



Research and Applications

Machine learning for early detection of sepsis: an internal and temporal validation study

Armando D. Bedoya ¹, Joseph Futoma,^{2,3} Meredith E. Clement,⁴ Kristin Corey,^{5,6} Nathan Brajer,^{5,6} Anthony Lin,^{5,6} Morgan G. Simons,^{5,6} Michael Gao,⁵ Marshall Nichols,⁵ Suresh Balu,^{5,6} Katherine Heller,² Mark Sendak ⁵ and Cara O'Brien⁷

¹Department of Medicine, Division of Pulmonary, Allergy, and Critical Care Medicine, Duke University, Durham, North Carolina, USA, ²Department of Statistics, Duke University, Durham, North Carolina, USA, ³John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA, ⁴Department of Medicine, Division of Infectious Diseases, Duke University, Durham, North Carolina, USA, ⁵Duke Institute for Health Innovation, Durham, North Carolina, USA, ⁶Duke University School of Medicine, Durham, North Carolina, USA and ⁷Department of Medicine, Durham, North Carolina, USA

Armando Bedoya and Joseph Futoma shared co-first authors.
Mark Sendak and Cara O'Brien shared co-senior authors.

Corresponding Author: Armando D. Bedoya, MD MMCI, Division of Pulmonary, Allergy, and Critical Care Medicine, Duke University, DUMC Box 102349, Hanes House, 330 Trent Drive, Room 117, Durham, NC 27710, USA; armando.bedoya@duke.edu

Received 18 July 2019; Revised 16 January 2020; Editorial Decision 12 February 2020; Accepted 10 March 2020

ABSTRACT

Objective: Determine if deep learning detects sepsis earlier and more accurately than other models. To evaluate model performance using implementation-oriented metrics that simulate clinical practice.

Materials and Methods: We trained internally and temporally validated a deep learning model (multi-output Gaussian process and recurrent neural network [MGP-RNN]) to detect sepsis using encounters from adult hospitalized patients at a large tertiary academic center. Sepsis was defined as the presence of 2 or more systemic inflammatory response syndrome (SIRS) criteria, a blood culture order, and at least one element of end-organ failure. The training dataset included demographics, comorbidities, vital signs, medication administrations, and labs from October 1, 2014 to December 1, 2015, while the temporal validation dataset was from March 1, 2018 to August 31, 2018. Comparisons were made to 3 machine learning methods, random forest (RF), Cox regression (CR), and penalized logistic regression (PLR), and 3 clinical scores used to detect sepsis, SIRS, quick Sequential Organ Failure Assessment (qSOFA), and National Early Warning Score (NEWS). Traditional discrimination statistics such as the C-statistic as well as metrics aligned with operational implementation were assessed.

Results: The training set and internal validation included 42 979 encounters, while the temporal validation set included 39 786 encounters. The C-statistic for predicting sepsis within 4 h of onset was 0.88 for the MGP-RNN compared to 0.836 for RF, 0.849 for CR, 0.822 for PLR, 0.756 for SIRS, 0.619 for NEWS, and 0.481 for qSOFA. MGP-RNN detected sepsis a median of 5 h in advance. Temporal validation assessment continued to show the MGP-RNN outperform all 7 clinical risk score and machine learning comparisons.

Conclusions: We developed and validated a novel deep learning model to detect sepsis. Using our data elements and feature set, our modeling approach outperformed other machine learning methods and clinical scores.

Key words: adult, sepsis/mortality, electronic health records/statistics and numerical data, machine learning, decision, support systems, clinical, emergency service, hospital/statistics and numerical data, hospitalization/statistics and numerical data, ROC curve, retrospective studies

INTRODUCTION

Mortality rates in patients with untreated sepsis can exceed 30%.^{1,2} As a leading cause of mortality,³ sepsis represents a significant burden to the patient, clinician, and healthcare system. Protocol-driven care bundles improve clinical outcomes,^{4,5} but require early detection of sepsis, which remains elusive even for experienced clinicians.

In 2016, a new consensus definition (Sepsis-3) was published, which utilizes the Sequential Organ Failure Assessment (SOFA) and a newly developed quick Sequential Organ Failure Assessment (qSOFA) to identify patients at risk for poor outcomes due to sepsis.⁶ The Sepsis-3 criteria have been criticized for detecting sepsis late in the clinical course.^{4,7,8} The Centers for Medicare and Medicaid Services (CMS) continue to use an older sepsis definition based on the presence of the systemic inflammatory response syndrome (SIRS) for the purposes of measuring compliance with the sepsis quality of care bundles (SEP-1 measure).^{4,7}

Quality improvement programs implemented at individual health systems have improved outcomes for patients with sepsis.^{9,10} However, overall compliance with recommended treatment remains poor. Deep learning is a suite of novel machine learning methods that have achieved performance on many challenging tasks.¹¹

The present study carries out 3 analyses to better characterize how a deep learning approach can detect sepsis early in the emergency department (ED) and pre-intensive care unit (ICU) inpatient setting. The deep learning model was specifically designed to detect the first episode of sepsis between presentation to the ED and discharge home, inpatient mortality, or transfer to an ICU. First, we compare the performance of our previously derived deep learning approach^{12,13} to clinical scores that are commonly used to identify patients at risk of sepsis. Second, we compare the performance of our model to previously published machine learning methods used to predict sepsis. Third, we test how well our model, clinical scores, and previously published machine learning methods generalize to a planned future implementation.

MATERIALS AND METHODS

Datasets

This retrospective, single-center study analyzed electronic health record (EHR) data from a quaternary academic hospital with 43 000 inpatient and 1 million outpatient visits annually. This study is reported as per the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines¹⁴ and was approved by the Duke University Health System Institutional Review Board (Pro00093721, Pro00080914).

The model development cohort consisted of all inpatient admissions that began in the ED between October 1, 2014 and December 1, 2015. Patients under the age of 18 were excluded. Hospital admissions that did not originate in the ED (eg, direct admission; scheduled surgery) and ED encounters that did not result in inpatient admission were also excluded from the model development cohort. Patients who developed sepsis within 1 h of presentation to the ED were excluded. Encounter data began at presentation to the ED. Encounters that did not result in sepsis ended at time of discharge, time of death, or time of ICU transfer. Encounters that did result in sepsis ended at time of the first sepsis episode. All data after discharge, the first sepsis episode, ICU transfer, or death were excluded from model development. Patients who developed sepsis after transfer to an ICU were included and treated as control cases. Curated features included structured static variables, such as demographic, encounter, and pre-admission

comorbidity data, as well as dynamic variables, such as vital sign, medication, and lab data. Vital sign measurements, medication administrations, and lab collections that occurred between the encounter start and end times were included.

There is no gold standard for the definition of sepsis. Various definitions of sepsis have been described in the literature which partition out specific populations to meet study or epidemiological needs. Sepsis was defined in our data by the presence of 2 or more SIRS criteria, a blood culture order, and at least one element of end-organ failure ([Supplementary Table S1](#)). Our definition was based upon prior efforts by our study team.¹⁵ A similar sepsis definition has been used for model development efforts at peer institutions that developed at least 2 other published models, and this definition aligns with the CMS definition.^{16,17}

We compared our sepsis definition with Sepsis 1, Sepsis 3, and the Centers for Disease Control (CDC) Adult Sepsis Event.¹⁷ The Sepsis-1 and Sepsis-3 definitions were computed using SIRS and qSOFA criteria. An order for any culture served as a proxy for clinician suspicion for infection to enable the Sepsis-1 and Sepsis-3 definition to be automatically computed from the EHR without manual chart review. The CDC Adult Sepsis Event surveillance definition is based on the Sepsis-3 framework of suspected infection with organ dysfunction.¹⁸ Sensitivity, specificity, positive predictive value (PPV), and negative predictive value were calculated for each definition using CDC Adult Sepsis Events as the gold standard.

A separate temporal validation cohort was curated from the same site. The cohort was not limited to inpatient admissions but included all ED encounters between March 1, 2018 and August 31, 2018. The same variables, inclusion and exclusion criteria, and outcome definition were applied. Unlike the model development cohort, the temporal validation cohort included encounters that began in the ED that did not result in inpatient admission.

A total of 86 variables were automatically curated¹⁹ for each cohort, including patient demographics, comorbidities, vital signs, medication administrations, and labs ([Supplementary Table S2](#)). In total, the model development cohort contained over 32 million data points.

Model development

We built on prior work coupling multi-output Gaussian processes (MGPs) and recurrent neural networks (RNNs) (hereafter called MGP-RNN).^{12,13} RNNs are a form of deep learning designed to ingest time series data and handle sequences of variable length.²⁰ A core feature of any deep learning method is the ability to capture complex relationships between input variables. RNNs can use a patient's complete pre-encounter and encounter data to predict an outcome while maintaining temporal relationships.^{21–23} RNNs generally require evenly spaced inputs, even if the overall lengths of encounters differ. A variety of imputation strategies have been used to model inputs that are irregularly sampled and often missing in EHR data,^{24–26} including multitask learning, which models relationships between time series.²⁷ An MGPs are a type of multitask learning that is probabilistic and maintains uncertainty about the true value.

Dynamic features (eg, vitals; labs) are sampled every hour from the MGP along with missingness indicator variables and fed into the RNN. Static features are replicated every hour and fed into the RNN. No minimum amount of data is required to generate a risk score. At each timepoint t , the likelihood of sepsis is computed and evaluated against whether or not the patient develops sepsis between time t and t plus 4 h.

Table 1. Baseline characteristics of internal development and validation cohorts (90% and 10% of full data), and of temporal validation cohort

| Baseline characteristic of cohort | Development, <i>n</i> (%) (<i>N</i> = 38 682) | Septic development, <i>n</i> (%) (<i>N</i> = 7347) | Internal validation, <i>n</i> (%) (<i>N</i> = 4297) | Septic internal validation, <i>n</i> (%) (<i>N</i> = 813) | Temporal validation, <i>n</i> (%) (<i>N</i> = 39 786) | Septic temporal validation, <i>n</i> (%) (<i>N</i> = 2562) |
|---|--|---|--|--|--|---|
| Age (years), mean ± SD | 55.9 (18.7) | 59.8 (17.1) | 56.2 (18.6) | 59.6 (17.3) | 50.4 (19.5) | 59.7 (17.0) |
| Sex male | 18 203 (47.1) | 4005 (54.5) | 2050 (47.7) | 434 (53.4) | 18 272 (45.9) | 1418 (55.3) |
| Weight (lbs), mean ± SD | 158.8 (19.7) | 160.2 (19.8) | 158.4 (19.4) | 159.7 (19.8) | 185.8 (72.1) | 184.4 (60.1) |
| Admission source | | | | | | |
| Home/non-healthcare facility | 30 063 (77.7) | 5072 (69.0) | 3363 (78.3) | 577 (71.0) | 34 848 (87.6) | 1854 (72.4) |
| Transfer from hospital | 4930 (12.7) | 1498 (20.4) | 534 (12.4) | 147 (18.1) | 2877 (7.2) | 538 (21.0) |
| Missing/other | 3689 (9.5) | 777 (10.6) | 400 (9.3) | 89 (10.9) | 2061 (5.2) | 170 (6.6) |
| Admission type | | | | | | |
| Elective | 11 854 (30.6) | 571 (7.8) | 1338 (31.1) | 60 (7.4) | 5620 (14.1) | 138 (5.4) |
| Emergency | 16 478 (42.6) | 4813 (65.5) | 1797 (41.8) | 522 (64.2) | 30 048 (75.5) | 1917 (74.8) |
| Urgent | 10 342 (26.7) | 1963 (26.7) | 1162 (27.0) | 231 (27.8) | 4118 (10.4) | 507 (19.8) |
| Race | | | | | | |
| Black or African American | 11 390 (29.4) | 2329 (31.7) | 1252 (29.1) | 255 (31.4) | 15 805 (39.7) | 931 (36.3) |
| Caucasian/White | 24 317 (62.9) | 4661 (63.4) | 2681 (62.4) | 499 (61.4) | 19 701 (49.5) | 1454 (56.8) |
| Missing/other | 2975 (7.7) | 357 (4.9) | 364 (8.5) | 59 (7.3) | 4280 (10.8) | 177 (6.9) |
| Comorbidities | | | | | | |
| Congestive heart failure | 5656 (14.6) | 1576 (21.5) | 627 (14.6) | 175 (21.6) | 3329 (8.4) | 469 (18.3) |
| Valvular disease | 5288 (13.7) | 1298 (17.7) | 544 (12.7) | 137 (16.9) | 1975 (5.0) | 232 (9.1) |
| Peripheral vascular disease | 4283 (11.1) | 1016 (13.8) | 474 (11.0) | 119 (14.6) | 1678 (4.2) | 198 (7.7) |
| Hypertension | 18 251 (47.2) | 4114 (56.0) | 2009 (46.8) | 445 (54.7) | 11 874 (29.8) | 1014 (39.6) |
| Other neurological disorders | 6725 (17.4) | 1974 (26.9) | 731 (17.0) | 226 (27.8) | 2538 (6.4) | 239 (9.3) |
| Pulmonary circulation disorders | 6917 (17.9) | 1830 (24.9) | 779 (18.1) | 192 (23.6) | 4047 (10.2) | 338 (13.2) |
| Diabetes mellitus without chronic complications | 6071 (15.7) | 1394 (19.0) | 685 (15.9) | 169 (20.8) | 2896 (7.3) | 225 (8.8) |
| Renal failure | 6188 (16.0) | 1876 (25.5) | 673 (15.7) | 216 (26.6) | 3829 (9.6) | 600 (23.4) |
| Solid tumor without Metastasis | 4711 (12.2) | 809 (11.0) | 525 (12.2) | 93 (11.4) | 3879 (9.7) | 395 (15.4) |
| Coagulopathy | 4503 (11.6) | 1588 (21.6) | 497 (11.6) | 173 (21.3) | 1558 (3.9) | 341 (13.3) |
| Obesity | 5542 (14.3) | 1203 (16.4) | 598 (13.9) | 133 (16.4) | 3213 (8.1) | 207 (8.1) |
| Fluid and electrolyte disorders | 10 204 (26.4) | 3221 (43.8) | 1110 (25.8) | 353 (43.4) | 4855 (12.2) | 668 (26.1) |
| Anemia | 9242 (23.9) | 2763 (37.6) | 1055 (24.6) | 309 (38.0) | 3327 (8.4) | 396 (15.5) |
| Depression | 6308 (16.3) | 1526 (20.8) | 715 (16.6) | 168 (20.7) | 2721 (6.8) | 174 (6.8) |
| Prior sepsis encounters in past year | | | | | | |
| 0 | 36 634 (94.7) | 6363 (86.6) | 4103 (95.5) | 727 (89.4) | 38 872 (97.7) | 2319 (90.5) |
| 1 | 1514 (3.9) | 688 (9.4) | 149 (3.5) | 64 (7.9) | 681 (1.7) | 165 (6.4) |
| 2 or more | 534 (1.4) | 296 (4.0) | 45 (1.0) | 22 (2.7) | 233 (0.6) | 78 (3.0) |
| Overall in-hospital mortality (%) | 1257 (3.2) | 696 (9.5) | 121 (2.8) | 59 (7.3) | 577 (1.5) | 337 (13.2) |
| Overall length of stay (h), median (25%–75%) | 95 (57–168) | 167 (95–318) | 95 (55–168) | 167 (94–315) | 13 (5–90) | 172 (95–342) |
| Overall rate of ICU admission (%) | 5870 (15.2) | 1530 (20.8) | 646 (15.0) | 166 (20.4) | 4598 (11.6) | 1148 (44.8) |
| Septic (%) | 7347 (19.0) | 7347 (100.0) | 813 (18.9) | 813 (100.0) | 2562 (6.4) | 2562 (100.0) |

Note: For each cohort, characteristics are also broken out among the subgroup of patients who acquire sepsis.

The model development cohort was divided into training, test, and internal validation subsets. The training subset contained 80% of all encounters. The remaining encounters were evenly split between a test subset for hyperparameter selection and an internal validation subset. The internal validation subset was blinded to all methods until final evaluation. Each model was trained on the training subset until time of sepsis. For control encounters, data until a randomly chosen timepoint mid-encounter was used. Every model generated a risk score each hour starting 1 h after admission.

The performance of MGP-RNN was assessed using 2 sets of comparisons. First, we compared the performance of the MGP-RNN to SIRS,²⁸ National Early Warning Score (NEWS),²⁹ and qSOFA.⁶ Next, we compared the performance of the MGP-RNN to a Lasso-penalized Cox regression (CR),³⁰ random forest (RF),³¹ and

penalized logistic regression.³² Both sets of comparisons assess global performance of methods as well as performance as time passes following presentation to the ED.²⁵ This analysis demonstrates the ability of the various approaches to detect sepsis as early in the hospital course as possible.

Temporal validation

Finally, we compared the performance of MGP-RNN against all 7 clinical scores and machine learning methods on a temporal validation cohort. The temporal validation cohort represents a planned future implementation in an adult ED.

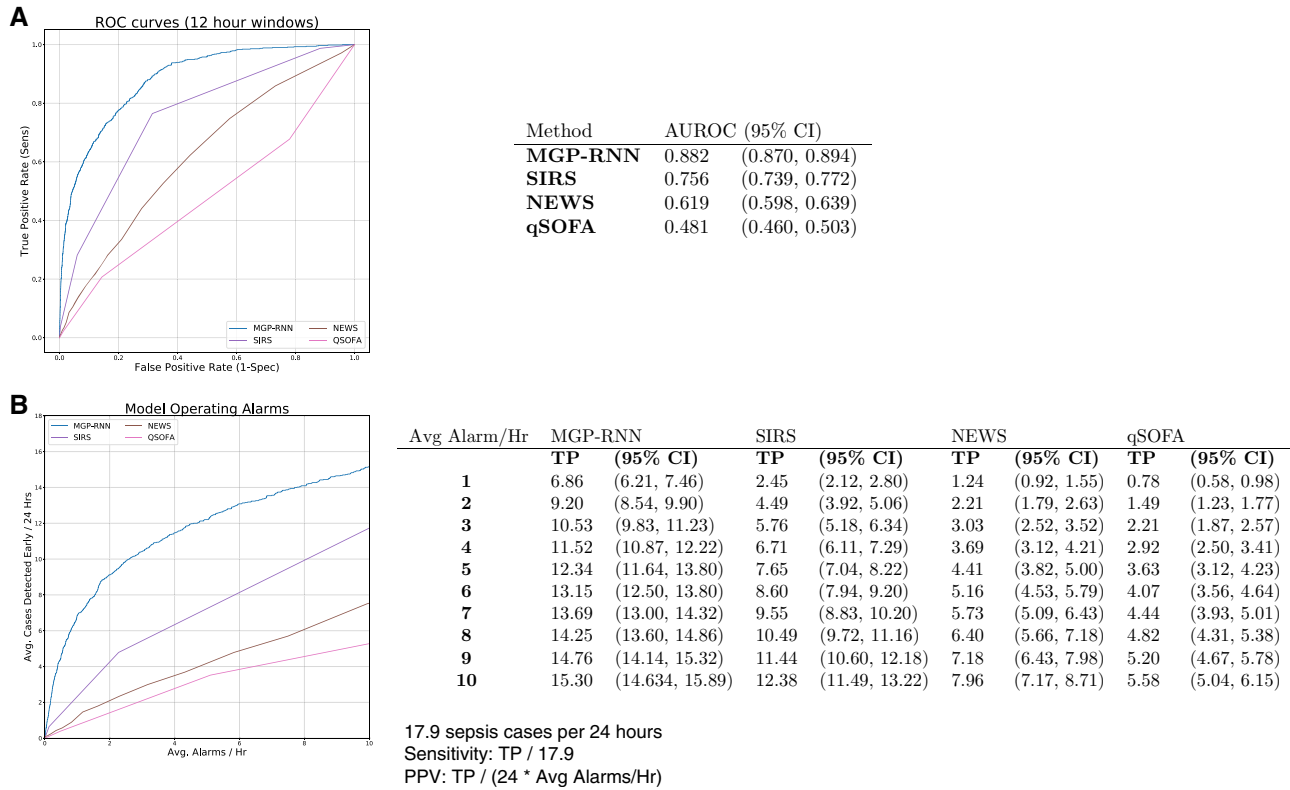


Figure 1. Results of our deep learning model compared with the clinical scores methods. (A) ROC curves for the MGP-RNN and the 3 clinical scores considered, SIRS, NEWS, and qSOFA is shown. The accompanying table lists C-statistics with bootstrap confidence intervals. (B) The average number of sepsis cases each day we expect to detect early before a definition for sepsis is met (ie, a more interpretable version of sensitivity), as a function of how many alarms each method would produce each hour is shown. We limit the average alarms per hour to less than 10, as this is the operating range at which we expect to use in practice. There were an average of 17.9 sepsis cases per 24-h period in the dataset, so sensitivity can be recovered by dividing the reported y-axis value in panel B by 17.9. Positive predictive value at a particular threshold can be recovered by dividing the reported y-axis value by 24 times the reported x-axis value (ie, the average number of alarms per 24-h period). MGP-RNN, multi-output Gaussian process and recurrent neural network; NEWS, national early warning score; qSOFA, quick Sequential Organ Failure Assessment; SIRS, systemic inflammatory response syndrome.

Statistical analysis

Evaluation metrics included area under the receiver operating characteristic curve (AUC). Lastly, we fix the number of alerts allowed per hour and report the number of sepsis cases identified early per day. This reflects the workflow constraint of needing to limit the number of alerts fired to front-line clinicians. Model performance is calculated on the 10% internal validation subset and on the temporal validation cohort.

Models generate risk scores every hour and we calculate performance using 2 approaches. To assess global performance, similar to prior work,^{13,26,33,34} metrics are calculated using the maximum score within independent 12-h windows. True positives are high-risk scores during 12-h blocks immediately preceding a sepsis event. False positives are high-risk scores during 12-h blocks not immediately preceding a sepsis event. To assess performance as time passes following presentation to the ED, metrics are calculated using the maximum score within windows ranging in size from 1-h to 12 h. True positives are high-risk scores during a window followed by a sepsis event within 4 h. False positives are high-risk scores during a window not followed by a sepsis event within 4 h. All model evaluations are completed without an alert ‘snooze’, a time period during which risk scores are suppressed and not considered.

All methods were implemented using the numpy (version 1.14.0), scikit-learn (version 0.18.1), and TensorFlow (version 1.6.0) python packages.

RESULTS

In the model development cohort, there were 42 979 admissions and sepsis developed in 8160 (19.0%) admissions. In the temporal validation cohort, there were 39 786 encounters and sepsis developed in 2562 (6.4%) encounters. Table 1 presents demographic and clinical characteristics of the model development, internal validation, and temporal validation cohorts. Sepsis was observed early in the hospital course. In the model development cohorts, 3100 (38%) sepsis cases occurred between presentation to the ED and inpatient admission. Furthermore, in the model development cohorts, 791 (9.7% overall; 25.5% of those in the ED) sepsis cases occurred within 1 h of presentation to the ED, and 372 (4.6%) sepsis cases occurred within 1 h of inpatient admission. Supplementary Figure S1 shows the full distribution of time of sepsis within both the model development and temporal validation cohorts. Supplementary Table S3 illustrates the performance of our sepsis definition, Sepsis-1, and Sepsis 3 in detecting CDC Adult Sepsis Events. Notably, our sepsis definition had the highest PPV for identifying patient that ultimately received 4 days of antibiotics to meet the CDC Adult Sepsis Event definition.

MGP-RNN outperformed SIRS, qSOFA, and NEWS. Figure 1A shows AUC and Figure 1B shows operational metrics fixing the number of alarms per hour. To minimize alarm fatigue, a workflow can be designed that limits the number of alerts prioritized for a clinician to review per hour. Allowing 3 alarms per hour, MGP-RNN

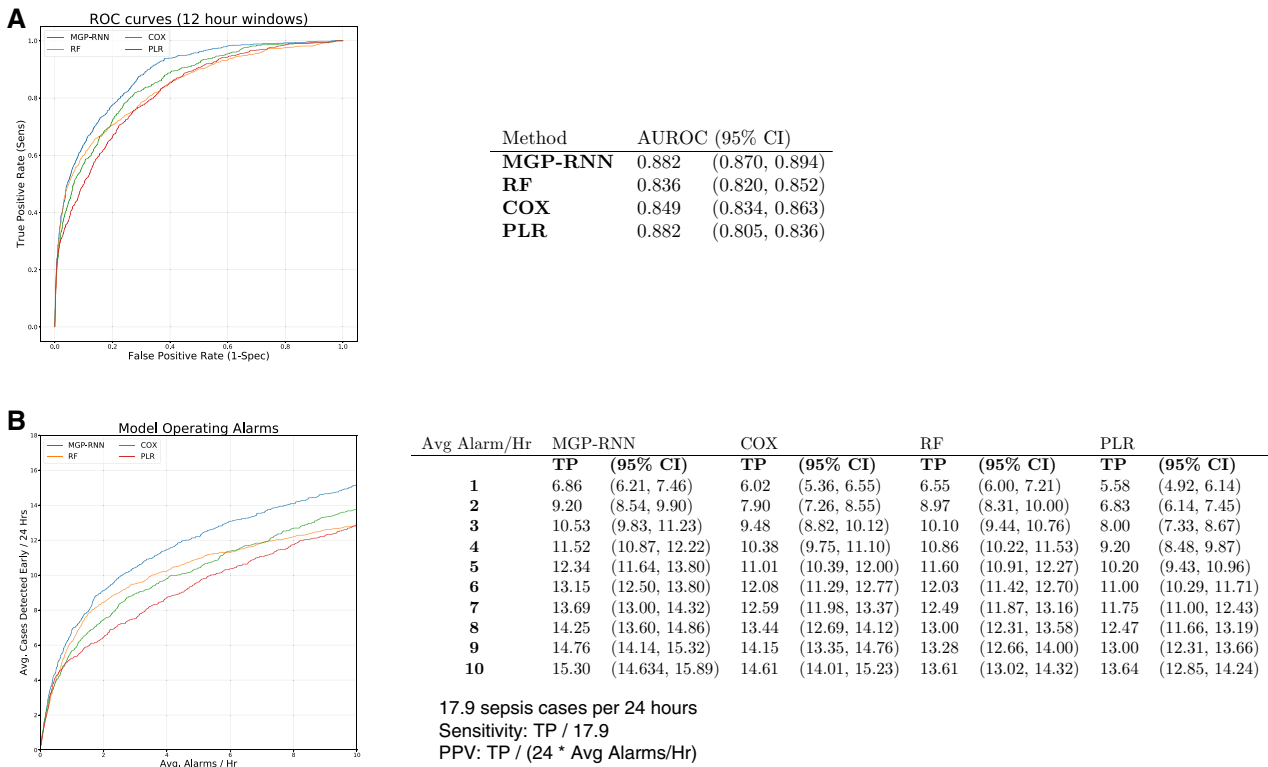


Figure 2. Results of our deep learning model compared with the other machine learning models. (A) ROC curves for the MGP-RNN and the 3 other machine learning models considered, Cox regression, penalized logistic regression, and random forest is shown. The accompanying table lists C-statistics with bootstrap confidence intervals. (B) The average number of sepsis cases each day we expect to detect early before a definition for sepsis is met (ie, a more interpretable version of sensitivity), as a function of how many alarms each method would produce each hour is shown. We limit the average alarms per hour to less than 10, as this is the operating range at which we expect to use in practice. There were an average of 17.9 sepsis cases per 24-h period in the dataset, so sensitivity can be recovered by dividing the reported y-axis value in panel B by 17.9. Positive predictive value at a particular threshold can be recovered by dividing the reported y-axis value by 24 times the reported x-axis value (ie, the average number of alarms per 24-h period). MGP-RNN, multi-output Gaussian process and recurrent neural network; PLR, penalized logistic regression; RF, random forest.

captured 10.5 out of 17.9 sepsis cases per day, compared to 5.76 for SIRS, 3.03 for NEWS, and 2.21 for qSOFA.

MGP-RNN also outperformed machine learning methods used in previously published sepsis prediction models. Figure 2A shows AUC for each approach and Figure 2B shows operational metrics fixing the number of alarms per hour. Allowing 3 alarms per hour, MGP-RNN captured 10.5 out of 17.9 sepsis cases per day, compared to 9.48 for CR, 8.00 for logistic regression (LR), and 10.10 for RF.

At this threshold yielding an average of 3 alarms per hour, MGP-RNN detects sepsis a median of 5 h in advance (with 25% and 75% quantiles of 2 and 20 h). Supplementary Figure S2 shows the full distribution of how far in advance MGP-RNN detects sepsis in both the internal and temporal validation cohorts. Supplementary Figure S3 also shows the precision-recall curves for MGP-RNN versus the clinical scores and machine learning methods on the internal cohort.

When applied to the temporal validation cohort, MGP-RNN continues to outperform all 7 clinical risk score and machine learning comparisons. Figure 3A highlights the AUC for each approach across internal and temporal validation cohorts; discrimination generally improves on the temporal cohort. Figure 3B and C shows AUC and PPV as a function of hours after presentation to the ED. Not only does MGP-RNN discriminate better than all comparisons on a cohort of all comers to an adult ED, but MGP-RNN performs

best across metrics at almost all points during encounters. Figure 4A and B illustrate the superior performance of MGP-RNN on a temporally distinct time period. Supplementary Figure S4 also shows the precision-recall curves for MGP-RNN versus the clinical scores and machine learning methods on the temporal cohort.

Additional results in the Supplementary Material show model interpretability, calibration, and the effect of shortening the size of the independent 12-h time windows used for evaluation (Supplementary Figures S5–S7).

DISCUSSION

We developed a deep learning approach to detect sepsis early and validated the model on a cohort of inpatient admissions as well as a temporal cohort of adults presenting to the ED. This approach uses comprehensive data from a patient's hospital encounter to accurately detect sepsis from presentation to the ED until ICU transfer or hospital discharge.

Consistent with prior studies^{16,32,35} we find that machine learning models predict sepsis more accurately than clinical scores. These findings are clinically important, because qSOFA has been recommended as the screening tool for clinicians to use to identify patients for evaluation and potential escalation of care.³⁶ We find that across metrics, qSOFA performs poorly at detecting sepsis early, also

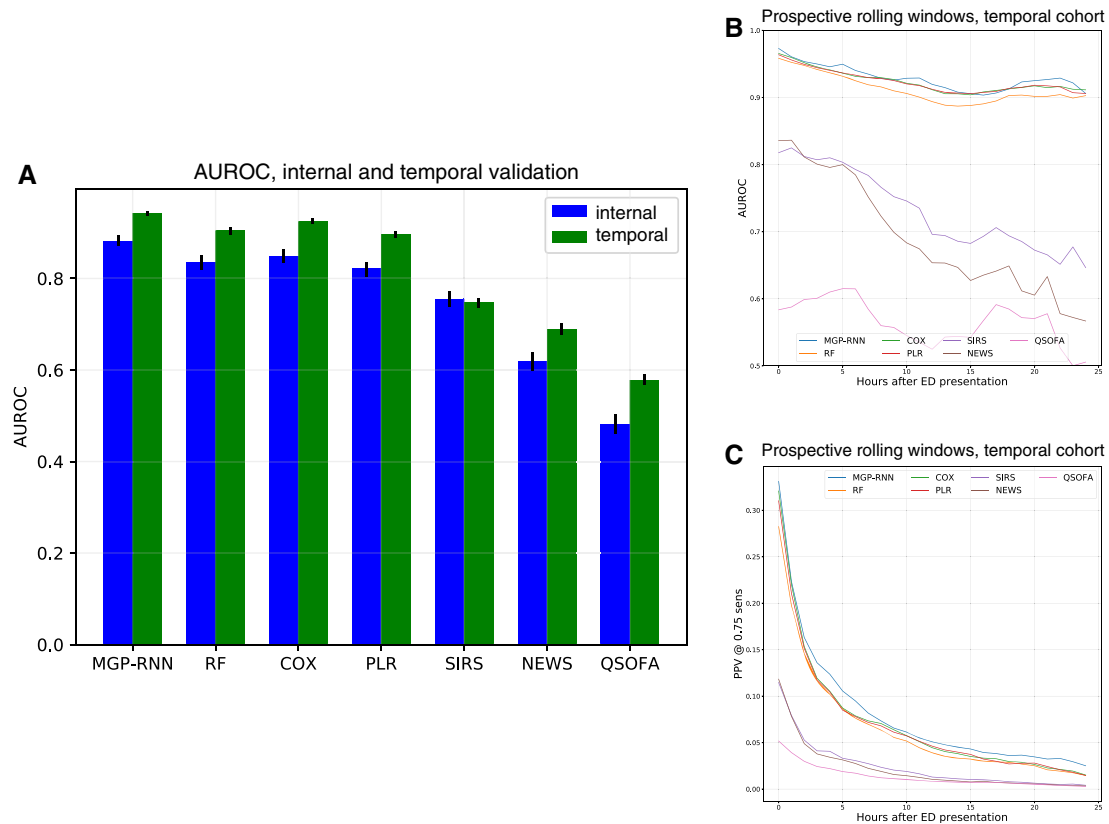


Figure 3. (A) Compares the AUROC obtained from the internal validation cohort with the AUROC from the temporal validation cohort for each method, along with bootstrap confidence intervals. (B) The AUROC as a function of hours after presentation to the ED for the temporal validation cohort for each method, limited to the first 24 h following initial presentation is shown. (C) The PPV at 75% sensitivity for each method as a function of number of hours after presentation to the ED is shown.

consistent with prior results.³⁷ Health systems with fixed workforce capacity looking to implement clinical decision support within an EHR may consider investment in infrastructure to leverage machine learning methods. Otherwise, fixing the number of alerts per hour, we find that SIRS consistently outperforms qSOFA in detecting sepsis early.

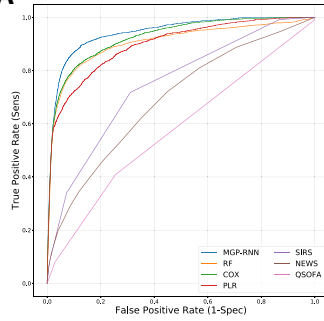
Compared to previously published machine learning methods (CR, LR, and RF), we demonstrated the superior performance of MGP-RNN. Across AUC and operational metrics, MGP-RNN surpassed these methods to detect sepsis within 4 h. MGP-RNN detects more sepsis cases than other machine learning models at every number of fixed alarms per hour (Figure 2B). This performance gain is likely due to the coupling of the MGP with the RNN to better impute continuous functions for all vital sign and lab data. If a lab value is missing, the MGP will use learned relationships from the other available continuous features to calculate a distribution of possible values for the specific patient.

This study compared multiple previously published machine learning methods head-to-head on the same dataset, because comparing models across studies is non-trivial. Prior studies use a variety of outcome definitions, cohort definitions, model inputs, and statistical methods. Most sepsis models were developed on cohorts of ICU patients^{16,30,38,39} and nearly all use the publicly available MIMIC dataset.⁴⁰ Many models use sepsis ICD codes as the outcome definition^{38,39,41,42} and predict sepsis at any point during an encounter, which is not directly actionable for frontline clinicians

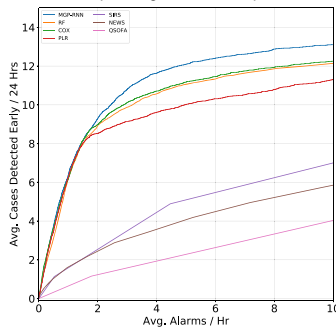
trying to follow SEP-1 bundle recommendations. In addition, nearly all models use static model inputs.^{30,38,39,41–43} While neural networks have been applied to sepsis prediction,^{44,45} none have been configured to use the entire time series of repeated measurements to detect sepsis within a window of time.

We further validated MGP-RNN on a more recent cohort that not only differs temporally but includes ED visits that do not result in admission. In comparison to the internal validation subset, performance characteristics improve for the temporal validation cohort. We suspect the improvement occurs because sepsis occurred in 19.0% of admitted patients, but only 6.4% of patients presenting to the ED. By including many low-risk patients, the improvement in AUC can be expected. The temporal validation results demonstrate the robustness of MGP-RNN within the implementation setting, where at the time of presentation it is unknown whether a patient will be admitted. The results further demonstrate MGP-RNN’s ability to detect sepsis better than all other methods at various points during the hospital course. These findings laid the groundwork for implementing MGP-RNN in the ED and a prospective evaluation is currently underway (ClinicalTrials.gov identifier: NCT03655626). Furthermore, our general approach can be scaled to other institutions, although each new local context would likely require retraining and possibly even the development of new models.

This study has a number of limitations. First, sepsis does not have a universally accepted definition. We adapted a definition similar to the clinical criteria outlined by CMS and this approach has potential weaknesses. Our definition does not address elevated but

A ROC Curves (12 Hour Windows), Temporal Cohort

| Method | AUROC | (95% CI) |
|---------|-------|----------------|
| MGP-RNN | 0.943 | (0.938, 0.948) |
| RF | 0.905 | (0.896, 0.913) |
| COX | 0.925 | (0.919, 0.931) |
| PLR | 0.897 | (0.889, 0.904) |
| SIRS | 0.748 | (0.737, 0.757) |
| NEWS | 0.690 | (0.678, 0.703) |
| qSOFA | 0.578 | (0.567, 0.590) |

B Model Operating Alarms, Temporal Cohort

| Avg Alarm/Hr | MGP-RNN | COX | RF | PLR | SIRS | NEWS | qSOFA |
|--------------|----------------------|----------------------|----------------------|----------------------|-------------------|-------------------|-------------------|
| | TP (95% CI) | TP (95% CI) | TP (95% CI) | TP (95% CI) | TP (95% CI) | TP (95% CI) | TP (95% CI) |
| 1 | 6.37 (6.04, 6.68) | 6.18 (5.86, 6.48) | 6.16 (5.81, 6.46) | 6.25 (5.95, 6.57) | 1.59 (1.43, 1.75) | 1.63 (1.43, 1.83) | 0.65 (0.56, 0.74) |
| 2 | 9.29 (8.96, 9.60) | 8.98 (8.67, 9.29) | 8.89 (8.56, 9.19) | 8.53 (8.23, 8.81) | 2.54 (2.36, 2.73) | 2.44 (2.20, 2.68) | 1.23 (1.07, 1.39) |
| 3 | 10.80 (10.48, 11.10) | 10.12 (9.79, 10.39) | 9.87 (9.57, 10.14) | 9.13 (8.84, 9.42) | 3.49 (3.27, 3.73) | 3.09 (2.85, 3.33) | 1.58 (1.43, 1.75) |
| 4 | 11.64 (11.40, 11.87) | 10.72 (10.43, 10.97) | 10.56 (10.28, 10.82) | 9.61 (9.32, 9.87) | 4.45 (4.16, 4.75) | 3.57 (3.31, 3.84) | 1.93 (1.77, 2.11) |
| 5 | 12.10 (11.84, 12.33) | 11.17 (10.89, 11.43) | 11.06 (10.79, 11.29) | 10.02 (9.73, 10.28) | 5.10 (4.81, 5.39) | 4.07 (3.79, 4.34) | 2.28 (2.11, 2.47) |
| 6 | 12.41 (12.18, 12.62) | 11.46 (11.19, 11.72) | 11.33 (11.08, 11.58) | 10.32 (10.02, 10.57) | 5.48 (5.20, 5.76) | 4.49 (4.21, 4.76) | 2.63 (2.45, 2.83) |
| 7 | 12.62 (12.41, 12.81) | 11.75 (11.49, 11.98) | 11.63 (11.38, 11.85) | 10.52 (10.25, 10.80) | 5.86 (5.59, 6.13) | 4.88 (4.59, 5.18) | 2.98 (2.78, 3.20) |
| 8 | 12.88 (12.62, 13.05) | 11.93 (11.69, 12.18) | 11.85 (11.61, 12.07) | 10.79 (10.51, 11.10) | 6.24 (5.97, 6.52) | 5.22 (4.92, 5.50) | 3.33 (3.11, 3.58) |
| 9 | 13.00 (12.80, 13.17) | 12.11 (11.87, 12.35) | 11.99 (11.76, 12.20) | 11.08 (10.81, 11.36) | 6.62 (6.35, 6.90) | 5.54 (5.23, 5.84) | 3.68 (3.43, 3.93) |
| 10 | 13.11 (12.93, 13.27) | 12.25 (12.02, 12.49) | 12.14 (11.90, 12.35) | 11.31 (11.02, 11.58) | 7.00 (6.73, 7.29) | 5.85 (5.52, 6.17) | 4.03 (3.77, 4.31) |

14.4 sepsis cases per 24 hours
Sensitivity: TP / 14.4
PPV: TP / (24 * Avg Alarms/Hr)

Figure 4. Results for the temporal validation cohort (analogous to Figures 1 and 2, which show results on the internal validation cohort.) (A) ROC curves and (B) the operating alarms are shown. There were an average of 14.4 sepsis cases per 24-h period in the dataset, so sensitivity can be recovered by dividing the reported y-axis value in panel B by 14.4. Positive predictive value at a particular threshold can be recovered by dividing the reported y-axis value by 24 times the reported x-axis value (ie, the average number of alarms per 24-h period).

stable vital signs or abnormal laboratory values due to chronic organ dysfunction. We also did not include markers of acute respiratory dysfunction, a component of the CMS SEP-1 measure, due to variable reliability of data capture within our EHR. Although multiple sepsis definitions were compared in a prior analysis,¹⁵ a single definition was selected to train all machine learning models. Future work will have to assess model performance across multiple sepsis definitions. Second, this is a single-site study that describes development, internal, and temporal validation all within the same hospital. Another limitation of our study is the low PPV at high sensitivities; however, the low PPV is similar to other EHR-based sepsis prediction models.^{46–48}

Although the model is not tested on a geographically distinct population, use of a temporal split cohort does demonstrate robustness of model performance.⁴⁹ Future work with external partners to evaluate model performance will need to be conducted to demonstrate geographic generalizability. Furthermore, for models intended to be implemented within a local setting, we have previously shown that machine learning methods developed on locally curated EHR data can outperform models developed on national datasets.¹⁹ Finally, because MGP-RNN does not infer causal relationships, frontline clinicians will not have insight into factors driving sepsis risk. We do provide a variable importance graph in the [Supplementary Figure S5](#), but the relationship between variables and sepsis is not necessarily causal.

In conclusion, this study couples probabilistic continuous function imputation for dynamic variables with a downstream deep learning model to calculate risk of sepsis. MGP-RNN is comprehensive, including repeated measurements of labs and vitals, as well as all administrations of medications from the entirety of a patient's

hospital encounter. We demonstrate that using our data elements and feature set, our modeling approach outperformed both clinical scores and previously published machine learning methods to detect sepsis early within cohorts of admitted patients and patients presenting to the ED.

FUNDING

This work was partially funded by the Duke Institute for Health Innovation; HHS | NIH | National Institute of Allergy and Infectious Diseases (NIAID) grant number T32-AI007392 (to MEC); National Defense Science and Engineering Graduate (NDSEG) Fellowship (to JF); Duke Institute for Health Innovation Clinical Research & Innovation Scholarship (to NB, AL, KC, and MGS); and NSF Faculty Early Career Development Program (CAREER) Award (to KH).

AUTHOR CONTRIBUTIONS

The authors meet criteria for authorship as recommended by the International Committee of Medical Journal Editors. ADB, JF, MEC, KC, MGS, and MS wrote the initial manuscript. JF designed and implemented the deep learning model and other machine learning models, conducted the evaluations, and did the statistical analyses. JF, MS, SB, MG, and MN designed the evaluation framework and experiments. KC, AL, MS, MG, MN, SB, MGS, and NB did the data preparation and data cleaning. KC and MS curated the temporal validation cohort. ADB, MEC, and CO'B did the manual clinical validation of the features. All authors contributed to the overall design of the study and contributed to the production of the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

ADB, JF, MEC, NB, AL, MG, MN, SB, KH, MS, and CO'B are the named inventors of the Sepsis Watch deep learning model, which was licensed from Duke University by Cohere Med, Inc. (but does not hold any equity in Cohere Med, Inc.). KC and MGS declared no conflict of interest statement.

REFERENCES

- Liu V, Escobar GJ, Greene JD, *et al.* Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA* 2014; 312 (1): 90–2.
- Rhee C, Dantes R, Epstein L, *et al.*; for the CDC Prevention Epicenter Program. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. *JAMA* 2017; 318 (13): 1241–9.
- Epstein L, Dantes R, Magill S, Fiore A. Varying estimates of sepsis mortality using death certificates and administrative codes—United States, 1999–2014. *MMWR Morb Mortal Wkly Rep* 2016; 65 (13): 342–5.
- Levy MM, Dellinger RP, Townsend SR, *et al.* The Surviving Sepsis Campaign: results of an international guideline-based performance improvement program targeting severe sepsis. *Intensive Care Med* 2010; 36 (2): 222–31.
- Seymour CW, Gesten F, Prescott HC, *et al.* Time to treatment and mortality during mandated emergency care for sepsis. *N Engl J Med* 2017; 376 (23): 2235–44.
- Seymour CW, Liu VX, Iwashyna TJ, *et al.* Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016; 315 (8): 762–74.
- Cortes-Puch I, Hartog CS. Opening the debate on the new sepsis definition change is not necessarily progress: revision of the sepsis definition should be based on new scientific insights. *Am J Respir Crit Care Med* 2016; 194: 16–8.
- Bhattacharjee P, Edelson DP, Churpek MM. Identifying patients with sepsis on the hospital wards. *Chest* 2017; 151 (4): 898–907.
- Schorr C, Odden A, Evans L, *et al.* Implementation of a multicenter performance improvement program for early detection and treatment of severe sepsis in general medical-surgical wards. *J Hosp Med* 2016; 11 (Suppl 1): S32–39.
- Arabi YM, Al-Dorzi HM, Alamry A, *et al.* The impact of a multifaceted intervention including sepsis electronic alert system and sepsis response team on the outcomes of patients with sepsis and septic shock. *Ann Intensive Care* 2017; 7 (1): 57.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25 (1): 44–56.
- Futoma J, Hariharan S, Heller K. Learning to detect sepsis with a multi-task Gaussian process RNN classifier. In: ICML'17: proceedings of the 34th International Conference on Machine Learning - Volume 70; August 2017: 1174–82.
- Futoma J, Hariharan S, Heller K, *et al.* An improved multi-output Gaussian process RNN with real-time validation for early sepsis detection. In: Finale D-V, Jim F, David K, Rajesh R, Byron W, Jenna W, eds. Proceedings of the 2nd Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research: PMLR; 2017: 243–54.
- Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; 350: g7594.
- Lin AL, Sendak M, Bedoya AD, *et al.* Evaluating sepsis definitions for clinical decision support against a definition for epidemiological disease surveillance. *bioRxiv* 2019: 648907.
- Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med* 2015; 7 (299): 299ra122.
- Amland RC, Sutariya BB. Quick sequential [sepsis-related] organ failure assessment (qSOFA) and St. John sepsis surveillance agent to detect patients at risk of sepsis: an observational cohort study. *Am J Med Qual* 2018; 33 (1): 50–7.
- Hospital Toolkit for Adult Sepsis Surveillance. Secondary Hospital Toolkit for Adult Sepsis Surveillance. 2018. https://www.cdc.gov/sepsis/pdfs/Sepsis-Surveillance-Toolkit-Mar-2018_508.pdf. Accessed July 8, 2019.
- Corey KM, Kashyap S, Lorenzi E, *et al.* Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. *PLoS Med* 2018; 15 (11): e1002701.
- Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9 (8): 1735–80.
- Lipton ZC, Kale DC, Elkan C, Wetzel RC. Learning to diagnose with LSTM recurrent neural networks. *CoRR* 2015; abs/1511.0.
- Lipton ZC, Kale DC, Wetzel RC. Directly modeling missing data in sequences with RNNs: improved classification of clinical time series. In: Machine Learning for Healthcare Conference; 2016: 253–70.
- Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 2017; 24 (2): 361–70. [27521897]
- Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. *Sci Rep* 2018; 8 (1): 6085.
- Meyer A, Zverinski D, Pfahringer B, *et al.* Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med* 2018; 6 (12): 905–14.
- Wiens J, Gutttag J, Horvitz E. Patient risk stratification with time-varying parameters: a multitask learning approach. *J Mach Learn Res* 2016; 17 (1): 2797–819.
- Bonilla E, Chai KM, Williams C. Multi-Task Gaussian Process Prediction. In: proceedings of the 20th Anniversary Conference on Neural Information Processing Systems; 2008: 153–60.
- Levy MM, Fink MP, Marshall JC, *et al.*; for the International Sepsis Definitions Conference. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Intensive Care Med* 2003; 29 (4): 530–8.
- Williams B, Alberti G, Ball C, Bell D, Binks R, Durham L. *National Early Warning Score (NEWS): Standardising the Assessment of Acute-Illness Severity in the NHS*. London: The Royal College of Physicians; 2012.
- Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med* 2018; 46 (4): 547–53.
- Taylor RA, Pare JR, Venkatesh AK, *et al.* Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016; 23 (3): 269–78.
- Desautels T, Calvert J, Hoffman J, *et al.* Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform* 2016; 4 (3): e28.
- Hyland SL, Faltys M, Hüser M, *et al.* Machine learning for early prediction of circulatory failure in the intensive care unit. *arXiv preprint arXiv:1904.07990*. 2019.
- Tomasev N, Glorot X, Rae JW, *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019; 572 (7767): 116–9.
- Delahanty RJ, Alvarez J, Flynn LM, Sherwin RL, Jones SS. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Ann Emerg Med* 2019; 73 (4): 334–44.
- Singer M, Deutschman CS, Seymour CW, *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016; 315 (8): 801–10.
- Askim A, Moser F, Gustad LT, *et al.* Poor performance of quick-SOFA (qSOFA) score in predicting severe sepsis and mortality—a prospective study of patients admitted with infection to the emergency department. *Scand J Trauma Resusc Emerg Med* 2017; 25 (1): 56.
- Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. *Comput Biol Med* 2017; 89: 248–55.

39. Calvert JS, Price DA, Chettipally UK, *et al.* A computational approach to early sepsis detection. *Comput Biol Med* 2016; 74: 69–73.
40. Johnson AE, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 160035.
41. Giannini HM, Chivers C, Draugelis M, *et al.* Development and implementation of a machine-learning algorithm for early identification of sepsis in a multi-hospital academic healthcare system. D15. Critical care: do we have a crystal ball? Predicting clinical deterioration and outcome in critically ill patients: American Thoracic Society, 2017: A7015.
42. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One* 2017; 12 (4): e0174708.
43. Umscheid CA, Chivers C, Bleich J, Draugelis M. In response to “Development, implementation and impact of an automated early warning and response system for sepsis”. *J Hosp Med* 2015; 10 (5): 341.
44. Kamaleswaran R, Akbilgic O, Hallman MA, West AN, Davis RL, Shah SH. Applying artificial intelligence to identify physiometers predicting severe sepsis in the PICU. *Pediatr Crit Care Med* 2018; 19 (10): e495–503.
45. Lin C, Zhang Y, Ivy J, *et al.* Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI). New York, NY: IEEE; 2018: 219–28.
46. Alsolamy S, Al Salamah M, Al Thagafi M, *et al.* Diagnostic accuracy of a screening electronic alert tool for severe sepsis and septic shock in the emergency department. *BMC Med Inform Decis Mak* 2014; 14 (1): 105.
47. Masino AJ, Harris MC, Forsyth D, *et al.* Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PLoS One* 2019; 14 (2): e0212665.
48. Rothman M, Levy M, Dellinger RP, *et al.* Sepsis as 2 problems: identifying sepsis at admission and predicting onset in the hospital using an electronic medical record-based acuity score. *J Crit Care* 2017; 38: 237–44.
49. Steyerberg EW, Moons KG, van der Windt DA, *et al.*; for the PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; 10 (2): e1001381.