HHS Public Access

Author manuscript

Crit Care Med. Author manuscript; available in PMC 2022 August 01.

Published in final edited form as:

Crit Care Med. 2021 August 01; 49(8): 1312–1321. doi:10.1097/CCM.0000000000004966.

A Simulated Prospective Evaluation of a Deep Learning Model for Real-Time Prediction of Clinical Deterioration Among Ward Patients

Parth K. Shah, BA[1], **Jennifer C. Ginestra, MD**[2], **Lyle H. Ungar, PhD**[3], **Paul Junker, MS**[4], **Jeff I. Rohrbach, MSN**[4], **Neil O. Fishman, MD**[5], **Gary E. Weissman, MD, MHSP**[2,5]

[1]Perelman School of Medicine, University of Pennsylvania

[2]Palliative and Advanced Illness Research Center, Perelman School of Medicine, University of Pennsylvania

[3]Computer and Information Science Department, University of Pennsylvania

[4]Clinical Effectiveness and Quality Improvement, Hospital of the University of Pennsylvania

[5]Department of Medicine, Hospital of the University of Pennsylvania

Abstract

Objective: The National Early Warning Score (NEWS), Modified Early Warning Score (MEWS), and quick Sepsis-related Organ Failure Assessment (qSOFA) can predict clinical deterioration. These scores exhibit only moderate performance and are often evaluated using aggregated measures over time. A simulated prospective validation strategy that assesses multiple predictions per patient-day would provide the best pragmatic evaluation. We developed a deep recurrent neural network (DRNN) deterioration model and conducted a simulated prospective evaluation.

Design: Retrospective cohort study.

Setting: Four hospitals in Pennsylvania.

Patients: Inpatient adults discharged between July 1, 2017 and June 30, 2019.

Interventions: None.

Measurements and Main Results: We trained a DRNN and logistic regression model (logit) using data from electronic health records to predict hourly the 24-hour composite outcome of transfer to intensive care unit (ICU) or death. We analyzed 146,446 hospitalizations with 16.75 million patient-hours. The hourly event rate was 1.6% (12,842 transfers or deaths, corresponding to 260,295 patient-hours within the predictive horizon). On a hold-out dataset, the DRNN achieved an area under the precision-recall curve (AUPRC) of 0.042 (95% CI 0.04 to 0.043), comparable to logit (0.043, 95% CI 0.041 to 0.045), and outperformed NEWS (0.034, 95% CI 0.032 to 0.035), MEWS (0.028, 95% CI 0.027 to 0.03), and qSOFA (0.021, 95% CI 0.021 to 0.022). For a fixed sensitivity of 50%, the DRNN achieved a positive predictive value (PPV) of 3.4% (95% CI 3.4 to 3.5) and outperformed logit (3.1%, 95% CI 3.1 to 3.2), NEWS (2.0%, 95% CI 2.0 to 2.0), MEWS (1.5%, 95% CI 1.5 to 1.5), and qSOFA (1.5%, 95% CI 1.5 to 1.5).

Conclusions: Commonly used early warning scores for clinical decompensation, along with a logit and a DRNN model, show very poor performance characteristics when assessed using a simulated prospective validation. None of these models may be suitable for real-time deployment.

Keywords

Early Warning Score; Deep Learning; Machine Learning; Clinical Deterioration; Electronic Health Records

Introduction

Many early warning scores (EWSs) have been developed for use in hospital wards to guide evaluation for clinical deterioration. Such scoring systems could be useful for identifying patients who may benefit from more intensive care, monitoring progression of disease, and triaging resources for rapid response teams (1, 2). Commonly used EWSs such as the National Early Warning Score (NEWS), Modified Early Warning Score (MEWS), and quick Sepsis-related Organ Failure Assessment (qSOFA) are simple to compute. However, these scores have not demonstrated reliable improvements in processes of care or clinical outcomes when deployed in practice (3–6). Lack of reported clinical benefit may be due, in part, to only moderate predictive performance of existing models. It is also possible to overinflate model performance with the use of evaluation strategies that utilize information that would not be available to clinicians in real-time, for example by retrospectively aggregating scores over a hospitalization, using only the worst score in a certain time period, computing scores at a predetermined time before clinical deterioration is observed, or by ignoring realistic use cases of EWS for ongoing serial monitoring (4, 7, 8). On the other hand, the few reports of EWS performance that do account for the low event rate have yielded positive predictive values (PPVs) less than 5% in predicting sepsis and cardiac arrest (9, 10).

There are at least two potential strategies to overcome these limitations in existing EWSs. First, conduct simulated prospective validation of a model to account for the pragmatic scenario in which a model is used to produce multiple predictions each patient day, e.g. every hour. Second, use a modeling strategy to account for temporal trends in clinical parameters. Most EWSs utilize data from a single point in time or manually determine temporal aggregations of vital signs and laboratory data (1, 11, 12). This approach requires *a priori* specification of the most useful aggregations, which are unknown; does not capture the different sampling rates of each variable, which are themselves often informative (13); and does not capture complex interactions between trends in variables. Modern electronic health records (EHRs) present an opportunity to utilize greater amounts of data and develop EWSs that are not limited to simple scoring systems. Several previously published EWSs have utilized machine learning techniques in the setting of the emergency room (14–16) where longitudinal data are less available and in the intensive care unit (9, 17–19) where data are abundant, but less attention has been given to patients on hospital wards (20, 21). Some recent EWSs have employed deep recurrent neural network (DRNN) models that are specifically designed for temporal data in other settings (9, 22, 23).

We sought to overcome existing barriers to developing high-performing EWSs that capture time-varying patterns in data and are evaluated under realistic clinical conditions. Specifically, we developed a DRNN model that could learn relevant trends from EHR data to predict clinical deterioration among hospital ward patients and performed a simulated prospective validation of our model to compare against commonly used EWSs.

Materials and Methods

Data Source

Data was acquired from the Clarity database that contains all clinical data stored from the EHR for our institution (24).

Study Population

All patients at least 18 years old admitted for inpatient stay lasting at least 24 hours at one of four PennMedicine hospitals were included. Data captured while a patient was in a non-ward setting (e.g. ICU, emergency room, or operating room) were excluded, and patients discharged from an ICU were treated like new admissions once they arrived on the floor. After removal of non-ward data, any encounters less than or equal to twelve hours were excluded to ensure times series had adequate length for analysis. Data were divided into a model training dataset used for training DRNN model weights, representing 80% of encounters (discharge dates from July 1, 2017 to February 8, 2019), a model validation dataset used for hyperparameter tuning and model selection, representing 10% of encounters (discharge dates from February 8, 2019 to April 19, 2019), and a testing dataset used for simulated prospective validation, representing the final 10% of encounters (discharge dates from April 19, 2019 to June 30, 2019).

Feature Generation

We extracted up to hour-level vital signs, laboratory studies, antibiotics usage, demographics, and other health process variables such as time and hospital unit for the 6 hours preceding each prediction. A total of 61 features were included in the DRNN model. Procedures used for processing outliers, imputing missing data, and feature engineering along with complete list of input variables are provided in Supplementary Digital Content 1.

The primary outcome of clinical deterioration was defined as a composite of transfer to the ICU or in-hospital mortality within 24 hours, similar to other EWSs (4, 11, 15, 25).

Model Development

The DRNN was trained using a two-step transfer learning approach as described in Supplementary Digital Content 2 (26–28). Model parameters for both the autoencoder and the final model were trained using the training dataset while layer dimensions and regularization parameters for both were chosen based on model performance in tuning grids using the model validation dataset (Supplementary Digital Content 2, Figures S5–S6).

A binomial logistic regression model (logit) without penalization was also trained to predict the composite outcome using the training dataset with the same inputs as the DRNN model,

i.e. six-hour window of 61 features flattened to a single vector, to serve as an additional comparator.

Simulated Prospective Validation

Predictions from the DRNN and logit as well as NEWS, MEWS, and qSOFA scores were computed for each patient-hour in the testing dataset (5, 29, 30). Although qSOFA is not a general early warning score, it is often deployed and evaluated as such (31). Therefore, we chose to include it here for comparison. Discrimination of the composite outcome for each model was assessed by the area under the precision-recall curve (AUPRC). Calibration for the DRNN and logit models were assessed by visualizing a calibration plot. Because the other EWSs do not produce predicted probabilities, a proxy for calibration was assessed by plotting event rate across the range of scores. Additional performance metrics were computed, including area under the receiver operating characteristic (AUROC) curve, PPV for a fixed sensitivity of 50%, and the number needed to evaluate (NNE), a policy-level measure of the value of a predictive model (32, 33).

Subgroup Analysis

A pre-planned subgroup analysis was conducted by computing AUPRC for the DRNN, logit, NEWS, MEWS, and qSOFA across several demographic and clinical groups. These subgroups were chosen 1) to understand how the model might perform in populations with a case-mix different from that reported here; and 2) to better understand how such a model may differentially affect clinical care in protected groups given the risk of reinforcing biases when deploying clinical prediction models.

Model Usability

To assess DRNN usability, a patient encounter was selected for narrative analysis from the subset of DRNN true positives in the testing dataset in which the DRNN accurately predicted deterioration more reliably than other EWSs. Alert burden was simulated assuming model silencing for consecutive alerts. To get insight into the contribution of each feature to the model's predictions, a measure of global feature importance was computed using a permutation-based algorithm (34).

95% confidence intervals for performance metrics were produced via bootstrap with 1,000 replicates. 2-sided $P < 0.05$ was considered statistically significant. All data analyses were performed using R statistical software (35). We adhered to the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines for best practices in development and reporting of clinical prediction models (36). The final DRNN model object with instructions for use are available online for download at https://github.com/weissman-lab/deterioration_drnn. This study was deemed exempt by the University of Pennsylvania's Institutional Review Board (Protocol #832918).

Results

We analyzed 146,446 hospitalizations from 103,930 unique patients and 16.75 million patient-hours. In total, there were 11,507 transfers to an ICU and 1,335 (0.9%) in-hospital

deaths (Table 1). Analyzed on a patient-hour basis with a 24-hour predictive time horizon, 260,295 of the 16.75 million patient-hours fell within the predictive horizon, resulting in an event rate of 1.6% for the composite outcome.

On the hold-out testing dataset, the DRNN achieved an AUPRC of 0.042 (95% CI 0.04 to 0.043), comparable to logit (0.043, 95% CI 0.041 to 0.045), and outperforming NEWS (0.034, 95% CI 0.032 to 0.035), MEWS (0.028, 95% CI 0.027 to 0.03), and qSOFA (0.021, 95% CI 0.021 to 0.022) though all were below 0.05 (Figure 1A). The calibration plot for the DRNN closely tracks the 45-degree line at low predicted probabilities and overestimates predictions at higher predicted probabilities (Figure 1B).

For a fixed sensitivity of 50%, the DRNN achieved a PPV of 3.4% (95% CI 3.4 to 3.5) corresponding to a NNE of 29, compared to logit (PPV 3.1%, 95% CI 3.1 to 3.2, NNE 32), NEWS (PPV 2.0%, 95% CI 2.0 to 2.0, NNE 50), MEWS (PPV 1.5%, 95% CI 1.5 to 1.5, NNE 67), and qSOFA (PPV 1.5%, 95% CI 1.5 to 1.5, NNE 67). The DRNN achieved the highest hourly AUROC of 0.72 (95% CI 0.72 to 0.73), compared to logit (0.71, 95% CI 0.71 to 0.71), NEWS (0.60, 95% CI 0.60 to 0.60), MEWS (0.57, 95% CI 0.57 to 0.58), and qSOFA (0.55, 95% CI 0.55 to 0.55) (Table 2).

Subgroup Analysis

The DRNN achieved a higher AUPRC than NEWS, MEWS, and qSOFA for all subgroups except among patients suspected to have sepsis at some point during their hospitalization (Figure 2). The DRNN also shows improved discrimination compared to the total population among Black patients and patients who received blood cultures at some point during their hospitalization and worse discrimination among older and female patients.

Model Usability

To demonstrate the usability of the DRNN as a real-time alert, predictions using the DRNN, logit, and other EWSs were computed for a single patient encounter (Supplemental Digital Content 4, Patient Vignette) according to the information known at each point in time (Figure 3). The large lead time predicted by the DRNN could have provided an opportunity to readdress goals of care with more prognostic information and/or to escalate ward-level care prior to ICU transfer and ultimate rapid clinical deterioration.

While the DRNN produced fewer new alerts per patient encounter than the other models at a sensitivity of 50%, it produced at least one alert for 58% of patient encounters, and the alert burden was distributed along the entire length of the encounter (Supplemental Digital Content 4, Figures S11–S12). Permutation-based feature importance for the DRNN demonstrated that three of the ten most important variables were derived from the respiratory rate (Supplemental Digital Content 4, Figure S13).

Discussion

We found that a simulated prospective validation strategy to evaluate EWSs for clinical deterioration among ward patients consistently revealed poor performance across all model types. These findings also confirm prior systematic observations that currently published

performance measures of such scores likely overestimate their utility in practice (8). These findings may explain, in part, why prospective evaluations of EWS have failed to reliably demonstrate improvements in clinical outcomes. The DRNN model learned patterns of time-varying clinical features and provides performance improvement over models with static features, suggesting potential opportunities to leverage similar model architectures for EWS applications; however, it too achieves poor performance and is unlikely to support real-world deployment. Thus, our findings have several implications for the development of predictive EWS systems.

First, all models including the DRNN show performance that may be insufficient to support real-time deployment in typical clinical workflows due to high false alert rate. High false alert rates lead to large amounts of additional work-up and wasted hospital resources. Alert burden for the DRNN is high and spread along the entire encounter for both outcome positive and outcome negative patients, which would make real-time interpretation difficult and contribute to alert fatigue. In evaluating a discrete-time logistic model for predicting ICU transfer or hospital mortality, Kipnis et al. report a PPV around 15% to achieve 50% sensitivity, but they report PPV based on episode-level predictions rather than hourly predictions (11). These findings underscore the importance of refining models prior to deployment so as not to squander scarce resources on ineffective predictive interventions (9, 10).

Second, models that capture time-varying features are likely to outperform those that do not, supporting work by Churpek et al. (12). In the simulated prospective validation, our DRNN significantly outperforms other EWSs in discrimination of the composite outcome. Overall, the DRNN shows good calibration of predictions to event rate with overestimation at higher predicted probabilities. The DRNN also achieves higher AUROC, better PPV and NNE at fixed sensitivity, and lower alert burden compared to other EWSs. While recurrent neural networks have face validity for modeling temporal clinical data, the DRNN only outperformed logistic regression at higher sensitivities, and the added complexity of the DRNN should be weighed against availability of local resources. These findings are consistent with other studies showing that complex machine learning models provide no or only marginal improvements in performance over traditional regression methods (38, 39). Of note, all AUPRC were less than 0.05.

Third, pre-specified subgroup analyses are necessary to understanding nuances of predictive model performance to ensure efficacy and equity in deployment. The DRNN exhibited slightly worse discrimination among older and female patients and slightly better discrimination among Black patients and those receiving blood cultures. Further work would be needed to better characterize mechanisms for these differences and to consider additional data gathering or data generating mechanisms to alleviate inequitable performance prior to deployment (40). That NEWS, MEWS, and qSOFA showed better discrimination among patients with suspected sepsis compared to DRNN may be explained by the fact that these scores are more closely aligned with the Sepsis-2 definition that were used to identify sepsis in our cohort.

Finally, simulated prospective validation is essential prior to deployment of an EWS. Although this approach seemingly worsens relevant clinical and policy performance metrics like PPV and NNE, it provides the best pragmatic assessment of a prediction model's real-world performance for hospitalized patients (33). Researchers developing EWSs should evaluate their models using a simulated validation strategy that is as close in implementation to a prospective clinical trial as possible to provide realistic assessments of model performance prior to deployment.

This study should be interpreted in light of several limitations. First, while the data used to train and test the models came from multiple institutions, including academic and community hospitals, all data came from one region. Because we wanted testing data to be prospective in time to the training data, testing data came from just a few months, which could limit generalizability. Second, the DRNN requires compilation of data from multiple sources in real-time for implementation, which may not be feasible in health systems without robust EHR systems and resource-limited settings. Third, this study utilized NEWS rather than the updated NEWS2, which has alternative scoring for patients with documented hypercapnic respiratory failure. Because judgement of hypercapnic respiratory failure is challenging to make for an automated early warning system with limited access to patient's historical information, NEWS was selected for evaluation, which may have lower specificity in that population (41). Fourth, the DRNN model architecture was selected *a priori*, given both successful application of such models in other uses and face validity for the modeling task at hand. However, it is possible that other model architectures that were not tested may provide a better fit and perhaps gains in performance. Finally, an individual alert from the DRNN model does not provide information about which feature(s) drove that prediction but rather serves as a general alert for closer examination.

Ultimately, clinicians and hospital policy makers must decide whether the PPV, NNE, and calibration for these prediction models are acceptable prior to full implementation into clinical practice. They must weigh the benefits of earlier detection of clinical deterioration against the risks and costs of false positives and education needed to change clinical practice.

Conclusions

Commonly used early warning scores for clinical decompensation show very poor performance characteristics when assessed for real-time use in wards using a pragmatic, prospective simulated validation. A deep neural network model that accounts for temporal trends in clinical data may not support real-time deployment despite outperforming traditional models in this scenario.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Conflicts of Interest and Source of Funding: Parth Shah was supported by the Medical Student Health Services and Policy Research Summer Research Fellowship through the Master of Science in Health Policy Research Program at the Perelman School of Medicine, University of Pennsylvania. Gary Weissman was supported in part by NIH K23HL141639 and a grant from the America Thoracic Society Research Foundation.

Copyright form disclosure: Dr. Shah's institution received funding from America Thoracic Society Research Foundation and the National Institutes of Health (NIH), and he received funding from Medical Student Health Services and Policy Research Summer Research Fellowship through the Master of Science in Health Policy Research Program at the Perelman School of Medicine, University of Pennsylvania. Drs. Shah, Ungar, and Weissman received support for article research from the NIH. Dr. Weissman also received support for article research from the American Thoracic Society Research Foundation. The remaining authors have disclosed that they do not have any potential conflicts of interest.

References

1. Smith MEB, Chiovaro JC, O'Neil M, et al.: Early Warning System Scores for Clinical Deterioration in Hospitalized Patients: A Systematic Review. Ann Am Thorac Soc 2014; 11:1454–1465 [PubMed: 25296111]
2. Duncan K, McMullan C, Mills B: Early warning systems: The next level of rapid response. Nursing (Lond) 2012; 42:38–44
3. Song J-U, Sin CK, Park HK, et al.: Performance of the quick Sequential (sepsis-related) Organ Failure Assessment score as a prognostic tool in infected patients outside the intensive care unit: a systematic review and meta-analysis. Crit Care Lond Engl 2018; 22:28
4. Churpek MM, Snyder A, Han X, et al.: Quick Sepsis-related Organ Failure Assessment, Systemic Inflammatory Response Syndrome, and Early Warning Scores for Detecting Clinical Deterioration in Infected Patients outside the Intensive Care Unit. Am J Respir Crit Care Med 2017; 195:906–911 [PubMed: 27649072]
5. Smith GB, Prytherch DR, Meredith P, et al.: The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation 2013; 84:465–470 [PubMed: 23295778]
6. Alam N, Hobbelink EL, van Tienhoven AJ, et al.: The impact of the use of the Early Warning Score (EWS) on patient outcomes: a systematic review. Resuscitation 2014; 85:587–594 [PubMed: 24467882]
7. Yu S, Leung S, Heo M, et al.: Comparison of risk prediction scoring systems for ward patients: a retrospective nested case-control study. Crit Care 2014; 18:R132 [PubMed: 24970344]
8. Gerry S, Bonnici T, Birks J, et al.: Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. BMJ 2020; 369
9. Kwon J, Lee Y, Lee Y, et al.: An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest [Internet]. J Am Heart Assoc 2018; 7[cited 2020 Oct 30] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6064911/
10. Topiwala R, Patel K, Twigg J, et al.: Retrospective Observational Study of the Clinical Performance Characteristics of a Machine Learning Approach to Early Sepsis Identification. Crit Care Explor 2019; 1:e0046 [PubMed: 32166288]
11. Kipnis P, Turk BJ, Wulf DA, et al.: Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. J Biomed Inform 2016; 64:10–19 [PubMed: 27658885]
12. Churpek MM, Adhikari R, Edelson DP: The value of vital sign trends for detecting clinical deterioration on the wards. Resuscitation 2016; 102:1–5 [PubMed: 26898412]
13. Agniel D, Kohane IS, Weber GM: Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018; 361
14. Goto T, Camargo CA, Faridi MK, et al.: Machine Learning–Based Prediction of Clinical Outcomes for Children During Emergency Department Triage. JAMA Netw Open 2019; 2:e186937–e186937 [PubMed: 30646206]

15. Raita Y, Goto T, Faridi MK, et al.: Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care Lond Engl 2019; 23:64
16. 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 Off J Soc Acad Emerg Med 2016; 23:269–278
17. Shickel B, Loftus TJ, Adhikari L, et al.: DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Sci Rep 2019; 9:1–12 [PubMed: 30626917]
18. Delahanty RJ, Kaufman D, Jones SS: Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients. Crit Care Med 2018; 46:e481–e488 [PubMed: 29419557]
19. Shillan D, Sterne JAC, Champneys A, et al.: Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care Lond Engl 2019; 23:284
20. Churpek MM, Yuen TC, Winslow C, et al.: Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Crit Care Med 2016; 44:368–374 [PubMed: 26771782]
21. Rojas JC, Carey KA, Edelson DP, et al.: Predicting Intensive Care Unit Readmission with Machine Learning Using Electronic Health Record Data. Ann Am Thorac Soc 2018; 15:846–853 [PubMed: 29787309]
22. Futoma J, Hariharan S, Sendak M, et al.: An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. Proc Mach Learn Healthc 2017; 68
23. Kaji DA, Zech JR, Kim JS, et al.: An attention based deep learning model of clinical events in the intensive care unit. PLOS ONE 2019; 14:e0211057 [PubMed: 30759094]
24. Epic Systems. Madison, WI: Epic Systems Corporation;
25. Shamout FE, Zhu T, Sharma P, et al.: Deep Interpretable Early Warning System for the Detection of Clinical Deterioration. IEEE J Biomed Health Inform 2019;
26. Sutskever I, Vinyals O, Le QV: Sequence to Sequence Learning with Neural Networks. In: Proc NIPS. Montreal, CA: 2014;
27. Beaulieu-Jones BK, Greene CS, Pooled Resource Open-Access ALS Clinical Trials Consortium: Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform 2016; 64:168–178 [PubMed: 27744022]
28. Sagheer A, Kotb M: Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems. Sci Rep 2019; 9:1–16 [PubMed: 30626917]
29. Singer M, Deutschman CS, Seymour CW, et al.: The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016; 315:801–810 [PubMed: 26903338]
30. Subbe CP, Kruger M, Rutherford P, et al.: Validation of a modified Early Warning Score in medical admissions. QJM Mon J Assoc Physicians 2001; 94:521–526
31. Liu VX, Lu Y, Carey KA, et al.: Comparison of Early Warning Scoring Systems for Hospitalized Patients With and Without Infection at Risk for In-Hospital Mortality and Transfer to the Intensive Care Unit. JAMA Netw Open 2020; 3:e205191 [PubMed: 32427324]
32. Liu VX, Bates DW, Wiens J, et al.: The number needed to benefit: estimating the value of predictive analytics in healthcare. J Am Med Inform Assoc JAMIA 2019; 26:1655–1659 [PubMed: 31192367]
33. Romero-Brufau S, Huddleston JM, Escobar GJ, et al.: Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. Crit Care 2015; 19
34. Fisher A, Rudin C, Dominici F: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. ArXiv180101489 Stat 2019;
35. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: 2018. Available from: https://www.r-project.org
36. Collins GS, Reitsma JB, Altman DG, et al.: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015; 162:55–63 [PubMed: 25560714]

37. National Early Warning Score (NEWS) Standardising the assessment of acute-illness severity in the NHS. Report of a working party. London: RCP: Royal College of Physicians; 2017.

38. Christodoulou E, Ma J, Collins GS, et al.: A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology 2019; 110:12–22 [PubMed: 30763612]

39. Desai RJ, Wang SV, Vaduganathan M, et al.: Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Netw Open 2020; 3:e1918962 [PubMed: 31922560]

40. Parikh RB, Teeple S, Navathe AS: Addressing Bias in Artificial Intelligence in Health Care. JAMA 2019; 322:2377–2378 [PubMed: 31755905]

41. Pedersen NE, Rasmussen LS, Petersen JA, et al.: Modifications of the National Early Warning Score for patients with chronic respiratory disease. Acta Anaesthesiol Scand 2018; 62:242–252 [PubMed: 29072311]

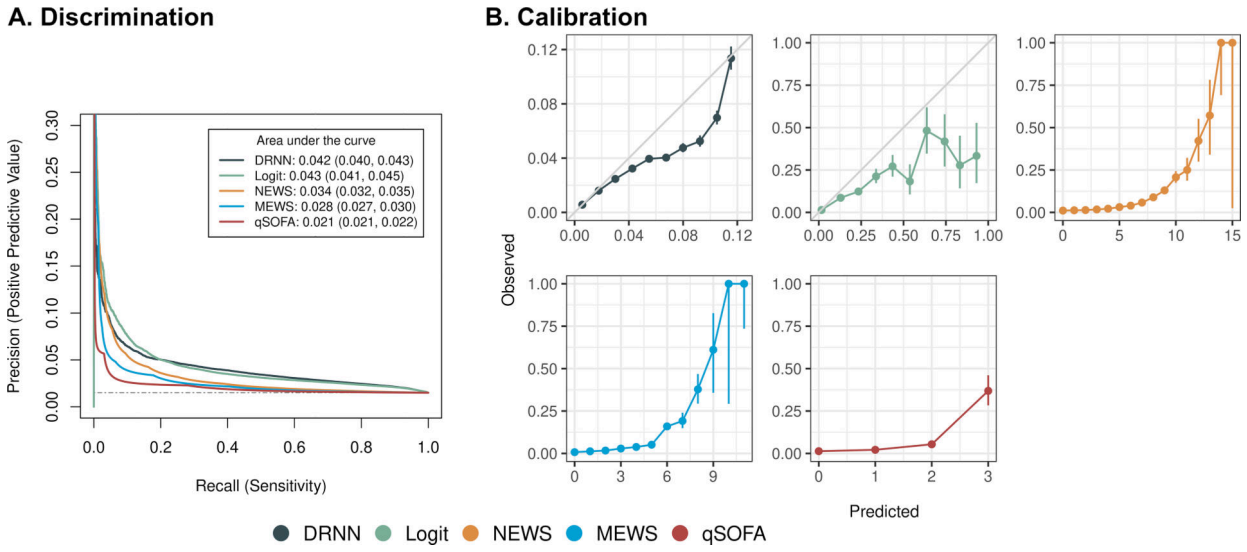

Figure 1:

(**A**) Discrimination of deep recurrent neural network (DRNN) compared to commonly used early warning scores for predicting the 24-hour composite outcome with 95% confidence intervals. Scores computed per hourly simulated prospective evaluation. Dashed gray line denotes performance of random classifier. (**B**) Model calibration is assessed by sorting model predictions and comparing prediction score to observed fraction of composite outcome with 95% confidence intervals. DRNN was divided into ten equally spaced bins while other EWSs were divided into one bin for each point. Solid gray line for DRNN reflects ideal calibration. Abbreviations: Logit, logistic regression model; NEWS, National Early Warning Score; MEWS, Modified Early Warning Score; qSOFA, quick Sepsis-related Organ Failure Assessment.

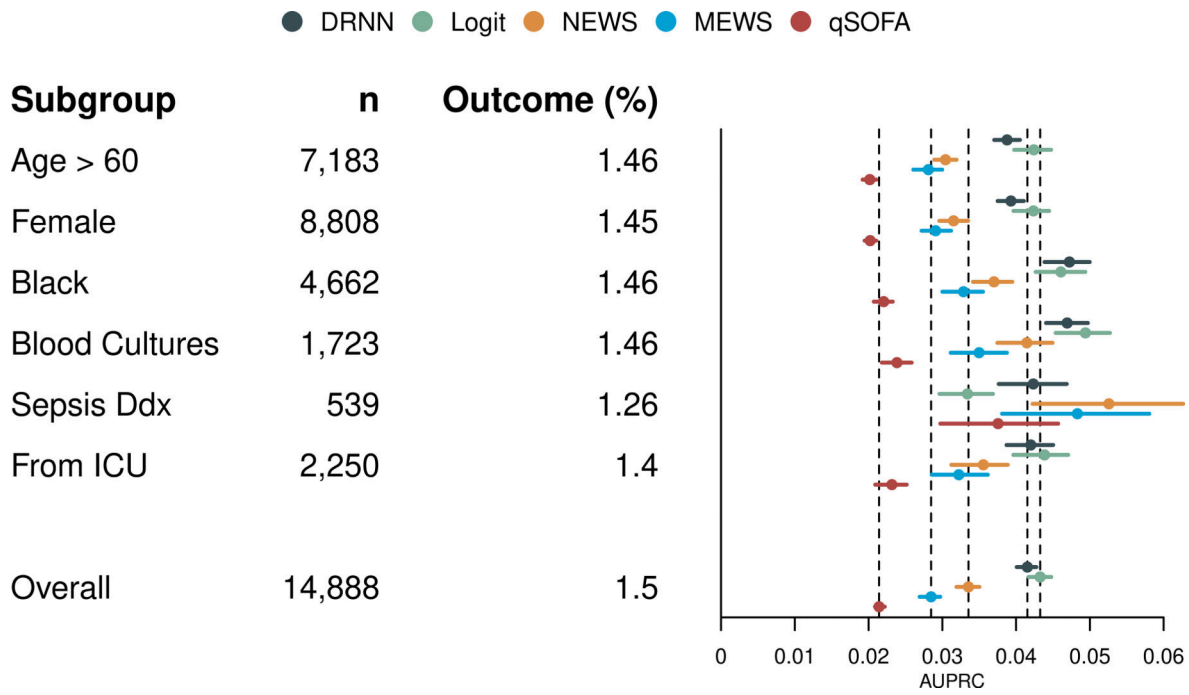

Figure 2:

For each subgroup, area under the precision-recall curve (AUPRC) is plotted with 95% confidence intervals for the deep recurrent neural network (DRNN), logistic regression model (Logit), National Early Warning Score (NEWS), Modified Early Warning Score (MEWS), and quick Sepsis-related Organ Failure Assessment (qSOFA). Abbreviations: n, number of hospitalizations; Outcome, prevalence of composite outcome; Blood Cultures, patients with blood cultures drawn at some point during their hospitalization; Sepsis Ddx, patients thought to have sepsis at some point during their hospitalization; From ICU, patients transferred to floor from an intensive care unit.

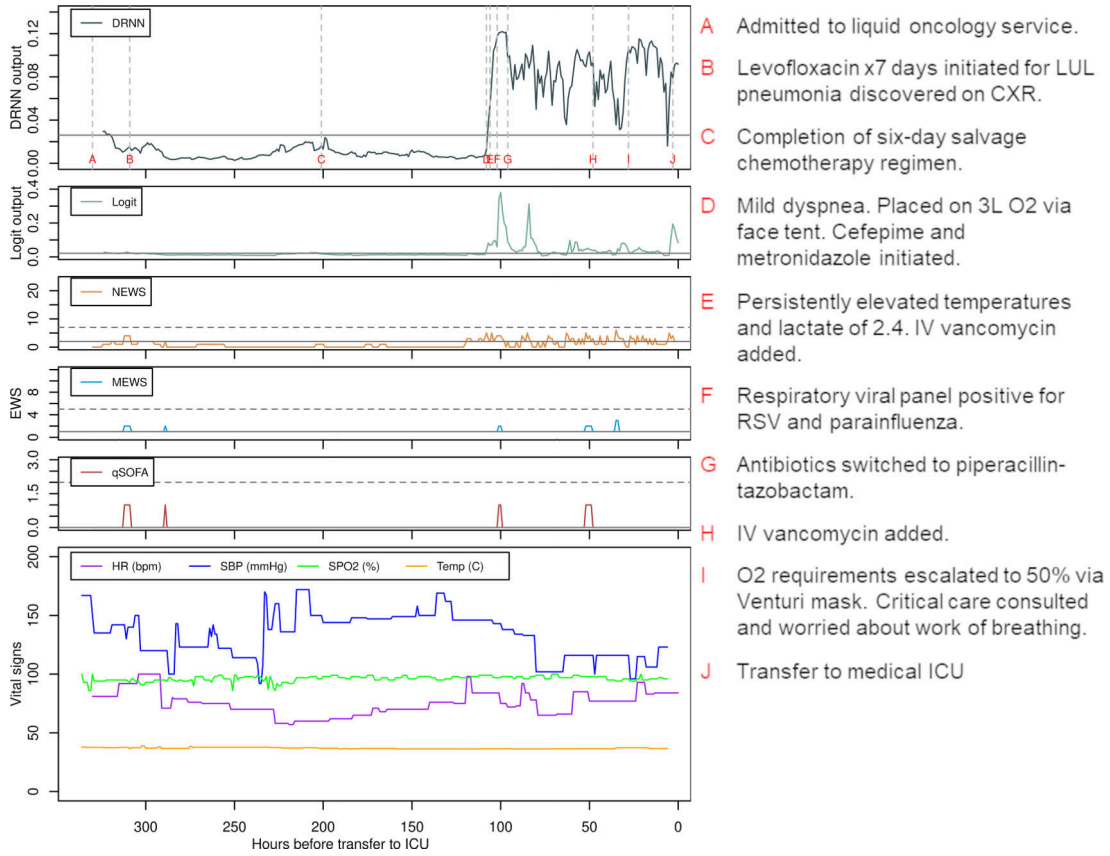

Figure 3:

Deep recurrent neural network (DRNN) and logistic regression (Logit) model predictions and other early warning scores (EWS) for a single patient from the testing dataset, correlated with vital signs and important events represented by letters A through G. Solid gray lines mark thresholds that yield a sensitivity of 50% while horizontal dashed gray line mark commonly used thresholds for early warning scores for outreach to critical-level care. Abbreviations: NEWS, National Early Warning Score; MEWS, Modified Early Warning Score; qSOFA, quick Sepsis-related Organ Failure Assessment; HR, heart rate; SBP, systolic blood pressure; SPO2, oxygen saturation; ICU, intensive care unit; LUL, left upper lobe; CXR, chest x-ray; RSV, respiratory syncytial virus; IV, intravenous.

Table 1:

Data characteristics

Variable	Value	Observations per encounter, median (IQR)	Encounters without any values, n (%)
Hospitalizations, n	146,446		
Patient-hours, n	16,750,411		
Female gender, n (%)	86,621 (59.1)		0 (0)
Age (years), median (IQR)	59 (37–71)		0 (0)
Race, n (%)			3,201 (2.1)
White	88,372 (56.8)		
Black	47,368 (30.5)		
Asian	3,986 (2.6)		
Other	3,519 (2.3)		
Hispanic, n (%)	5,949 (4.1)		3,201 (2.1)
LOS (days), median (IQR)	3.6 (2.4–6.2)		0 (0)
Hospital Mortality, n (%)	1,335 (0.9)		0 (0)
Transfers to an ICU, n	11,507		
Patient-hour level variables			
Respiratory Support, n (%)	880,222 (5.3)		0 (0)
Heart Rate (beats/min), median (IQR)	81 (72–92)	11 (6–24)	574 (0.4)
Lactate (mmol/L), median (IQR)	1.2 (1.2–1.2)	0 (0–1)	113,251 (72.5)
Platelet Count (10^9/L), median (IQR)	207 (153–271)	80 (52–146)	3,413 (2.2)
Total Bilirubin (mg/dL), median (IQR)	0.5 (0.4–0.6)	43 (0–116)	68,502 (43.9)
Glasgow Coma Scale, median (IQR)	15 (15–15)	46 (0–119)	65,410 (41.9)
Creatinine (mg/dL), median (IQR)	0.87 (0.67–1.18)	77 (39–145)	19,903 (12.7)
SOFA Score, median (IQR)	3 (1–5)	88 (57–151)	0 (0)
Blood Cultures Drawn, n (%)	36,025 (0.2)		0 (0)
Antibiotics Usage, n (%)	544,964 (3.3)		0 (0)

Patient-hour level values are reported after removal of outliers and imputation of missing data. For outlier ranges and imputation strategies utilized, see supplement. Abbreviations: n, number of observations; IQR, interquartile range; ICU, intensive care unit.

Table 2:

Performance metrics for DRNN and other early earning scores

Performance metric	DRNN	Logistic Model	NEWS	MEWS	qSOFA
Area under precision-recall curve	0.042 (0.040, 0.043)	0.043 (0.041, 0.045)	0.034 (0.032, 0.035)	0.028 (0.027, 0.030)	0.021 (0.021, 0.022)
Positive predictive value	0.034 (0.034, 0.035)	0.031 (0.031, 0.032)	0.020 (0.020, 0.020)	0.015 (0.015, 0.015)	0.015 (0.015, 0.015)
Number needed to evaluate	29 (29, 30)	32 (32, 33)	50 (49, 51)	67 (66, 67)	67 (66, 67)
Area under receiver operating characteristics curve	0.723 (0.720, 0.726)	0.711 (0.708, 0.714)	0.600 (0.596, 0.604)	0.574 (0.570, 0.577)	0.551 (0.548, 0.554)

Performance metrics reported with 95% confidence intervals. Positive predictive value and number needed to evaluate were calculated for a fixed sensitivity of 50%. Abbreviations: DRNN, deep recurrent neural network; NEWS, National Early Warning Score; MEWS, Modified Early Warning Score; qSOFA, quick Sepsis-related Organ Failure Assessment.