MDMP&P
Policy & Practice

MDM Policy & Practice
1–10
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2381468319899663
journals.sagepub.com/home/mdm
SAGE

Development, Implementation, and Evaluation of an In-Hospital Optimized Early Warning Score for Patient Deterioration

Cara O'Brien, Benjamin A. Goldstein, Yueqi Shen, Matthew Phelan, Curtis Lambert, Armando D. Bedoya, and Rebecca C. Steorts

Background. Identification of patients at risk of deteriorating during their hospitalization is an important concern. However, many off-shelf scores have poor in-center performance. In this article, we report our experience developing, implementing, and evaluating an in-hospital score for deterioration. **Methods.** We abstracted 3 years of data (2014–2016) and identified patients on medical wards that died or were transferred to the intensive care unit. We developed a time-varying risk model and then implemented the model over a 10-week period to assess prospective predictive performance. We compared performance to our currently used tool, National Early Warning Score. In order to aid clinical decision making, we transformed the quantitative score into a three-level clinical decision support tool. **Results.** The developed risk score had an average area under the curve of 0.814 (95% confidence interval = 0.79–0.83) versus 0.740 (95% confidence interval = 0.72–0.76) for the National Early Warning Score. We found the proposed score was able to respond to acute clinical changes in patients' clinical status. Upon implementing the score, we were able to achieve the desired positive predictive value but needed to retune the thresholds to get the desired sensitivity. **Discussion.** This work illustrates the potential for academic medical centers to build, refine, and implement risk models that are targeted to their patient population and work flow.

Keywords
clinical decision support, electronic health records, predictive models

Date received: October 3, 2019; accepted: December 7, 2019

Early Warning Scores (EWSs) have become an important component of managing inpatient care. They provide a means to assess changes in a patient's clinical status alerting clinicians to the need for intervention. One commonly used EWS is the National Early Warning Score (NEWS), which was designed to detect risk of patient deterioration.[1] Seeking to improve clinical decision support, our institution integrated automated calculation and reporting of the NEWS into its electronic health record (EHR) system. An internal pre–post evaluation showed that implementation of the NEWS had no meaningful impact on patient outcomes.[2] The overall performance of the

Department of Medicine, Duke University, Durham, North Carolina (CO, ADB); Department of Biostatistics & Bioinformatics, Duke University, Durham, North Carolina (BAG, RCS); Center for Predictive Medicine, Duke Clinical Research Institute, Durham, North Carolina (BAG, MP); Department of Statistical Sciences, Duke University, Durham, North Carolina (YS, RCS); Duke Health Technology Solutions, Duke University Health System, Durham, North Carolina (CL, ADB); Department of Computer Sciences, Duke University, Durham, North Carolina (RCS). The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Duke Center for Integrative Health (CO, RCS, BAG).

Corresponding Author
Benjamin A. Goldstein, Department of Biostatistics & Bioinformatics, Duke University, 2424 Erwin Rd, Suite 9023, Durham, NC 27705, USA (ben.goldstein@duke.edu).

NEWS was relatively low with the average area under the curve (AUC) of 0.74 for prediction of unanticipated intensive care unit (ICU) transfer or death within the first 2 days after admission. Others have found that similar off-the-shelf EWSs have similarly mixed performance,[3] and that alerts were generally ignored.[4]

As a clinical decision support tool, there are two key shortcomings for the NEWS. First, EWSs such as the NEWS were not designed to fully utilize the capabilities of modern EHR systems. Instead, NEWS was intended to be easily hand calculated and, therefore, uses only seven vital signs. Such hand calculation has proven to be challenging in clinical environments.[5] However, modern EHR systems are capable of collecting and analyzing patient data on a variety of factors within a real-time environment. For example, each time a blood pressure measurement is taken or a laboratory test is ordered, this information is stored within a running catalog of the EHR system. We can combine such data using more sophisticated machine learning methods to develop robust risk scores. Second, as a general score, NEWS is not optimized for any particular patient population. As EHR data become more readily accessible, the opportunity to develop more locally tailored scores has increased.[6] Moreover, attempts to translate general scores that were developed in "cleaner" environments to local EHRs have been challenging.[7] As such, when considering local implementation, there is room to improve risk models for the local environment. It is possible for each institution to develop its own robust risk score and not rely on simpler off-the-shelf scores. Other have illustrated how they can achieve increased performance by developing more sophisticated risk scores from single-center data.[8–10]

Our goal in this article is to demonstrate how we developed, implemented, and evaluated a risk model for patient deterioration. Since we tailored this score for our local patient population, we do not suggest that it is optimal for any other patient group. Instead, we highlight the gains that one realize by implementing a locally optimized risk score. Moreover, we show how we took a quantitative risk score—that can provide hard to interpret output without clear clinical guidance—and developed a more clinically interpretable display that produces a clear clinical decision support aid.

Materials and Methods

Available Data

Analytic Cohort. We drew data from the Duke University Hospital (DUH) EHR system, an Epic Systems Corporation[11] (EPIC) based health system, installed in late 2013. We extracted data on patient hospital stays from January 1, 2014, to December 30, 2016. We focus on patients admitted to general medical-surgical wards, that is, an environment where patients are not receiving constant monitoring, as is the case in an ICU. We planned to implement the model into the EPIC Acuity Scoring module.[11] This environment allows one to generate risk models using data generated in real time. One designates the clinical feature to be used and assigns a weight (i.e., a beta coefficient) to that value. Therefore, this environment is designed to handle regression-based models as opposed to more complicated machine learning models. However, since the calculation is embedded in the EPIC environment the results can be directly fed back to providers.

Outcome of Interest. Our primary outcome, which we term patient deterioration, is a composite of inpatient mortality and transfer to the ICU. We chose this composite because it captures most forms of adverse outcomes and allows for the clinical team to make assessments regarding the best course of action.[12,13] At our institution, a decompensation requiring transfer to the ICU typically involves calling a Rapid Response Team (RRT). We considered including RRT calls as part of the outcome but found that these data were not reliably captured. We removed events that did not occur on one of the general medical-surgical wards of interest. For example, if a patient died during a surgical procedure, we did not consider this an event of interest. Similarly, if a person went to the surgical ICU immediately postoperatively we did not consider this an event. Since most events occurred early in the admission, for analytical purposes, patients were censored either at the time of discharge or after 7 days of a hospital stay.

Predictor Variables. We extracted variables that we could easily incorporate into a real-time alert system. In particular, we did not include any variables that would depend on previous encounter information. The extracted predictors consisted of demographics, vital signs, comorbidities, medication therapeutic class, and laboratory tests. In total there were 50 predictors (see Supplemental Table 1). We extracted all time varying data with time stamps. We also tracked patient unit location to ensure a patient was in a unit of interest, that is, not in surgery. We describe how we set up the data in the supplemental materials.

Setting up the Data

We took a number of steps to set up the time-varying data for analytic purposes. Our overarching goal was to have the analytic data reflect the real-time streaming data. First, we dealt with implausible vital (0.15% of all vitals) and laboratory values ($<1\%$) by simply removing this small fraction of values. A second consideration was handling of laboratory values (see Supplemental Table 2). Because not all labs were taken on all patients, we created a four-level categorical variable for each lab of "not ordered," "normal," "high," and "low." Categorizing lab values this way has the advantage of naturally handling unmeasured laboratory tests—which we expect to be informative—and allowing for a U-like relationship in the test value. Because there is a lag between when a test is ordered and when the test is resulted, we created an indicator for the when the test was ordered. Because the results of most laboratory tests are clinically relevant for 24 to 72 hours after the test is ordered, we had the results reset to "not ordered" after a set period of time. Finally, we noted that the vitals were updated frequently but at irregular intervals ranging from several minutes to several hours with a median interval of 1.9 hours. To avoid having irregularly observed data, we set up the data in 2-hour blocks. If a patient had more than one measurement in a 2-hour period, we use the most recent measurement. We chose the most recent measurement because the EPIC Acuity Scoring does not have a *memory* of previous values. If a patient did not have a measurement within a block, we simply carried forward the most recent measurement. We used an imputation of a *normal* value (e.g., systolic blood pressure of 120) before a first vital was measured.

Development of Predictive Model

To estimate the predictive model, we split the data into training and testing sets, reserving the last 6 months of data, July 2016 to December 2016, as the testing data. While a variety of machine learning methods could have been used to fit the predictive model, the EPIC Acuity Scoring platform most readily handles (logistic) regression-based models where one simply inputs a weight (i.e., beta coefficient) for each variable. Therefore, using the training data, we fit a regularized logistic regression model, using least absolute shrinkage and selection operator (LASSO) to estimate more stable coefficients. We included time after admission as a predictor, categorizing time into piecewise constants of <1 day, 1 to 2 days, or ≥ 2 days. The NEWS table suggests there are U-shape relationships between four of the vitals (heart rate, blood pressure, temperature, and respiration) and risk, with both low and high values conferring added risk. To account for these potential U-shape relationships, we include quadratic terms for the vitals. We also considered interaction terms between vitals and time after admission. We used 10-fold cross-validation on the training data, assessing fit via the AUC. Results suggested that the optimal fit incorporated quadratic terms but no interactions between time and vitals.

Retrospective Evaluation

We evaluated the model based on its 12-hour performance. The clinical team determined a 12-hour horizon to be a clinically relevant time point that was both near-term enough to be considered a salient risk and over enough of a horizon to be actionable. Using the test data, we calculated the AUC over the next 12 hours, varying this over time. We compared this to the performance of the NEWS.[1] We used the bootstrap to construct 95% confidence intervals (CIs) and compare model performance[14] using the pROC[15] package in R.

We next considered the performance of an alert. We had a series of conversations with nursing staff asking how they would like to interact with an alert. Based on these discussions, we choose a three-tiered alert system that we coded as red/yellow/green. Due to the low event rates, even high-risk patients would have 12-hour predicted probabilities less than 1%. As such, the nursing staff felt uncomfortable interpreting absolute percentage risk scores. Instead, we wanted to develop a system that aided clinical decision support. We ultimately envisioned a primary use case for the risk score as a rounding tool for the RRT. We wanted a "red" alert to be a likely event, so we chose a threshold that would have a positive predictive value (PPV) of 10%. We wanted the green alert to be a likely nonevent so we chose the yellow and red combination to have a sensitivity of 80%, that is, 80% of all events would be a yellow or red. Using the 6 months of testing data, we calculated the average number of daily alerts, false positives, true positives, and false negatives per day.

Prospective Evaluation

Finally, we implemented the model into the EPIC system in a "silent" mode. After 10 weeks, we evaluated both the performance of the predictive model as well as the clinical decision support tool. We evaluated the model's performance based on its ability to obtain the prespecified thresholds clinically desired operating characteristics.

Table 1 Baseline Characteristics

N	87,897
Demographics	
Age (Median, 25th–75th)	61 (49–71)
Percent female	48.0%
Race	
African American	29.1%
White or Caucasian	65.0%
Other	5.9%
Groupers	
Diabetes	30.2%
Malignancy	29.4%
Chronic kidney disease	20.0%
Chronic obstructive pulmonary disease	11.9%
Myocardial infarction	3.8%
Stroke	5.9%
HIV	1.1%
Do not attempt resuscitation	9.8%
Transplant	1.8%
Outcomes	
Discharged	96.3%
ICU transfer	2.8%
Expired	0.9%
Time to event (days)	
Median (25th–75th)	3.66 (1.90–6.45)

ICU, intensive care unit.

We performed all analyses in R 3.4.2. Our institutional review board approved this work. The authors have no conflicts to report.

Results

From January 1, 2014, to December 31, 2016, there were 87 897 individual patient hospitalizations; 53% were on surgical wards. Of these hospitalizations, 2.5% resulted in an ICU transfer and 0.9% of which resulted in an inpatient death. Most events happened within the first 2 days of the admission (71%). The median length of stay was 3.7 days. Table 1 reports patient characteristics. Figure 1 presents a cumulative incidence curve.

Model Performance

We used 73,215 individuals to form the training data set and the remaining 14,682 as the testing data set. Figure 2 shows model coefficients, based on 10-fold cross-validation of a LASSO logistic regression. Some of the most predictive variables include systolic blood pressure, pulse, supplemental oxygen, type of admission, and respiratory rate. In general, variables used to calculate NEWS were the most

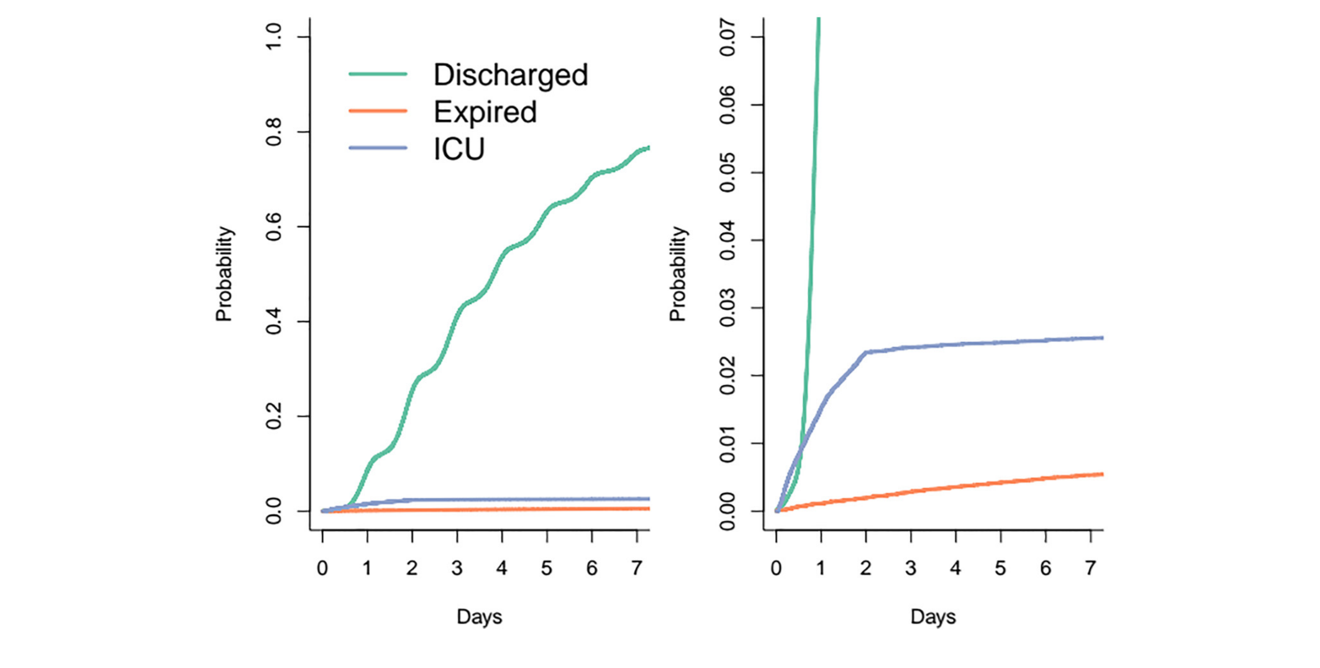


Figure 1 Cumulative incidence curves for time to intensive care unit (ICU) transfer, death, and discharged. Most events happen within the first 2 days of the admission.

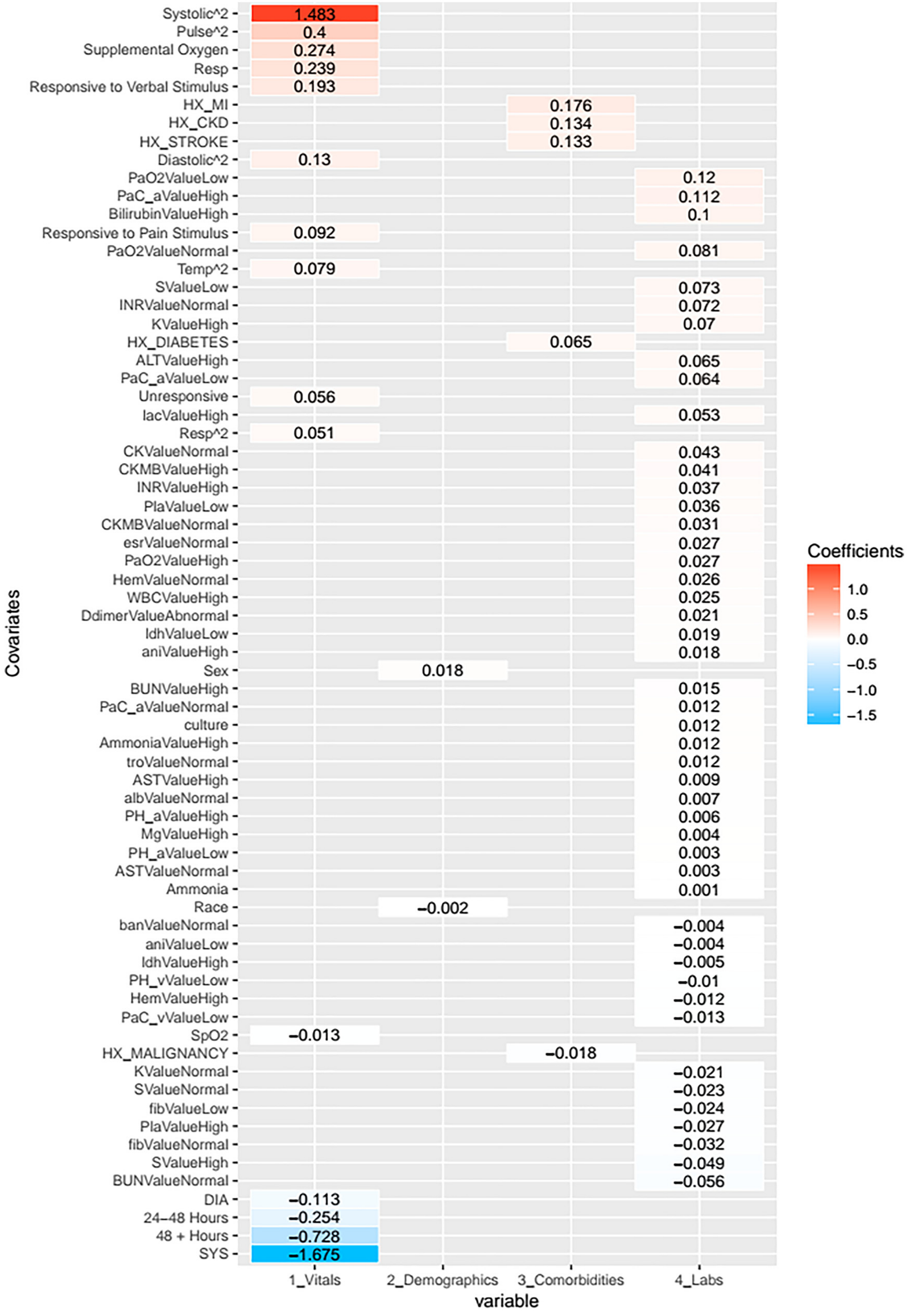


Figure 2 Standardized beta coefficients from the LASSO regression model fit. Variables are standardized unit variance to be comparable across. The color represents the magnitude of the coefficient. The strongest predictors are vital signs.

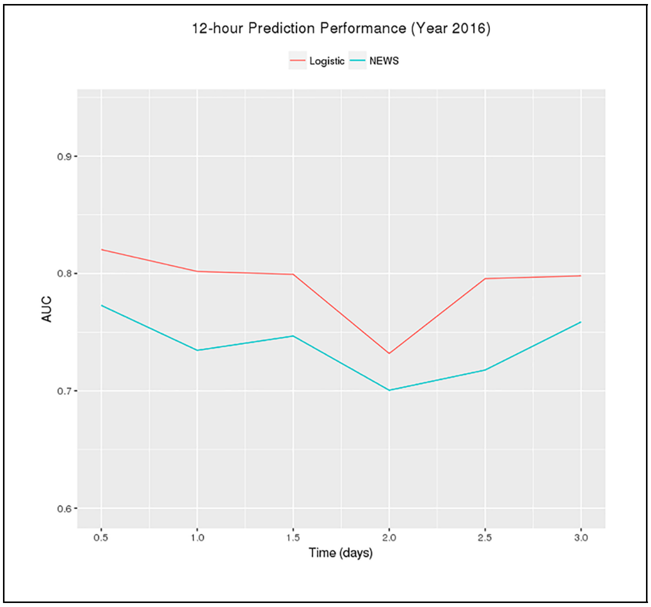


Figure 3 Predictive performance over time from admission based on AUC over the developed model (red) compared to NEWS (blue). The developed model has better overall predictive performance.

predictive variables. Figure 3 shows the time-varying AUC for our internally constructed risk score and the NEWS, within the testing data. The average AUC for the proposed model is 0.814 (95% CI = 0.79–0.83) compared with 0.740 (95% CI = 0.72–0.76) for the NEWS (test for difference $P < 0.001$). Of note, the locally derived model performs better earlier in a patient's admission (first 60 hours) when most events occur. After 60 hours, both models have similarly strong performance.

To illustrate why the proposed model performs better, we chose four patients who had an adverse event and compared their time-varying NEWS score to their time-varying rapid-deterioration score (Figure 4). We see that our score represents patient's health condition more thoroughly and is more sensitive to changes in patient's health condition.

Clinical Decision Support

After discussion with nursing staff, when implementing the model, we decided not to display the predicted risk as an absolute number. Instead, we constructed a red/yellow/green alert system that would indicate level of

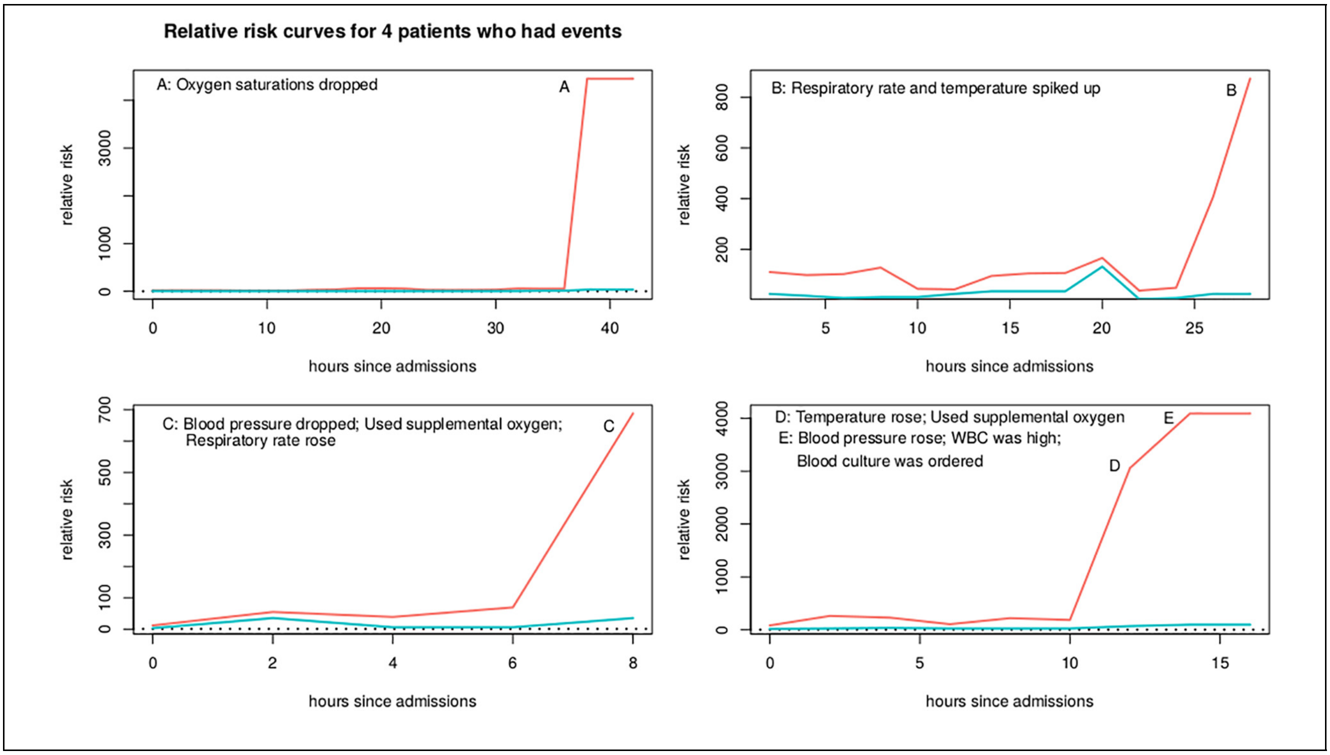
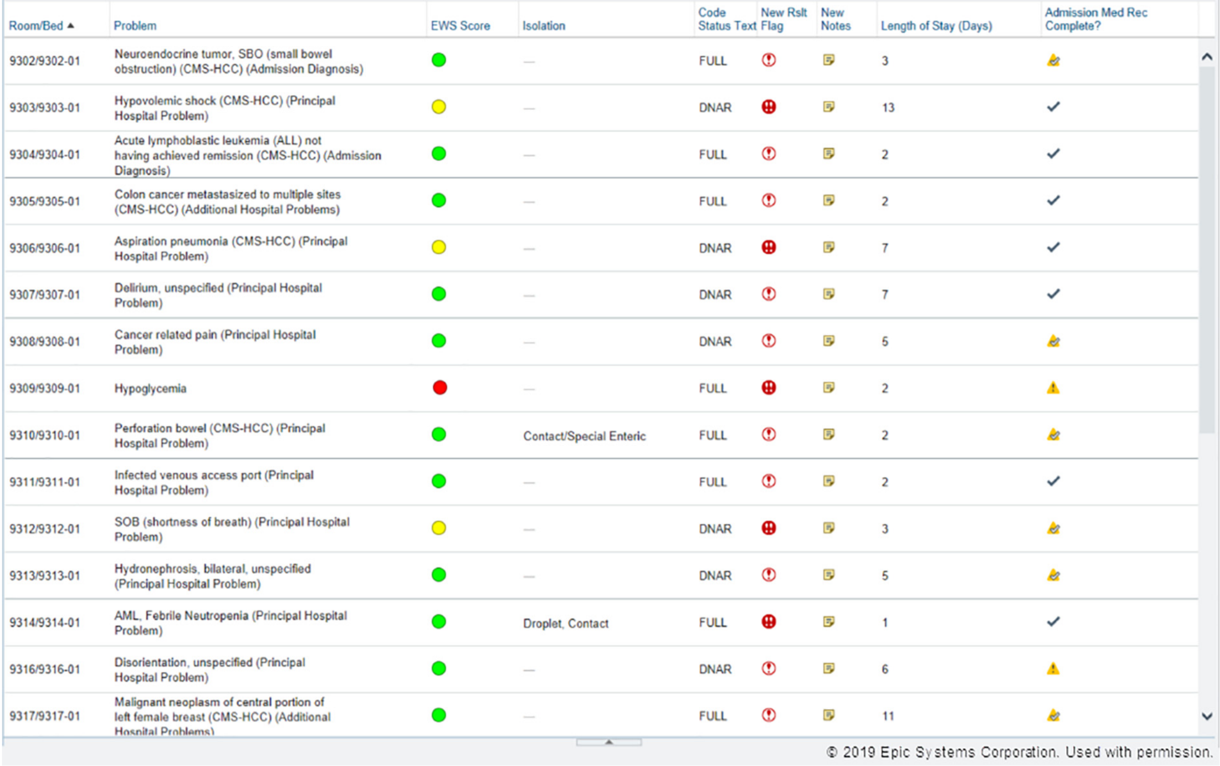


Figure 4 Risk curves—based on relative risk—for four selected individuals with events based on the developed model (red) and NEWS (blue). Events happened at the end of the time interval. Annotation indicates what changed in the patient's risk profile. In general, the developed model generates higher predicted risks.



Room/Bed ▲	Problem	EWS Score	Isolation	Code Status Text	New Rslt Flag	New Notes	Length of Stay (Days)	Admission Med Rec Complete?
9302/9302-01	Neuroendocrine tumor, SBO (small bowel obstruction) (CMS-HCC) (Admission Diagnosis)	🟢	—	FULL	!	✎	3	✎
9303/9303-01	Hypovolemic shock (CMS-HCC) (Principal Hospital Problem)	🟡	—	DNAR	❗	✎	13	✓
9304/9304-01	Acute lymphoblastic leukemia (ALL) not having achieved remission (CMS-HCC) (Admission Diagnosis)	🟢	—	FULL	!	✎	2	✓
9305/9305-01	Colon cancer metastasized to multiple sites (CMS-HCC) (Additional Hospital Problems)	🟢	—	FULL	!	✎	2	✓
9306/9306-01	Aspiration pneumonia (CMS-HCC) (Principal Hospital Problem)	🟡	—	DNAR	❗	✎	7	✓
9307/9307-01	Delirium, unspecified (Principal Hospital Problem)	🟢	—	DNAR	!	✎	7	✓
9308/9308-01	Cancer related pain (Principal Hospital Problem)	🟢	—	DNAR	!	✎	5	✎
9309/9309-01	Hypoglycemia	🔴	—	FULL	❗	✎	2	⚠
9310/9310-01	Perforation bowel (CMS-HCC) (Principal Hospital Problem)	🟢	Contact/Special Enteric	FULL	!	✎	2	✎
9311/9311-01	Infected venous access port (Principal Hospital Problem)	🟢	—	FULL	!	✎	2	✓
9312/9312-01	SOB (shortness of breath) (Principal Hospital Problem)	🟡	—	DNAR	❗	✎	3	✎
9313/9313-01	Hydronephrosis, bilateral, unspecified (Principal Hospital Problem)	🟢	—	DNAR	!	✎	5	✎
9314/9314-01	AML, Febrile Neutropenia (Principal Hospital Problem)	🟢	Droplet, Contact	FULL	❗	✎	1	✓
9316/9316-01	Disorientation, unspecified (Principal Hospital Problem)	🟢	—	DNAR	!	✎	6	⚠
9317/9317-01	Malignant neoplasm of central portion of left female breast (CMS-HCC) (Additional Hospital Problems)	🟢	—	FULL	!	✎	11	✎

© 2019 Epic Systems Corporation. Used with permission.

Figure 5 The developed EPIC dashboard. Red lights are people with high risk, yellow lights people with moderate risk, and green lights people with low risk.

risk. By creating a "red" bucket clinical staff would be able to easily identify high-risk individuals. For the red alert, we chose cut-point thresholds that would generate a PPV of 10%. For the yellow alert, we chose a threshold such that the red/yellow combination would have a sensitivity of 80%, capturing most events. To choose the proper thresholds, we evaluated the performance of the clinical support tool on a daily basis, as of 7 AM. Figure 5 provides a screen shot of the risk score as visualized by clinical staff within the EPIC interface.

Prospective Evaluation

After fitting the model, we ran the algorithm prospectively for 10 weeks. During this period there were 4210 patient encounters, with 33 deaths and 97 patients transfers to an ICU floor from a location of interest. Results indicated that the score had strong predictive performance, particularly in the near term (Table 2). The 2-hour AUC was 0.794 (95% CI: = 0.71–0.88). This decreased to 0.750 (95% CI = 0.73–0.78) and 0.731 (95% CI = 0.71–0.75) for 12-hour and 24-hour risk, respectively. The predictive performance for the longer term horizons was significantly better than the NEWS ($P < 0.01$). We also evaluated the performance of the clinical decision support tool. Our "red" category had a PPV of 10.8%. However, the red and yellow combination had a sensitivity of 32.9%, suggesting that we need to lower the yellow threshold. Overall, compared with the NEWS decision rule, our decision had better PPV and sensitivity.

Discussion

Our results highlight the potential for using one's own EHR data to develop a risk model as opposed to relying on off-the-shelf scores such as the NEWS. We were able to use retrospective data to develop a multivariable risk model for patient deterioration that incorporates not only vital signs but also demographics, laboratory tests, and comorbidities. Moreover, we were able to tailor this to our clinical environment and workflow. While this

Table 2 Predictive Performance for Death/ICU Transfer Based on Prospective Data

	Implemented Model	NEWS	*P* Value
AUC			
2-Hour window	0.794 (0.71–0.88)	0.732 (0.68–0.79)	.24
6-Hour window	0.778 (0.74–0.81)	0.715 (0.68–0.76)	.022
12-Hour window	0.750 (0.73–0.78)	0.69 (0.66–0.72)	.003
24-Hour window	0.731 (0.71–0.75)	0.68 (0.66–0.70)	<.001
PPV (%)			
Green	0.28	0.37	
Yellow	1.70		
Red	10.81	1.85	
Sensitivity (%)			
Green	67.07	76.12	
Yellow	28.05		
Red	4.88	23.88	

AUC, area under the curve; ICU, intensive care unit; NEWS, National Early Warning Score; PPV, positive predicted value.

degree of customization limits the ability to port this score to other environments, the process and lessons learned are transferable.

One of the motivations for this work was the poor performance of the NEWS in our patient population. Our initial analyses suggested that the NEWS variables were associated with deterioration, but that the coefficients, or variable weights, were not optimal for our patient population.[2] We confirmed this finding in this analysis. Even after incorporating additional predictors such as demographics, comorbidities, medications, and laboratory values, the strongest predictors are still those vital signs incorporated in the original NEWS. Similarly, previous work has found that vital signs are most predictive of near-term outcomes.[16] Therefore, even though others have found that additional risk factors are useful for assessing patient deterioration,[17,18] the creators of the original NEWS likely identified the correct risk factors.

An important consideration in our work was how to handle the time-varying data. Modern EHR systems allow one to extract time-stamped data on vital signs and laboratory measures. Due to the constraints of the implementation environment, we were limited to a regression-based approach. By using a LASSO-based model, we were able to incorporate complex effects (i.e., nonlinear terms and interactions) while maintaining a stable model. While there are a variety of machine learning methods that we could have used to model these data, previous work has shown that regularized models often perform quite well in comparison.[19] Another modeling constraint was how to handle the longitudinal information. We were only able to use the last clinical value. While there has been interest in understanding trajectories of clinical measurements,[20] some work has suggested that these do not provide much added value from a predictive context.[21,22]

A unique component of our approach is how we chose to implement the risk score. Since the event is relatively rare, we did not want to have a hard decision rule that would likely have poor operating characteristics. We considered simply reported individual risk but decided that the low predicted probabilities would be confusing to clinical staff. Instead, we implemented a three-tiered system. The highest tier (red) would be a rare group that would be more likely to have an event, though only still at an approximate 10% rate. The middle group (yellow)—along with the red group—would cover 80% of all events. This would leave the green group of patients to be those who clinical staff could feel more comfortable would not deteriorate. In fact, the expected PPV of this green group based on the testing data is 0.2%, indicating that while there are events it is likely to be quite rare. Such a clinical decision support system has a variety of advantages. First, we were able to customize these thresholds for our own environment and workflow. A model for a different outcome that has a different event rate may choose either higher or lower thresholds. Second, it makes decision making easier for clinical staff that are already overloaded with too many alerts and may not be familiar with probabilities.[23] Third, it addresses one of the primary critiques from nurses regarding our original NEWS implementation, which was there were too many false positives. Our retrospective analysis found that over 85% of alerts were ignored.[2] By creating a three-level system, where the top level has been a high PPV as well as known event rate, we hope that the alert will have a more positive impact. Fourth, by performing ongoing monitoring of the score, these thresholds are easily alterable over time to ensure that the score maintains the desired operating characteristics.

An important consideration is how to set up the data when developing the predictive model. We processed the data into 2-hour blocks to capture the time-varying nature of the data. Since not all laboratory tests were collected on all patients, and the presence of a laboratory is likely informative,[24] we transformed the data into a categorical variable. It is important to note that this likely led to loss of information,[25] but avoided bias from an imputation strategy.

There are some limitations to our work. As we have stressed, the estimated model coefficients are specific to our clinical environment and are not transferable to other institutions, instead requiring external validation.[26] Instead, the most transferable component of this work is the analytic approach taken. Second, even the analytic approach is likely not optimal. The EPIC system constrained us to use a linear-based model. It is likely that more sophisticated machine learning methods, for example, deep learning, would yield a better model, as has been developed by others.[27] In parallel, we are actively assessing the added value of these methods and hope to implement them when newer versions of EPIC become available. Third, while our prospective evaluation suggests that our model has the desired operating characteristics, this does not necessarily mean that this will translate into improved patient care. Deterioration represents a highly heterogeneous population, and it is not necessarily clear what ought to be done, even for an at-risk patient. Related, we were not able to capture all types of decompensations (e.g., calls to a rapid response team that did not result in an ICU transfer) nor were we able to distinguish between expected and unexpected mortalities.

Conclusion

Overall, this work illustrates how health systems can use their available EHR data to develop tailored risk prediction tools. One should expect that locally derived models would perform better and have greater flexibility than off-the-shelf scores. However, the off-the-shelf scores can be used as a reasonable starting point for model development.

ORCID iD

Benjamin A. Goldstein https://orcid.org/0000-0001-5261-3632

Supplementary Material

Supplementary material for this article is available on the *Medical Decision Making Policy & Practice* website at http://journals.sagepub.com/home/mpp.

References

1. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. 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(4):465–70. doi:10.1016/j.resuscitation.2012.12.016
2. Bedoya AD, Clement ME, Phelan M, Steorts RC, O'Brien C, Goldstein BA. Minimal impact of implemented Early Warning Score and best practice alert for patient deterioration. *Crit Care Med*. 2019;47(1):49–55. doi:10.1097/CCM.0000000000003439
3. Alam N, Hobbelink EL, van Tienhoven AJ, van de Ven PM, Jansma EP, Nanayakkara PWB. The impact of the use of the Early Warning Score (EWS) on patient outcomes: a systematic review. *Resuscitation*. 2014;85(5):587–94. doi:10.1016/j.resuscitation.2014.01.013
4. Yiu CJ, Khan SU, Subbe CP, Tofeec K, Madge RA. Into the night: factors affecting response to abnormal Early Warning Scores out-of-hours and implications for service improvement. *Acute Med*. 2014;13(2):56–60.
5. Kolic I, Crane S, McCartney S, Perkins Z, Taylor A. Factors affecting response to national early warning score (NEWS). *Resuscitation*. 2015;90:85–90. doi:10.1016/j.resuscitation.2015.02.009
6. Ding X, Gellad ZF, Mather C 3rd, et al. Designing risk prediction models for ambulatory no-shows across different specialties and clinics. *J Am Med Inform Assoc*. 2018;25(8):924–30. doi:10.1093/jamia/ocy002
7. Kolek MJ, Graves AJ, Xu M, et al. Evaluation of a prediction model for the development of atrial fibrillation in a repository of electronic medical records. *JAMA Cardiol*. 2016;1(9):1007–13. doi:10.1001/jamacardio.2016.3366
8. 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. doi:10.1126/scitranslmed.aab3719
9. 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;64:10–9. doi:10.1016/j.jbi.2016.09.013
10. Green M, Lander H, Snyder A, Hudson P, Churpek M, Edelson D. Comparison of the Between the Flags calling criteria to the MEWS, NEWS and the electronic Cardiac Arrest Risk Triage (eCART) score for the identification of deteriorating ward patients. *Resuscitation*. 2018;123:86–91. doi:10.1016/j.resuscitation.2017.10.028
11. Epic. Available from: https://www.epic.com/.

12. Churpek MM, Wendlandt B, Zadravecz FJ, Adhikari R, Winslow C, Edelson DP. Association between intensive care unit transfer delay and hospital mortality: a multicenter investigation. *J Hosp Med*. 2016;11(11):757–62. doi: 10.1002/jhm.2630
13. Liu V, Kipnis P, Rizk NW, Escobar GJ. Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. *J Hosp Med*. 2012;7(3):224–30. doi:10.1002/jhm.964
14. Venkatraman E. A distribution-free procedure for comparing receiver operating characteristic curves for a paired experiment. *Biometrika*. 1996;83(4):835–48. doi:10.1093/biomet/83.4.835
15. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. doi:10.1186/1471-2105-12-77
16. Goldstein BA, Pencina MJ, Montez-Rath ME, Winkelmayer WC. Predicting mortality over different time horizons: which data elements are needed? *J Am Med Inform Assoc*. 2017;24(1):176–81. doi:10.1093/jamia/ocw057
17. Jo S, Yoon J, Lee JB, Jin Y, Jeong T, Park B. Predictive value of the National Early Warning Score—lactate for mortality and the need for critical care among general emergency department patients. *J Crit Care*. 2016;36:60–8. doi:10.1016/j.jcrc.2016.06.016
18. Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation*. 2016;102:1–5. doi:10.1016/j.resuscitation.2016.02.005
19. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol*. 2013;66(4):398–407. doi:10.1016/j.jclinepi.2012.11.008
20. Goldstein BA, Assimes T, Winkelmayer WC, Hastie T. Detecting clinically meaningful biomarkers with repeated measurements: an illustration with electronic health records. *Biometrics*. 2015;71(2):478–86. doi:10.1111/biom.12283
21. Goldstein BA, Pomann GM, Winkelmayer WC, Pencina MJ. A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis. *Stat Med*. 2017;36(17):2750–63. doi:10.1002/sim.7308
22. Sweeting MJ, Barrett JK, Thompson SG, Wood AM. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study. *Stat Med*. 2017;36(28):4514–28. doi:10.1002/sim.7144
23. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R; With the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Med Inform Decis Mak*. 2017;17(1):36. doi:10.1186/s12911-017-0430-8
24. Phelan M, Bhavsar NA, Goldstein BA. Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. *EGEMS (Wash DC)*. 2017;5(1):22. doi:10.5334/egems.243
25. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Stat Med*. 2016;35(23):4124–35. doi:10.1002/sim.6986
26. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–7. doi:10.1016/j.jclinepi.2015.04.005
27. 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. doi:10.1016/S2213-2600(18)30300-X