, the models were trained and a high-risk threshold was chosen on the training set, the high-risk

patients were identified on the test set and the adjusted association between timely antibiotics and in-hospital mortality was estimated among high-risk patients in the test set. Then, these estimates were averaged across the 20 random test sets, the regularization strength was chosen that maximized this average and the models were retrained using all of the retrospective data and the selected regularization strength.

To determine whether a patient in our study population fell into the high-risk cohort, both logistic regression models were applied to the patient's baseline variables, and the patient was identified as high risk if the ratio of the two predicted probabilities was above the high-risk threshold. Note that these models were used to identify high-risk patients only for analysis and were not presented to the frontline users as part of the TREWS system.

Statistical analyses. For in-hospital mortality, we used logistic regression to estimate the ARD and ARR between study and comparison arms[41]. We used linear regression to estimate the adjusted difference in mean SOFA progression and quantile regression to estimate the adjusted difference in median length of stay[42]. To account for nonlinearities, continuous lab values and vital signs were included as piecewise linear terms according to the thresholds used in APACHE II (ref. [40]). For all analyses, heteroskedasticity-robust estimators were used to estimate standard errors and construct CIs. All reported CIs are 95% CIs, and all *P* values correspond to two-sided tests. All statistical models were implemented via the Python packages 'scikit-learn' v.0.24.2 and 'statsmodels' v.0.12.1 (refs. [43,44]).

Secondary and sensitivity analyses. As a secondary analysis, we estimated the association between time from alert to first antibiotic administration and patient outcomes. Despite previous work demonstrating an association between timely antibiotics and mortality when time-to-first antibiotics is measured from admission, onset or recognition[5–10], it is important to establish that there is still an association when the TREWS alert is used as time zero and validate the potential to use a TREWS alert to initiate treatments in future studies. For example, suppose that providers can recognize most sepsis patients who need immediate antibiotics before the alert occurs. In this case, we might expect to see little association between antibiotic timing and patient outcomes in the remaining sepsis patients who did not receive antibiotics before their alert (that is, our study population). Using the same methods described above, we compared adjusted outcomes between patients in the target population who received their first antibiotic within 3 h of their alert with those who did not. In addition, for the purposes of direct comparison with previous work on antibiotic timing, we emulated the analyses from two previous retrospective studies on antibiotic timing[7,25]. These studies were chosen for their size and replicability. We estimated the per-hour adjusted odds ratio between time from alert to first antibiotic administration and in-hospital mortality using logistic regression and including all variables described above. We repeated this estimate including patients who receive their first antibiotics within 6 and 12 h, corresponding to the two previous studies.

Last of all, we tested the robustness of our analyses by repeating the primary analysis (adjusted comparison of patient outcomes between arms) under several different modifications. First, to ensure that our conclusions were not dependent on our choice of statistical method, we repeated the analysis using an alternative method, inverse probability of treatment weighting[45–47]. We used logistic regression for the probability of treatment model and stabilized and truncated weights at the 1st and 99th weight percentiles to reduce estimation variance[48]. Second, because a patient's problem list reflects real-time diagnoses made by providers, it is less reliable than final diagnoses that reflect all information from a patient's stay. Similarly, clinicians may not always mark a present-on-arrival diagnosis as such. We tested the sensitivity to these issues by replacing ICD-10 codes in the patient's problem list with final diagnosis ICD-10 codes. Final diagnosis codes are more comprehensive but may include information from well after the patient's alert, leading to potential over-adjustment. Third, because we included each patient encounter separately in our data, we wished to ensure that our conclusions were not overly sensitive to a few patients with multiple encounters. We tested this sensitivity by repeating the analysis using only the first encounter for each patient. Fourth, the COVID-19 pandemic impacted both characteristics of patients in our data and the care those patients received. We tested the sensitivity to inclusion of data during the COVID-19 pandemic by repeating the analysis using only patients admitted before 1 April 2020. Finally, the list of antibiotics used to determine the time of first antibiotic administration was intentionally inclusive. To ensure that our results are not sensitive to the specific antibiotics included, we repeated the analysis using a restricted list of antibiotics which removes some antibiotics that are unlikely to be used either to treat sepsis or on adult patients. The restricted list removes amoxicillin, azithromycin, cefotaxime, dapsone, erythromycin, neomycin and rifampin from the antibiotics listed in Extended Data Table 5.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data are not publicly available because they are from EHRs approved for limited use by Johns Hopkins University investigators. Making the data publicly available without additional consent, ethical or legal approval might compromise patients' privacy and the original ethical approval. To perform additional analyses using this data, researchers should contact A.W.W. or S.S. to apply for an IRB-approved research collaboration and obtain an appropriate data use agreement.

Code availability

The TREWS early warning system described in the present study is available from Bayesian Health, New York. The underlying source code is proprietary intellectual property and is not available. Code for the primary statistical analyses and development of the high-risk cohort can be found at https://github.com/royadams/adams_et_al_2022_code.

References

36. Jordan, M. I. & Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. *Proceedings of International Conference on Neural Networks* **2**, 1339–1344 (1993).
37. Seymour, C. W. et al. Assessment of clinical criteria for sepsis for the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* **315**, 762–774 (2016).
38. Rhee, C., Dantes, R. B., Epstein, L. & Klompas, M. Using objective clinical data to track progress on preventing and treating sepsis: CDC's new adult sepsis event surveillance strategy. *BMJ Qual. Saf.* **28**, 305–309 (2019).
39. Vincent, J. L. et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intens. Care Med.* **22**, 707–710 (1996).
40. Knaus, W. A., Draper, E. A., Wagner, D. P. & Zimmerman, J. E. APACHE II: a severity of disease classification system. *Crit. Care Med.* **13**, 818–829 (1985).
41. Norton, E. C., Miller, M. M. & Kleinman, L. C. Computing adjusted risk ratios and risk differences in Stata. *Stata J.* **13**, 492–509 (2013).
42. Peng, L. Quantile regression for survival data. *Annu. Rev. Stat. Its Appl.* **8**, 413–437 (2021).
43. Seabold, S. & Perktold, J. statsmodels: econometric and statistical modeling with python. In van der Walt, S. & Millman, J. (Eds.) *Proc. 9th Python in Science Conference* 92–96 (2010).
44. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Horvitz, D. G. & Thompson, D. J. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **47**, 663–685 (1952).
46. Robins, J. M. Marginal structural models versus structural nested models as tools for causal inference. In Halloran, M. E. & Berry, D. (Eds.) *Statistical Models in Epidemiology, the Environment, and Clinical Trials* 95–133 (Springer, 2000).
47. Hernán, M. A. & Robins, J. M. *Causal Inference: What If* (Chapman & Hall/CRC, 2020).
48. Lee, B. K., Lessler, J. & Stuart, E. A. Weight trimming and propensity score weighting. *PLoS ONE* **6**, e18174 (2011).
49. World Health Organization. *ICD-10 : international statistical classification of diseases and related health problems : tenth revision* (World Health Organization, 2004).

Acknowledgements

We thank Y. Ahmad, M. Yeo and Y. Karklin whose work significantly contributed to early iterations of the development of the deployed system. Further, we thank R. Demski, K. D'Souza, A. Kachalia, A. Chen and clinical and quality stakeholders who contributed to tool deployment, education and championing the work. We gratefully acknowledge the following sources of funding: the Gordon and Betty Moore Foundation (award no. 3926), the National Science Foundation (NSF) Future of Work at the Human-technology Frontier (award no. 1840088) and the Alfred P. Sloan Foundation research fellowship (2018). This information or content and conclusions are those of the authors and should not be construed as the official position or policy of, nor should any endorsements be inferred by, the NSF of the US Government.

Author contributions

K.E.H., R.A., A.W.W. and S.S. contributed to the initial study design and preliminary analysis plan. S.S. led the development and deployment efforts for the TREWS software. K.E.H., H.S., N.R., A.Z., A.S., R.C.L., L.J., M.H., S.M., D.N.H., A.R.A., A.W.W. and S.S. contributed to the system deployment. K.E.H., R.A., E.Y.K., S.E.C., E.S.C., D.N.H., A.W.W. and S.S. contributed to the review and analysis of the results. All authors contributed to the final preparation of the manuscript.

Competing interests

Under a license agreement between Bayesian Health and the Johns Hopkins University, K.E.H., S.S. and Johns Hopkins University are entitled to revenue distributions. In addition, the university owns equity in Bayesian Health. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its

conflict-of-interest policies. S.S. also has grants from the Gordon and Betty Moore Foundation, the NSF, the National Institutes of Health, Defense Advanced Research Projects Agency, the Food and Drug Administration and the American Heart Association; she is a founder of and holds equity in Bayesian Health; she is the scientific advisory board member for PatientPing; and she has received honoraria for talks from a number of biotechnology, research and health-tech companies. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflict-of-interest policies. D.N.H. discloses salary support and funding to his institution from the Marcus Foundation for the conduct of the vitamin C, thiamine and steroids in sepsis trial. S.E.C. received consulting fees from Basilea for work on an infection adjudication committee for a *Staphylococcus aureus* bacteremia trial. The remaining authors declare no disclosures of conflicts of interest.

Additional information

Extended data are available for this paper at https://doi.org/10.1038/s41591-022-01894-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-01894-0.

Correspondence and requests for materials should be addressed to Albert W. Wu or Suchi Saria.

Peer review information *Nature Medicine* thanks Derek Angus, Melanie Wright and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling editor: Michael Basson in collaboration with the *Nature Medicine* team.

Extended Data Table 1 | Sample statistics by hospital

	Primary analysis cohort					
	All	HCGH	SH	BMC	JHH	SMC
No. encounters	6,877	2,174	1,025	1,478	1,565	635
No. patients	6,264	1,970	953	1,361	1,441	597
Age (yr)	66.0 ± 17.2	68.3 ± 16.8	74.3 ± 15.7	63.7 ± 16.0`	57.8 ± 16.3	70.7 ± 16.6
Documented male	3,231 (47.0%)	1,020 (46.9%)	494 (48.2%)	697 (47.2%)	723 (46.2%)	297 (46.8%)
Admit from ED	6,582 (95.7%)	2,173 (100.0%)	1,012 (98.7%)	1,456 (98.5%)	1,317 (84.2%)	624 (98.3%)
ED discharge	86 (1.3%)	38 (1.7%)	3 (0.3%)	30 (2.0%)	8 (0.5%)	7 (1.1%)
Ever admitted to ICU	3,084 (44.8%)	794 (36.5%)	449 (43.8%)	875 (59.2%)	741 (47.3%)	225 (35.4%)
Mean CCI	6.5 ± 4.0	6.4 ± 3.9	7.5 ± 3.8	6.2 ± 3.9	6.2 ± 4.2	7.0 ± 3.8
At admission‡						
APACHE II	13.5 ± 8.4	12.0 ± 7.7	14.8 ± 8.5	16.4 ± 9.3	13.3 ± 8.4	10.5 ± 5.9
SOFA	4.7 ± 3.8	4.2 ± 3.6	5.1 ± 3.9	5.7 ± 4.2	4.7 ± 3.9	3.8 ± 2.9
GCS < 15	2,732 (39.7%)	691 (31.8%)	470 (45.9%)	868 (58.7%)	569 (36.4%)	134 (21.1%)
Lactate > 2 mmol/L	4,959 (72.1%)	1,773 (81.6%)	440 (42.9%)	1,214 (82.1%)	1,107 (70.7%)	425 (66.9%)
Septic shock	2,897 (42.1%)	847 (39.0%)	376 (36.7%)	804 (54.4%)	681 (43.5%)	189 (29.8%)
At alert						
APACHE II	11.8 ± 7.7	10.4 ± 7.0	12.9 ± 7.8	14.7 ± 8.7	11.4 ± 7.6	9.1 ± 5.1
SOFA	4.1 ± 3.4	3.7 ± 3.2	4.4 ± 3.4	5.0 ± 3.8	4.0 ± 3.4	3.2 ± 2.5
GCS < 15	2,288 (33.3%)	563 (25.9%)	388 (37.9%)	783 (53.0%)	460 (29.4%)	94 (14.8%)
Lactate > 2 mmol/L	4,790 (69.7%)	1,757 (80.8%)	336 (32.8%)	1,197 (81.0%)	1,086 (69.4%)	414 (65.2%)
Death in hospital	1,126 (16.4%)	273 (12.6%)	256 (25.0%)	254 (17.2%)	237 (15.1%)	106 (16.7%)

All discrete values are reported as 'count (percent)' and all continuous values are reported either as 'mean ± std' or 'median (IQR)'.
‡ 'At admission' values were calculated based on measurements taken within 24 hours ED triage or admission.

Extended Data Table 2 | Extended sample statistics

	Primary analysis cohort			High-risk cohort		
	Alert conf. within 3 hrs	Alert not conf. within 3 hrs	p-value†	Alert conf. within 3 hrs	Alert not conf. within 3 hrs	p-value†
No. encounters	4,220	2,657	---	1,430	935	---
No. patients	3,925	2,577	---	1,379	923	---
Age (yr)	66.6 ± 17.5	65.1 ± 16.8	< 0.001	67.9 ± 16.0	65.1 ± 15.6	< 0.001
Documented male	1,992 (47.2%)	1,239 (46.6%)	0.661	701 (49.0%)	437 (46.7%)	0.296
Admit from ED	4,148 (98.3%)	2,434 (91.6%)	< 0.001	1,414 (98.9%)	866 (92.6%)	< 0.001
ED discharge	62 (1.5%)	24 (0.9%)	0.052	14 (1.0%)	9 (1.0%)	0.862
Ever admitted to ICU	1,811 (42.9%)	1,273 (47.9%)	< 0.001	1,147 (80.2%)	731 (78.2%)	0.254
Mean CCI	6.5 ± 3.9	6.5 ± 4.0	0.372	7.5 ± 4.0	6.9 ± 4.1	< 0.001
Hospital						
HCGH	1,493 (35.4%)	681 (25.6%)	< 0.001	410 (28.7%)	209 (22.4%)	0.001
SH	677 (16.0%)	348 (13.1%)	0.001	265 (18.5%)	149 (15.9%)	0.117
BMC	941 (22.3%)	537 (20.2%)	0.043	426 (29.8%)	250 (26.7%)	0.119
JHH	831 (19.7%)	734 (27.6%)	< 0.001	255 (17.8%)	242 (25.9%)	< 0.001
SMH	278 (6.6%)	357 (13.4%)	< 0.001	74 (5.2%)	85 (9.1%)	< 0.001
Patient history						
Dementia	688 (16.3%)	329 (12.4%)	< 0.001	244 (17.1%)	115 (12.3%)	0.002
Diabetes w/o complications	1,663 (39.4%)	1,013 (38.1%)	0.300	606 (42.4%)	359 (38.4%)	0.060
Diabetes w/ complications	960 (22.7%)	603 (22.7%)	0.982	415 (29.0%)	232 (24.8%)	0.028
Malignancy	1,104 (26.2%)	765 (28.8%)	0.018	431 (30.1%)	265 (28.3%)	0.373
Metastatic solid tumor	623 (14.8%)	419 (15.8%)	0.272	258 (18.0%)	155 (16.6%)	0.389
Comorbidities present on arrival						
End stage renal disease	12 (0.3%)	7 (0.3%)	0.940	7 (0.5%)	3 (0.3%)	0.769
Chronic obstructive pulmonary disease	43 (1.0%)	18 (0.7%)	0.181	13 (0.9%)	8 (0.9%)	0.929
Congestive heart failure	50 (1.2%)	31 (1.2%)	0.963	19 (1.3%)	14 (1.5%)	0.871
Acute liver disease	2 (0.0%)	3 (0.1%)	0.602	1 (0.1%)	3 (0.3%)	0.347
Gastrointestinal bleed	0 (0.0%)	0 (0.0%)	1.000	0 (0.0%)	0 (0.0%)	1.000
Metastatic solid tumor	10 (0.2%)	3 (0.1%)	0.385	4 (0.3%)	1 (0.1%)	0.662
Septic shock §	2,167 (43.5%)	730 (38.6%)	< 0.001	1,344 (76.7%)	436 (71.1%)	0.007
At alert						
APACHE II	11.7 ± 7.5	11.9 ± 8.1	0.555	19.1 ± 7.4	19.7 ± 7.9	0.13
SOFA	4.0 ± 3.4	4.3 ± 3.4	< 0.001	7.7 ± 3.0	7.8 ± 3.0	0.159
GCS < 15	1,381 (32.7%)	907 (34.1%)	0.237	949 (66.4%)	645 (69.0%)	0.199
Lactate > 2 mmol/L	3,160 (74.9%)	1,630 (61.3%)	< 0.001	1,099 (76.9%)	595 (63.6%)	< 0.001
Systolic blood pressure	117.0 ± 28.3	119.5 ± 28.4	< 0.001	107.1 ± 30.3	113.7 ± 32.7	< 0.001
Heart rate	102.9 ± 22.5	98.9 ± 22.7	< 0.001	104.3 ± 25.5	99.9 ± 26.5	< 0.001
Respiration rate	21.4 ± 6.8	20.9 ± 6.6	< 0.001	22.8 ± 8.0	22.1 ± 7.8	0.057
White blood cell count	15.0 ± 8.6	13.2 ± 8.1	< 0.001	14.8 ± 9.3	13.3 ± 8.5	< 0.001
Temperature	99.0 ± 2.4	98.4 ± 2.0	< 0.001	98.4 ± 3.1	97.8 ± 2.6	< 0.001
On vasopressors	901 (21.4%)	490 (18.4%)	0.004	851 (59.5%)	473 (50.6%)	< 0.001
On mechanical ventilation	511 (12.1%)	454 (17.1%)	< 0.001	473 (33.1%)	402 (43.0%)	< 0.001

All discrete values are reported as 'count (percent)' and all continuous values are reported either as 'mean ± std' or 'median (IQR)'.
† P-values were based on Pearson's chi-square and two-sided Wilcoxon rank-sum tests for categorical and continuous variables, respectively. P-values were not adjusted for multiple comparisons.
§ Septic shock was defined as being both positive for sepsis and either receiving vasopressors or having a lactate > 4 mmol/L within 48 hours of the first signs of organ dysfunction.

Extended Data Table 3 | Sensitivity analyses

Outcome	Original analysis (N=6,877)	Using IPTW (N=6,877)	Using final diagnosis codes (N=6,877)	First admission only (N=4,989)	Pre-COVID only (N=4,945)	Restricted antibiotics (N=6,855)
In-hospital mortality						
ARD	-3.34% [-5.10%, -1.67%]***	-2.86% [-4.68%, -1.09%]***	-2.77% [-4.42%, -1.09%]***	-3.73% [-5.76%, -1.73%]**	-3.51% [-5.50%, -1.50%]***	-3.48% [-5.21%, -1.70%]***
ARR	-18.18% [-26.31%, -9.65%]***	-15.78% [-24.21%, -6.43%]***	-15.36% [-23.39%, -6.44%]***	-20.49% [-29.49%, -10.27%]**	-19.89% [-29.12%, -9.16%]***	-18.75% [-26.63%, -9.74%]***
SOFA at 72 hours						
ARD	-0.26 [-0.42, -0.11]**	-0.28 [-0.46, -0.12]***	-0.19 [-0.34, -0.05]*	-0.30 [-0.48, -0.12]**	-0.24 [-0.42, -0.06]**	-0.29 [-0.44, -0.13]***
Median length of stay (hrs)						
ARD	-11.58 [-18.13, -5.03]**	-10.0 [-22.3, -4.0]***	-13.25 [-19.86, -6.64]***	-12.16 [-19.43, -4.89]**	-13.52 [-20.80, -6.24]***	-12.20 [-18.76, -5.63]***

Associations between treatment arm and patient outcomes under different modifications to the analysis. Associations are reported as either an adjusted risk difference (ARD) or adjusted relative reduction (ARR) and are presented as 'association [95% confidence interval]'. P-values are denoted: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$. P-values for in-hospital mortality were based on non-parametric bootstrap resampling using 5,000 bootstrap samples. P-values for SOFA progression and median length of stay were based on t-tests. All tests were two-sided and were not adjusted for multiple comparisons.

Extended Data Table 4 | Associations between antibiotics timing and patient outcomes

	ABx <= 3 hrs	Abx > 3 hrs	Adjusted risk difference or adjusted relative reduction	p-value §
All included	N=4,987	N=1,890		
In-hospital mortality	776 (15.6%)	350 (18.5%)	ARD -3.52% [-5.43%, -1.62%]	< 0.001
			ARR -18.59% [-26.73%, -9.29%]	< 0.001
SOFA progression at 72 hours ‡	-0.8 ± 2.7	-0.4 ± 3.0	ARD -0.32 [-0.48, -0.15]	< 0.001
Median length of stay (hrs) †	161 (101 - 265)	192 (120 - 330)	ARD -15.6 [-22.6, -8.6]	< 0.001
High-risk cohort	N=1,752	N=613		
In-hospital mortality	524 (29.9%)	218 (35.6%)	ARD -6.40% [-10.47%, -2.25%]	0.001
			ARR -17.71%; [-26.74%, -6.83%]	0.001
SOFA progression at 72 hours ‡	-1.4 ± 3.4	-1.2 ± 3.8	ARD -0.33 [-0.70, 0.05]	0.086
Median length of stay (hrs) †	211 (130 - 360)	263 (164 - 456)	ARD -28.91 [-49.86, -7.97]	0.007

Mortality is reported as 'count (percent)', SOFA progression is reported as 'mean ± std', and length of stay is reported as 'median (IQR)'. Associations are reported as either an adjusted risk difference (ARD) or adjusted relative reduction (ARR) and presented as 'association [95% confidence interval].

‡ SOFA progression at 72 hrs excludes patients who were discharged to hospice, left against medical advice, or were transferred to another acute care facility within 72 hours of the alert.

† Median length of stay was calculated only on patients who did not die in the hospital.

§ P-values for in-hospital mortality were based on non-parametric bootstrap resampling using 5,000 bootstrap samples. P-values for SOFA progression and median length of stay were based on t-tests. All tests were two-sided and were not adjusted for multiple comparisons.

Extended Data Table 5 | List of included antibiotics

Amikacin	Ertapenem
Amoxicillin	Erythromycin
Ampicillin	Gentamicin
Ampicillin/sulbactam	Levofloxacin
Azithromycin	Linezolid
Aztreonam	Meropenem
Cefazolin	Metronidazole
Cefepime	Minocycline
Cefotaxime	Moxifloxacin
Cefotetan	Neomycin
Ceftaroline	Oxacillin
Ceftazidime	Penicillin
Ceftolozane/tazobactam	Penicillin G
Ceftriaxone	Piperacillin/tazobac
Cefuroxime	Polymyxin B
Ciprofloxacin	Quinupristin/dalfopristin
Clindamycin	Rifampin
Colistimethate	Tigecycline
Dapsone	Tobramycin
Daptomycin	Trimethoprim
Doxycycline	Vancomycin

nature portfolio

Corresponding author(s): Albert Wu and Suchi Saria

Last updated by author(s): Mar 24, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specialized software was used for data collection.
Data analysis	The TREWS early warning system described in this study is available from Bayesian Health, New York, New York. The underlying source code is proprietary intellectual property and is not available. Code for the primary statistical analyses and development of the high-risk cohort can be found at https://github.com/royadams/adams_et_al_2022_code.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data are not publicly available because they are from electronic health records approved for limited use by Johns Hopkins University investigators. Making the data publicly available without additional consent, ethical, or legal approval might compromise patients' privacy and the original ethical approval. To perform additional analyses using this data, researchers should contact AWW or SS to apply for an IRB-approved research collaboration and obtain an appropriate data use agreement.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample used was a convenience sample including all data available from the deployment of the tool under study, however, a post-hoc power analysis performed using the multiple logistic regression test in G*Power gave power 0.993 at the alpha=0.05 level.
Data exclusions	No data was excluded beyond the inclusion criteria stated in the paper.
Replication	Due to the cost and time involved in deploying a clinical decision support tool and gathering electronic medical record data no attempts were made to replicate this prospective cohort study. However, we tested the robustness of our analyses by repeating the primary analysis (adjusted comparison of patient outcomes between arms) under five modifications to the analysis or population. First, to ensure that our conclusions were not dependent on our choice of statistical method, we repeated the analysis using an alternative method, inverse probability of treatment weighting. We used logistic regression for the probability of treatment model and we stabilized and truncated weights at the 1st and 99th weight percentiles to reduce estimation variance. Second, because a patient's problem list reflects real-time diagnoses made by providers, it is less reliable than final diagnoses which reflect all information from a patient's stay. Similarly, clinicians may not always mark a present on arrival diagnosis as such. We tested the sensitivity to these issues by replacing ICD-10 codes in the patient's problem list with final diagnosis ICD-10 codes. Final diagnosis codes are more comprehensive but may include information from well after the patient's alert, leading to potential over-adjustment. Third, because we included each patient encounter separately in our data, we wished to ensure that our conclusions were not overly sensitive to a few patients with multiple encounters. We tested this sensitivity by repeating the analysis using only the first encounter for each patient. Fourth, the COVID-19 pandemic impacted both characteristics of patients in our data and the care those patients received. We tested the sensitivity to inclusion of data during the COVID-19 pandemic by repeating the analysis using only patients admitted before April 1, 2020. Finally, the list of antibiotics used to determine the time of first antibiotic administration was intentionally inclusive. To ensure that our results are not sensitive to the specific antibiotics included, we repeated the analysis using a restricted list of antibiotics that removes some antibiotics that are either unlikely to be used to treat sepsis or unlikely to be used on adult patients. The restricted list removes amoxicillin, azithromycin, cefotaxime, dapsone, erythromycin, neomycin, and rifampin from the antibiotics listed in Extended Data Table 5. These sensitivity analyses did not change the direction or significance of our findings.
Randomization	The study was not randomized and thus covariate adjustment was used. The relationship between antibiotics timing and patient outcomes may be confounded by several aspects of patient presentation. We adjusted for a range of patient variables to account for potential confounding. Similar to previous studies, we adjusted for patient age, documented sex, comorbidity history, and measurements taken during the current encounter. To account for differences due to pre-existing conditions and patient history, we adjusted for the patient's Charlson Comorbidity Index (CCI) as well as individual indicators for a history of diabetes (with and without complications), dementia, malignant tumors, or metastatic solid tumors. These conditions were identified from International Classification of Diseases, Tenth Revision (ICD-10) codes included in the patient's medical history. To account for differences in treatment arising from patient presentation at the current visit, we adjusted for several sepsis-relevant labs, vital signs, and treatments including systolic blood pressure, Glasgow Coma Score (GCS) below 15 (indicating any abnormal mental activity), temperature, white blood cell count, lactate above 2 mmol/L, and indicators for vasopressors and mechanical ventilation. We additionally adjusted for composite measures of acute patient severity including individual SOFA score components and APACHE II score. For labs and vital signs, we used the most recent measurement taken in the 24 hours prior to the alert and imputed a normal value if no measurement was observed during this period. SOFA and APACHE II scores were calculated using the worst measurements from the 24 hours prior to the alert. In cases where the alert occurred within 12 hours of presentation to the ED or admission to an in-patient unit (whichever came first), we used the worst measurements recorded within 12 hours of presentation/admission to allow for lab processing and recording delays. In addition to the variables used in previous studies, we included several additional patient and hospital related variables. In medically complex patients, clinicians may wish to obtain more information on the patients before ordering antibiotics. To account for this, we further adjusted for the presence of several comorbidities that may obscure a sepsis diagnosis including metastatic cancer, end stage renal disease, congestive heart failure, acute liver disease, gastrointestinal bleeding, and chronic obstructive pulmonary disease. These comorbidities were identified from ICD-10 codes listed as part of the patient's problem list and marked as present on arrival. Additionally, for ED patients, we adjusted for whether or not the trauma team was activated upon arrival. To account for differences in clinical practice between hospitals, we included an indicator for the admitting hospital. Finally, we included a binary indicator for presentation at the hospital after April 1, 2020, to account for differences in treatment patterns arising from the COVID-19 pandemic.
Blinding	This study was observational and thus no blinding was used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	6,877 patient encounter involving 6,264 unique patients were included in the primary analysis. The average age was 66.0 (std 17.2) and 47.0% were document as male. The average Charlson Commorbidity Index was 6.5 (std 4.0) and, at the time of the alert, the average APACHE II score was 11.8 (std 7.7) and the average SOFA score was 4.1 (std 3.4). 95.7% presented at the emergency department, 1.3% were discharged directly from the emergency department, and 44.8% were admitted to the ICU during the encounter. 42.1% met the criteria for septic shock during the encounter. For additional details, see Table 1 and Supplementary Table 1.
Recruitment	This study used electronic health records gathered naturally over the course of patient care and thus patients were not directly recruited to the study. A waiver of consent was obtained from the Johns Hopkins University internal review board (IRB No. 00252594).
Ethics oversight	The Johns Hopkins University internal review board (IRB No. 00252594).

Note that full information on the approval of the study protocol must also be provided in the manuscript.