ORIGINAL ARTICLE

Detecting Deteriorating Patients in the Hospital
Development and Validation of a Novel Scoring System

Marco A. F. Pimentel[1]*, Oliver C. Redfern[2]*, James Malycha[2], Paul Meredith[3], David Prytherch[4], Jim Briggs[4], J. Duncan Young[2], David A. Clifton[1], Lionel Tarassenko[1], and Peter J. Watkinson[2,5]

[1]Institute of Biomedical Engineering, Department of Engineering Science, and [2]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom; [3]Research and Innovation Department, Portsmouth Hospitals University National Health Service Trust, Portsmouth, United Kingdom; [4]Centre for Healthcare Modelling and Informatics, University of Portsmouth, Portsmouth, United Kingdom; and [5]Kadoorie Centre for Critical Care Research and Education, Oxford University Hospitals National Health Service Trust, Oxford, United Kingdom

Abstract

Rationale: Late recognition of patient deterioration in hospital is associated with worse outcomes, including higher mortality. Despite the widespread introduction of early warning score (EWS) systems and electronic health records, deterioration still goes unrecognized.

Objectives: To develop and externally validate a Hospital-wide Alerting via Electronic Noticeboard (HAVEN) system to identify hospitalized patients at risk of reversible deterioration.

Methods: This was a retrospective cohort study of patients 16 years of age or above admitted to four UK hospitals. The primary outcome was cardiac arrest or unplanned admission to the ICU. We used patient data (vital signs, laboratory tests, comorbidities, and frailty) from one hospital to train a machine-learning model (gradient boosting trees). We internally and externally validated the model and compared its performance with existing scoring systems (including the National EWS, laboratory-based acute physiology score, and electronic cardiac arrest risk triage score).

Measurements and Main Results: We developed the HAVEN model using 230,415 patient admissions to a single hospital. We validated HAVEN on 266,295 admissions to four hospitals. HAVEN showed substantially higher discrimination (c-statistic, 0.901 [95% confidence interval, 0.898–0.903]) for the primary outcome within 24 hours of each measurement than other published scoring systems (which range from 0.700 [0.696–0.704] to 0.863 [0.860–0.865]). With a precision of 10%, HAVEN was able to identify 42% of cardiac arrests or unplanned ICU admissions with a lead time of up to 48 hours in advance, compared with 22% by the next best system.

Conclusions: The HAVEN machine-learning algorithm for early identification of in-hospital deterioration significantly outperforms other published scores such as the National EWS.

Keywords: machine learning; physiological monitoring; retrospective observational study; early warning scores

(Received in original form July 7, 2020; accepted in final form February 1, 2021)

*Joint first authors.

Supported by the Health Innovation Challenge Fund grants HICF-R9-524 and WT-103703/Z/14/Z; a parallel funding partnership between the UK Department of Health and Social Care and the Wellcome Trust; the National Institute for Health Research Biomedical Research Centre, Oxford (P.J.W., D.A.C., and L.T.); and Drayson Research fellowships (M.A.F.P. and O.C.R.). The views expressed in this publication are those of the authors and not necessarily those of the UK Department of Health and Social Care or the Wellcome Trust.

This article, which is the result of a 5-year collaboration, primarily between the Oxford Critical Care Research Group and the Portsmouth Clinical Outcomes Research Team, is dedicated to the memory of Mrs. Isabelle Tolman (née Tarassenko), who was a patient in the Portsmouth ICU for the last week of her life and is the sister of one of the Oxford co-authors.

Model availability: The authors make a commitment to sharing the algorithm with investigators seeking to validate it for academic and noncommercial purposes. Requests should be made to Professor Peter J. Watkinson (peter.watkinson@ndcn.ox.ac.uk) or Dr. Oliver C. Redfern (ccrg.research@ndcn.ox.ac.uk).

Author Contributions: Study design: D.P., J.B., J.D.Y., D.A.C., L.T., and P.J.W. Data collection: M.A.F.P. and P.M. Data analysis: M.A.F.P., O.C.R., and J.M. Data interpretation and writing the paper: all authors.

Correspondence and requests for reprints should be addressed to Oliver C. Redfern, B.Sc., M.B. B.S., Ph.D., Nuffield Department of Clinical Neurosciences, University of Oxford, Level 6, West Wing, John Radcliffe Hospital, Oxford OX3 9DU, UK. E-mail: oliver.redfern@ndcn.ox.ac.uk.

This article has a related editorial.

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Am J Respir Crit Care Med Vol 204, Iss 1, pp 44–52, Jul 1 2021

Originally Published in Press as DOI: 10.1164/rccm.202007-2700OC on February 1, 2021
Internet address: www.atsjournals.org

At a Glance Commentary

Scientific Knowledge on the Subject: Late recognition of patient deterioration in hospital is associated with worse patient outcomes. Current early warning score systems based purely on vital sign measurements still do not identify the majority of deteriorations without also generating many false alerts.

What This Study Adds to the Field: We used a machine-learning algorithm to combine patients' vital signs with additional physiological measurements, comorbidities, and frailty to create the Hospital-wide Alerting via Electronic Noticeboard scoring system. This model substantially increased the precision with which deteriorating patients could be identified when compared with previously published scores.

Over 60,000 patients annually deteriorate on UK hospital wards to the extent that they require ICU admission (1). Late or missed recognition of deterioration is associated with worse patient outcomes, including higher mortality (2–4). Over the past 20 years, healthcare systems worldwide have implemented alerting systems to improve the detection of patients at risk of deterioration (5–7). Most are based on abnormalities in patients' vital signs, usually by combining them into an early warning score (EWS). Clinicians are alerted when the EWS rises above a given threshold. Many hospitals also employ rapid response teams to respond to the most critically unwell patients (8). However, there is conflicting evidence that implemented EWS systems or rapid response teams improve patient outcomes (8–11).

Current EWSs were designed to be calculated easily at the bedside when most hospitals recorded observations on paper charts. This simplicity means EWSs cannot account for trends over time, patients with chronically abnormal physiology, or other indicators of deterioration (e.g., acute kidney injury). Consequently, EWSs commonly generate false alerts, risking alarm fatigue and increasing the likelihood that deteriorating patients are missed (12).

Increased uptake of electronic health records (EHRs) facilitates the development of sophisticated EWSs incorporating additional routinely collected patient data. For example, our group and others have shown that combining laboratory results with vital sign measurements increases the precision with which deteriorating patients can be detected (13–19). Many newer risk scores exploit machine-learning algorithms (13, 15, 17, 20–24). However, few are externally validated (25–27) and fewer still are implemented in the EHR (23). Those that have are often subject to proprietary licenses, which can limit the research community's ability to validate them (22, 23, 28, 29). Some algorithms also use data, such as detailed nursing assessments, that are not routinely recorded in the EHR (28). A key reason predictive machine-learning models are not clinically implemented is the failure to consider whether they add value in clinical practice (15, 30, 31). Indeed, we previously argued that even current EWS systems are not optimized to identify patients with reversible deterioration; namely, where intervention is likely to change patient outcomes (32).

In this study, we describe the development and external validation of the Hospital-wide Alerting via Electronic Noticeboard (HAVEN) system to identify patients with potentially reversible deterioration. HAVEN provides a real-time risk assessment, which is continuously updated using patients' vital signs, laboratory test results, and medical histories.

Methods

Study Design

A multicenter retrospective development and external validation of a prognostic model. It is reported following the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines (33).

Ethical Approval

This work received Health Research Authority, Research Ethics Committee (REC) (reference number 16/SC/0264 from the South Central Oxford C REC, and 08/02/1394 from the Isle of Wight, Portsmouth, and South East Hampshire REC), and Confidentiality Advisory Group (16/CAG/0066) approval.

Setting

Patient data were collected retrospectively from two separate UK hospital groups: Portsmouth Hospitals National Health Service (NHS) Trust and Oxford University Hospitals NHS Trust. Data were extracted, linked, and deidentified before being made available to the research team.

Portsmouth Hospitals NHS Trust is a large, acute, district general hospital (hospital A) with approximately 1,250 beds, which provides a full range of elective and emergency medical and surgical services to a local population of around 675,000 (34). Oxford University Hospitals NHS Trust is a hospital group with approximately 1,465 beds, which serves a local population of around 655,000. We included the tertiary referral center for trauma, cardiology, and neurosurgery, which also provides general acute medical and surgical services (hospital B); the specialist renal transplant and cancer referral center (hospital C); and the district general hospital (hospital D). We excluded a hospital performing predominantly elective orthopedic procedures.

Data Sources

The routinely collected data stored across different clinical information systems in all four hospitals were extracted. Data included admissions' administrative information (including dates and timings for admission, discharge, and any transfers within the hospital site), diagnoses as 10th-revision International statistical Classification of Diseases and related health problems (ICD-10) codes, laboratory results (including hematology, biochemistry, and microbiology results), vital signs, and patient demographics.

Participants

We included all patients (aged 16 or above) admitted to hospital A from January 2012 to December 2017 or admitted to hospitals B–D from January 2016, to December 2017.

Admissions with no recorded vital signs were excluded to ensure a minimum required data set for score computation.

The training cohort comprised admissions to hospital A from January 2012 to December 2015. The primary test cohort combined admissions from hospitals A–D between January 2016, and December 2017.

Outcomes

Our primary outcome was a composite of in-hospital cardiac arrest and unplanned admission to the ICU not preceded by surgery in the prior 24 hours. ICU admissions shortly after surgery were excluded, as deterioration may happen during the procedure rather than on the ward. Secondary outcomes were unplanned admission to the ICU not preceded by surgery in the initial 24 hours and in-hospital cardiac arrest separately. We included a third secondary outcome of all unplanned admissions to the ICU to determine the effect of including unplanned ICU admissions preceded by surgery within 24 hours.

Predictors

We identified potential variables for inclusion in the model by a systematic literature search (35) and expert suggestions, followed by an expert panel review. The expert panel comprised critical care nurses and doctors, alongside a statistician and senior general physician. The panel undertook a modified Delphi process to consider additional variables useful in determining patients' risk of deterioration. A consensus was reached after two discussion rounds, resulting in a final 76-candidate variable list.

Each patient admission was represented by *static* (time-invariant) and *dynamic* (time-varying) variables.

As *static* variables, we included the patient's age and sex at admission to the hospital. We also encoded the presence or absence of comorbidities using ICD-10 diagnosis codes. Because diagnostic coding in the United Kingdom typically occurs after discharge from the hospital, this information is not available electronically unless the patient has previously been admitted to the same hospital. We extracted unique diagnostic codes from previous admissions over the 2 years before hospital admission under study. Diagnostic codes were grouped into 30 categories according to Elixhauser (36), comprising 30 binary features encoding whether patients had common chronic diseases, such as congestive heart failure or chronic lung disease. We further calculated: smoking status (using the ICD-10 codes F17, Z716, and Z720), the Hospital Frailty Risk Score (37), and the total length of all hospitalizations in the 2 prior years.

As *dynamic* variables, we included commonly measured laboratory values and vital signs and the estimated inspired oxygen fraction. A variable list is provided in the online supplement (Section D).

We designed HAVEN to recalculate a patient's deterioration risk each time a new variable is recorded. When one time-varying variable is measured, other variables often are not. We therefore included the most recent measured value for each physiological and laboratory result variable at each time point (equivalent to a last value carried forward imputation). To capture how variables change over time, we also calculated two derived features before imputation: a 24-hour variability index for physiological variables (38) (defined as the difference between the maximum and minimum values over the preceding 24 h) and the maximum and minimum values of laboratory results recorded during the patient's admission before the time point (both including the current measurement).

Missing Data

Distributions of variables were inspected manually. A clinical expert panel identified "biologically implausible" ranges, with values outside these ranges defined as missing.

The remaining missing values were imputed with the median (or mode for dichotomous variables) of each variable from the training set. Although other methods were considered, such as multiple imputations (39), we used the median and mode to simulate the HAVEN implementation within a live clinical system.

Statistical Analysis

Model development. We trained the HAVEN system by generating the set of features for each time point in which a new measurement (vital sign or laboratory test) occurred. We labeled each time point as "positive" if the primary outcome occurred within 24 hours. We used a gradient boosting machine with decision trees, as implemented in the XGBoost library (40). XGBoost has a number of hyperparameters (e.g., the depth of the component decision trees) that are modifiable to produce the best model. One of these hyperparameters changes the relative weighting between the positive and negative classes, which can improve model performance in unbalanced data sets. To discover the optimal hyperparameters, we used a random search (500 permutations) and selected the model with the highest c-statistic (using a fivefold cross-validation procedure), using the first 3 years of data in the training set.

Optimal model predictions were recalibrated on the training set's final year of data to reflect the frequency of observed outcomes using isotonic regression (41). Uncalibrated and calibrated predictions were compared using reliability plots (41).

In addition to the gradient boosting machine, we trained, optimized, and validated four alternative machine-learning models: a Random Forest, a Generalized Additive Model, and both L1-regularized (Lasso regression) and L2-regularized logistic regression models (*see* Table EA6 in the online supplement).

Model evaluation. We evaluated risk prediction model performance using the test set containing data from all four hospitals. In line with Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guidance, we report results for individual hospitals and for the three hospitals not used to develop HAVEN (33). We report model performance using discrimination and calibration metrics computed at both the "observation" and "patient admission" levels. We designed HAVEN to identify patients at risk of deterioration on hospital wards rather than identifying direct admissions from the emergency department—for this reason, scores generated from emergency department measurements were excluded.

At the observation level, we calculated the area under the curve (AUC) for the receiver operating curve (ROC) for our outcome measures occurring within the subsequent 12-, 24-, and 48-hour periods of each measurement (i.e., each time a measurement is recorded). The ROC AUC (c-statistic) measures discrimination, corresponding to the probability that patients who experience the outcome will be ranked above those who do not. As the outcomes are relatively rare (there are many more patients who go home without an event than there are patients who have an unplanned ICU admission or a cardiac arrest), we also computed the AUC for the precision-recall curve (PR), which can be informative in class-imbalanced data sets (42, 43). The PR AUC shows the trade-off between precision (positive predictive value) and recall (sensitivity) at each threshold. The closer the PR AUC is to 1, the greater the ability of the score or model to detect true cases (recall) with high precision over the range of thresholds. Calibration curves for selected models were

determined for outcome occurrence within 24 hours of each measurement.

The sequential nature of predictions means the total number of positive time steps (in which the outcome occurs within *n* hours) does not directly correspond to the number of patients experiencing the outcome. Multiple positive time steps may be associated with a single adverse event. To assess the clinical applicability of the proposed model, we calculated the "patient admission sensitivity" at different degrees of precision (5%, 10%, 20%). These precisions correspond to evaluating 20, 10, and 5 patients, respectively, for each true-positive result—also known as the number needed to evaluate (NNE) (44). For each degree of precision, a patient admission was considered a false-positive result if they had at least one score above the threshold and no adverse event occurred. True-positive results were patient admissions with at least one score above the threshold in the *n* hours before an adverse event. We examined the sensitivity of the model over different time prediction windows preceding the event (up to 48 h). To further evaluate clinical utility, we performed a decision curve analysis (45–48).

All 95% confidence intervals (CIs) were calculated using bootstrapping (200 samples) (49). We used the Shapley additive explanation algorithm (50) to calculate the relative "importance" of each predictor in the final model (*see* Section F of the online supplement).

Comparison with Published Risk Scoring Systems

We compared HAVEN score performance with established EWS systems: the centile-based EWS (51), the modified EWS (52), the standardized EWS (53), the National EWS (NEWS) (54), and the cardiac arrest risk triage (CART) score (55). We also compared it with three physiological scoring systems: the NEWS:LDTEWS (13), the electronic CART (eCART) score (56), and the laboratory-based acute physiology score (LAPS-2) (57). We excluded scoring systems in which the coefficients were unpublished or where data (e.g., nursing assessments) were not routinely recorded in our study sites (22, 58). Further details of EWSs and other scoring systems are shown in in the online supplement (Section C).

Results

After exclusions, we included 496,710 unique admissions to four hospitals. The training set included 230,415 admissions (from 113,450 patients) to hospital A.

There were 266,295 admissions (159,182 patients) to four hospitals (A–D) in the test set. The two cohorts have similar patient characteristics (Table 1), both with a slightly higher proportion of female patients (of around 53%) and a median age of 62–63 years.

Table 1. Summary Description Statistics for the Cohorts

	Training	Test
Patients		
Unique patients	113,450	159,182
Age, yr*	63 (44–77)	62 (43–77)
Sex		
Males	52,720 (46.5%)	74,812 (47.0%)
Females	60,730 (53.5%)	84,370 (53.0%)
Ethnicity		
White	93,853 (82.7%)	120,706 (75.8%)
Mixed	337 (0.3%)	883 (0.6%)
Black	437 (0.4%)	1,196 (0.8%)
Asian	593 (0.5%)	2,698 (1.7%)
Other	543 (0.5%)	1,280 (0.8%)
Unknown	17,687 (15.6%)	32,421 (20.4%)
Admissions within period		
Hospital sites	1†	4†
Period	Jan 2012 to Dec 2015	Jan 2016 to Dec 2017
Unique admissions per patient	230,415	266,295
Median	1 (1–2)	1 (1–2)
Average	2.03 (2.55)	1.67 (1.66)
Length of stay, d	1.77 (0.54–5.26)	1.36 (0.36–4.76)
Elective admissions	81,703 (35.5%)	82,402 (30.9%)
Surgical admissions	102,603 (44.5%)	116,459 (43.7%)
In-hospital deaths	7,436 (3.2%)	7,880 (3.0%)
ICU admissions	2,863 (1.2%)	4,098 (1.5%)
Unpl.		
Unpl. Med.	2,004 (0.9%)	2,527 (0.9%)
Cardiac arrests	808 (0.4%)	647 (0.2%)
Primary outcome‡	2,695 (1.2%)	3,105 (1.2%)

Definition of abbreviations: NHS = National Health Service; Unpl. = admissions to the ICU defined as unplanned (or unanticipated); Unpl. Med. = admissions to ICU defined as unplanned and not preceded by a visit to the theater in the preceding 24 hours.
The median and interquartile range are shown for continuous variables.
*When multiple admissions are present, the age of the patient at the first admission is used.
†In total, data from four hospitals were included: three hospitals from one organization (Oxford University Hospitals NHS Foundation Trust) and one hospital from a different organization (Portsmouth Hospitals NHS Trust). Data from the latter (collected within different periods), was used for training and calibration.
‡Primary (composite) outcome is defined as the occurrence of a cardiac arrest and/or an unplanned admission to the ICU.

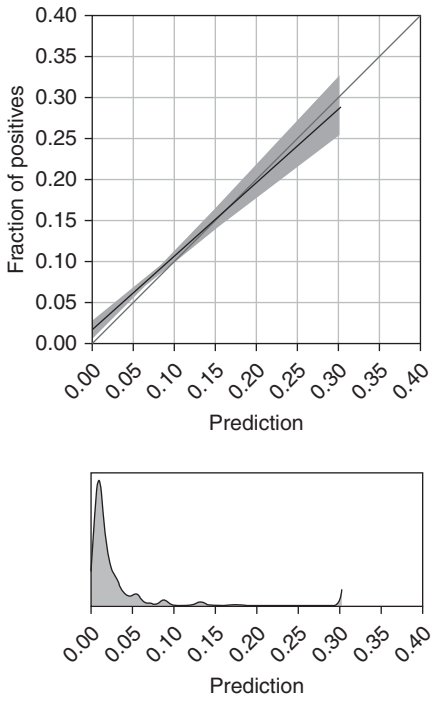

Figure 1. Calibration curve for Hospital-wide Alerting via Electronic Noticeboard predictions in the test set (top), alongside the distribution of Hospital-wide Alerting via Electronic Noticeboard predictions (bottom). The calibration curve (black) shows the locally estimated (scatterplot smoothing) smoothed observed probability versus estimated probability of adverse events (with 95% confidence bands). The diagonal line (gray) shows ideal calibration.

In the test cohort, 31% of admissions to the four hospitals (A–D) were elective, with a median hospital stay of 1.36 (interquartile range, 0.36–4.76) days. Hospital mortality was approximately 3%. In approximately 1% of admissions, patients had an unplanned ICU admission without visiting the operating theater in the preceding 24 hours. A cardiac arrest occurred during 647 admissions (0.2%). There was some variability in patient characteristics across the four hospitals (*see* Table EA1). Hospital C had a higher proportion of elective admissions (55.6%), a lower mortality rate (1.9%), and a higher rate of unplanned ICU admissions (3.9%) than the other hospitals. Class imbalance and the extent of missing data are reported in online supplement (Section E).

The calibration curve in the combined test set is shown in Figure 1. Table 2 shows HAVEN model performance on the test set for predicting the observation-level primary outcome (unplanned ICU admission or cardiac arrest) within different time windows. ROC AUC values increase as the time window moves closer to the event, from 0.881 (95% CI, 0.879–0.883) within the following 48 hours to 0.921 (95% CI, 0.919–0.924) within the following 12 hours. A similar trend in ROC AUC values occurs for the individual secondary outcomes (Table 2). HAVEN model performance (either by ROC or PR AUCs) was higher for predicting unplanned ICU admissions than for cardiac arrests (Table 2). The average contributions ("feature importance") of individual predictors are shown in the online supplement (Section F).

HAVEN performance was higher than all other published EWS and risk scores when predicting the primary outcome measured by either the ROC AUC or the PR AUC (Table 3). For example, for a time window of 24 hours, HAVEN had a ROC AUC of 0.901 (95% CI, 0.898–0.903), whereas LAPS-2, the next best-performing scoring system, had a ROC AUC of 0.863 (0.860–0.865). This improved performance remained when testing was restricted to individual hospitals (Tables EA2 and EA4) and to the three test hospitals (B–D) where HAVEN had not been developed (Table EA5). HAVEN performed as well or better than all other EWS and other risk scores for the individual secondary outcomes (*see* Table EA3).

Figure 2 shows the patient admission level sensitivity of HAVEN for different prediction time windows for three fixed degrees of precision. A greater proportion of events were correctly predicted, as outcomes are included closer to the prediction point. At 10% precision (NNE = 10), HAVEN identified 42% of adverse events occurring in the subsequent period of <1–48 hours and 27% of adverse events occurring between 12 and 48 hours after the prediction point. In comparison, LAPS-2 identified 22% and 14% of adverse events in the

Table 2. Model Performance Using the Entire Test Set

AUC by Time Window	Composite Outcome*	Unplanned ICU Admission	Cardiac Arrest
ROC AUC (95% CI)			
12 h	0.921 (0.919–0.924)	0.939 (0.936–0.941)	0.831 (0.823–0.840)
24 h	0.901 (0.898–0.903)	0.921 (0.919–0.923)	0.807 (0.800–0.814)
48 h	0.881 (0.879–0.883)	0.902 (0.900–0.904)	0.772 (0.765–0.779)
PR AUC (95% CI)			
12 h	0.073 (0.069–0.078)	0.076 (0.071–0.081)	0.006 (0.003–0.010)
24 h	0.080 (0.076–0.084)	0.083 (0.079–0.087)	0.006 (0.003–0.008)
48 h	0.081 (0.078–0.084)	0.084 (0.080–0.087)	0.006 (0.003–0.008)

Definition of abbreviations: AUC = area under the curve; CI = confidence interval; PR = precision recall; ROC = receiver operating characteristic. ROC AUC and PR AUC performance when predicting the risk of future adverse event (and each event separately, namely, unplanned admission to ICU and cardiac arrest) across different time windows.
*Composite outcome is defined as the occurrence of a cardiac arrest and/or an unplanned admission to the ICU.

Table 3. Comparative Performance of Scoring Systems Using the Entire Test Set

Scoring System	ROC AUC (95% CI)	PR AUC (95% CI)
CEWS	0.838 (0.834–0.841)	0.031 (0.028–0.033)*
MEWS	0.836 (0.833–0.839)	0.031 (0.028–0.033)
NEWS	0.842 (0.839–0.845)	0.028 (0.025–0.030)
SEWS	0.791 (0.788–0.795)	0.026 (0.024–0.028)
NEWS:LDTEWS	0.860 (0.858–0.863)†	0.029 (0.026–0.031)
CART	0.700 (0.696–0.704)	0.023 (0.021–0.025)
eCART	0.796 (0.792–0.800)	0.026 (0.024–0.029)
LAPS-2	0.863 (0.860–0.865)†	0.031 (0.028–0.033)*
HAVEN	0.901 (0.898–0.903)†	0.080 (0.076–0.084)*

Definition of abbreviations: AUC = area under the curve; CART = cardiac arrest risk triage; CEWS = centile-based EWS; CI = confidence interval; eCART = electronic CART; EWS = early warning score; HAVEN = Hospital-wide Alerting via Electronic Noticeboard; LAPS-2 = laboratory-based acute physiology score 2; LDTEWS = Laboratory Decision Tree EWS; MEWS = modified EWS; NEWS = National EWS; PR = precision recall; ROC = receiver operating characteristic; SEWS = standardized EWS.
ROC AUC and PR AUC performance when predicting the risk of future composite adverse event (unplanned admission to ICU and cardiac arrest) within 24 hours.
*Top 3 performing systems according to the PR AUC.
†Top 3 performing systems according to the ROC AUC.

corresponding time periods (Figure EB1). NEWS and LAPS-2 performed similarly. The total number of events becomes smaller as the window duration decreases. Nearly all patients were in the hospital for an hour before an event, but progressively fewer were hospitalized as the prediction horizon increased (roughly 60% of events occurred more than 24 h after admission to a general ward).

Decision curve analysis showed HAVEN had a higher net benefit than all other scoring systems over a range of risk thresholds (*see* Figures EB3 and EB4). Including unplanned ICU admissions preceded by a theater visit decreased the performance of HAVEN and all other scoring systems (Table EA4).

Discussion

Main Findings

In this large, retrospective, observational study, we developed a novel risk score (HAVEN) to identify hospitalized patients at risk of potentially reversible deterioration. HAVEN had higher discrimination than all previously published EWSs and physiological scoring systems we tested (Tables 2 and 3) and was well calibrated (Figure 1). At 10% precision, the model identified nearly twice as many adverse outcomes in advance of the event (depending on the prediction horizon) (Figure 2) as the next best scoring system, LAPS-2 (Figure EB1).

Strengths and Limitations

Our study used data from four large hospitals and follows the latest recommendations for developing and validating prediction models and EWSs (45, 59). We used a composite primary outcome of unplanned admission to the ICU and in-hospital cardiac arrest as a proxy for potentially reversible clinical deterioration, as no well-defined indicator of "reversible" deterioration is recorded. This contrasts with other studies that either used only one of these two outcomes or used in-hospital mortality (60–62). We deliberately excluded in-hospital mortality from our composite outcome. In the United Kingdom, 40–50% of deaths occur in hospitals and only 3.6% of these are estimated to be avoidable (63, 64). Excluding in-hospital mortality reduces the risk that our model would be optimized to predict inevitable, rather than potentially preventable, deterioration. The importance of outcome selection has been noted previously by ourselves and others (32, 61). LAPS-2 was optimized to predict in-hospital mortality, which may have impacted its performance in our study.

We excluded unplanned ICU admissions preceded by an operating theater visit from the primary outcome. We assessed the impact of this exclusion on HAVEN performance, finding (as with other scoring systems) lower performance when including unplanned ICU admissions preceded by a theater visit. This decrease was particularly marked in hospital C, a dedicated center for cancer and renal services (including transplants). Notwithstanding the case-mix differences in comparison with the other three hospitals (*see* Table EA1 and Figure EB2), certain surgical

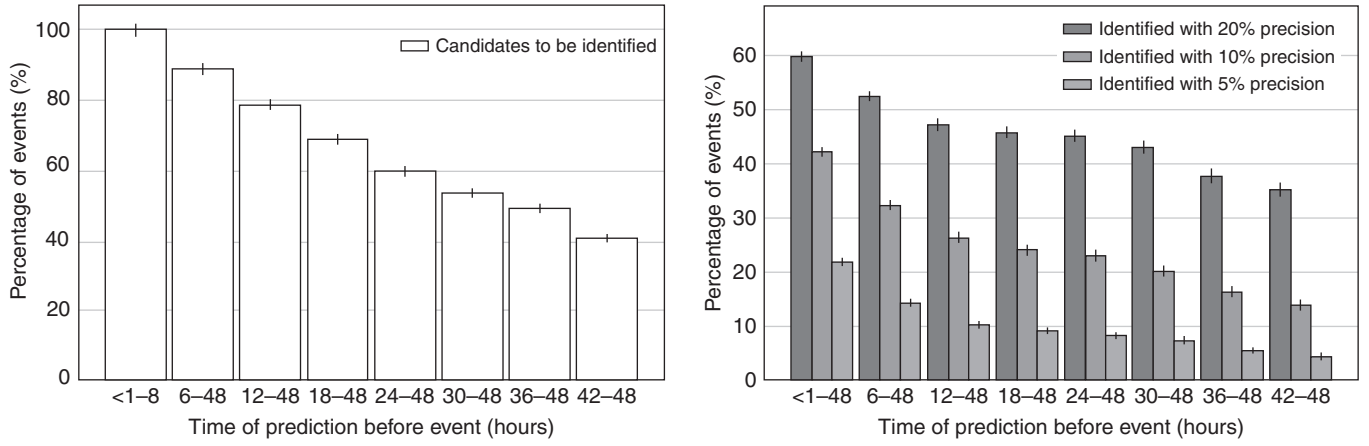

Figure 2. Patient admission level sensitivity: average proportion of (candidate) adverse events to be identified within each window (left); and the average proportion of adverse events identified ahead of time for Hospital-wide Alerting via Electronic Noticeboard at different precision levels (5%, 10%, and 20%) (right). The error bars denote 1 SD.

procedures are undertaken on physiologically stable patients, who are routinely transferred to the ICU postoperatively and coded as an unplanned ICU admission. This again demonstrates the importance of selecting the appropriate outcome when evaluating risk scoring systems.

There are limitations to using unplanned ICU admission and cardiac arrest as outcomes. These outcomes are affected by existing treatment-limitation plans and "do not attempt cardiopulmonary resuscitation" decisions. Electronic coding of these decisions varies between hospitals and is currently insufficiently robust for inclusion in a generalizable model. A recent systematic review found that ICU admission can be affected by clinicians' experience, the perception of benefit, and organizational factors (e.g., bed availability) (65). Training our model on retrospective data risks incorporating these potential "cultural biases." We sought to reduce bias (e.g., against older patients) by including a broad range of patient factors (comorbidities, frailty) in our model. Indeed, Figure EF3 shows that although, on average, patients aged over 80 years have a decreasing likelihood of either cardiac arrest or ICU transfer, there is wider variation in the overall predicted risk for each age value above 80 years.

To further evaluate HAVEN's predictive performance, we computed the percentage of adverse events identified ahead of time (Figure 2). We used a patient-level approach to determine the sensitivity of the model at different degrees of precision. As HAVEN was targeted at patients who deteriorate on general wards (rather than direct ICU admissions), we only included time periods after patients were transferred to a general ward. Our results therefore cannot be applied to patients who deteriorated in the emergency department. Despite the low prevalence of the outcome, the HAVEN model identified 42% of adverse events up to 48 hours in advance at an NNE of 10. Although nearly twice as good as the next best system (LAPS-2), seeing 10 patients to detect 1 would still create a significant workload. However, decision curve analysis (Figures EB4 and EB5) showed that HAVEN has higher net benefit than the next three highest scoring systems (including NEWS, in common use in the United Kingdom). Together, these findings suggest that implementing the HAVEN model should improve patient care.

Studies of EWSs and other risk scores for identifying deteriorating patients vary in the outcomes and statistical methods used to validate their performance, making comparisons difficult (22, 43, 45, 66). A large retrospective study of the Advanced Alert Monitor (AAM) score showed that the AAM score had a discrimination (ROC AUC) of 0.82 in comparison with discrimination of 0.79 and 0.76 for electronic CART and NEWS, respectively, for predicting unplanned ICU admissions within 12 hours (22). In contrast, the discrimination of HAVEN was 0.92 for predicting unplanned ICU admissions within 24 hours, outperforming the AAM score over a longer prediction horizon.

Conclusions

HAVEN performed significantly better than other published scores, such as NEWS and LAPS-2, when externally validated on an independent test set. Through the use of an ensemble of "weak learners" (gradient boosting decision trees) as our machine-learning algorithm, we were better able to model patients' physiological measurements in the context of their known comorbidities and frailty. We plan further external validation to ensure HAVEN model performance is sustained in other hospitals before a prospective evaluation on patient outcomes. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

References

1. Intensive Care National Audit and Research Centre. Unit mortality and acute hospital mortality of admissions direct from the ward to adult, general critical care units in England and Wales. 2014 [accessed 2021 Jun 23]. Available from: https://www.icnarc.org/DataServices/Attachments/Download/4e7031c4-f74a-e411-a65b-d48564544b14.
2. Barwise A, Thongprayoon C, Gajic O, Jensen J, Herasevich V, Pickering BW. Delayed rapid response team activation is associated with increased hospital mortality, morbidity, and length of stay in a tertiary care institution. *Crit Care Med* 2016;44:54–63.
3. Findlay GP, Shotton H, Kelly K, Mason M. National Confidential Enquiry into Patient Outcomes and Death. Time to Intervene? A review of patients who underwent cardiopulmonary resuscitation as a result of an in-hospital cardiorespiratory arrest. London, UK: National Confidential Enquiry into Patient Outcomes and Death; 2012.
4. Keogh B. Review into the quality of care and treatment provided by 14 hospital trusts in England: overview report. Leeds, UK: National Health Service; 2013.
5. Gerry S, Birks J, Bonnici T, Watkinson PJ, Kirtley S, Collins GS. Early warning scores for detecting deterioration in adult hospital patients: a systematic review protocol. *BMJ Open* 2017;7:e019268.
6. Smith GB, Prytherch DR, Jarvis S, Kovacs C, Meredith P, Schmidt PE, *et al.* A comparison of the ability of the physiologic components of medical emergency team criteria and the U.K. National early warning score to discriminate patients at risk of a range of adverse clinical outcomes. *Crit Care Med* 2016;44:2171–2181.
7. Pedersen NE, Rasmussen LS, Petersen JA, Gerds TA, Østergaard D, Lippert A. A critical assessment of early warning score records in 168,000 patients. *J Clin Monit Comput* 2018;32:109–116.
8. National Institute for Health and Care Excellence. Emergency and acute medical care in over 16s: service delivery and organization. London, UK: National Institute for Health and Care Excellence; 2018.
9. McGaughey J, Alderdice F, Fowler R, Kapila A, Mayhew A, Moutray M. Outreach and early warning systems (EWS) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. *Cochrane Database Syst Rev* 2007;3:CD005529.
10. Solomon RS, Corwin GS, Barclay DC, Quddusi SF, Dannenberg MD. Effectiveness of rapid response teams on rates of in-hospital cardiopulmonary arrest and mortality: a systematic review and meta-analysis. *J Hosp Med* 2016;11:438–445.
11. McNeill G, Bryden D. Do either early warning systems or emergency response teams improve hospital patient survival? A systematic review. *Resuscitation* 2013;84:1652–1667.
12. 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;47:49–55.
13. Redfern OC, Pimentel MAF, Prytherch D, Meredith P, Clifton DA, Tarassenko L, *et al.* Predicting in-hospital mortality and unanticipated admissions to the intensive care unit using routinely collected blood tests and vital signs: development and validation of a multivariable model. *Resuscitation* 2018;133:75–81.
14. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016;44:368–374.
15. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25:1337–1340.
16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.

17. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30–36.
18. Kang MA, Churpek MM, Zadravecz FJ, Adhikari R, Twu NM, Edelson DP. Real-time risk prediction on the wards: a feasibility study. *Crit Care Med* 2016;44:1468–1473.
19. Dummett BA, Adams C, Scruth E, Liu V, Guo M, Escobar GJ. Incorporating an early detection system into routine clinical practice in two community hospitals. *J Hosp Med* 2016;11:S25–S31.
20. Rojas JC, Carey KA, Edelson DP, Venable LR, Howell MD, Churpek MM. Predicting intensive care unit readmission with machine learning using electronic health record data. *Ann Am Thorac Soc* 2018;15:846–853.
21. Arnold J, Davis A, Fischhoff B, Yecies E, Grace J, Klobuka A, *et al.* Comparing the predictive ability of a commercial artificial intelligence early warning system with physician judgement for clinical deterioration in hospitalised general internal medicine patients: a prospective observational study. *BMJ Open* 2019;9:e032187.
22. Kipnis P, Turk BJ, Wulf DA, LaGuardia JC, Liu V, Churpek MM, *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.
23. Sendak MP, D'Arcy J, Kashyap S, Gao M, Nichols M, Corey K, *et al.* A path for translation of machine learning products into healthcare delivery. *EMJ Innov* [online ahead of print] 27 Jan 2020; DOI: 10.33590/emjinnov/19-00172.
24. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019;572:116–119.
25. O'Brien C, Goldstein BA, Shen Y, Phelan M, Lambert C, Bedoya AD, *et al.* Development, implementation, and evaluation of an in-hospital optimized early warning score for patient deterioration. *MDM Policy Pract* 2020;5:2381468319899663.
26. Rothman MJ. The emperor has no clothes. *Crit Care Med* 2019;47:129–130.
27. Paulson SS, Dummett BA, Green J, Scruth E, Reyes V, Escobar GJ. What do we do after the pilot is done? Implementation of a hospital early warning system at scale. *Jt Comm J Qual Patient Saf* 2020;46:207–216.
28. Finlay GD, Rothman MJ, Smith RA. Measuring the modified early warning score and the Rothman index: advantages of utilizing the electronic medical record in an early warning system. *J Hosp Med* 2014;9:116–119.
29. Bartkowiak B, Snyder AM, Benjamin A, Schneider A, Twu NM, Churpek MM, *et al.* Validating the electronic cardiac arrest risk triage (eCART) score for risk stratification of surgical inpatients in the postoperative setting: retrospective cohort study. *Ann Surg* 2019;269:1059–1063.
30. Lindsell CJ, Stead WW, Johnson KB. Action-informed artificial intelligence-matching the algorithm to the problem. *JAMA* 2020;323:2141–2142.
31. Adibi A, Sadatsafavi M, Ioannidis JPA. Validation and utility testing of clinical prediction models: time to change the approach. *JAMA* 2020;324:235–236.
32. Tarassenko L, Clifton DA, Pinsky MR, Hravnak MT, Woods JR, Watkinson PJ. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation* 2011;82:1013–1018.
33. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73.
34. Care Quality Commission. Inspection report: Oxford University Hospitals NHS Foundation Trust. 2019 [accessed 2020 May 20]. Available from: https://www.cqc.org.uk/provider/RTH/.
35. Malycha J, Bonnici T, Clifton DA, Ludbrook G, Young JD, Watkinson PJ. Patient centered variables with univariate associations with unplanned ICU admission: a systematic review. *BMC Med Inform Decis Mak* 2019;19:98.
36. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care* 1998;36:8–27.
37. Gilbert T, Neuburger J, Kraindler J, Keeble E, Smith P, Ariti C, *et al.* Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study. *Lancet* 2018;391:1775–1782.
38. Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation* 2016;102:1–5.
39. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
40. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, *et al.* xgboost: extreme gradient boosting. 2019 [accessed 2020 May 20]. Available from: https://xgboost.readthedocs.io/en/latest/ (accessed 20-05-2020).
41. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning: ICML '05, New York: ACM Press; 2005. pp. 625–632.
42. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432.
43. Churpek MM, Yuen TC, Winslow C, Robicsek AA, Meltzer DO, Gibbons RD, *et al.* Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med* 2014;190:649–655.
44. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Crit Care* 2015;19:285.
45. Gerry S, Bonnici T, Birks J, Kirtley S, Virdee PS, Watkinson PJ, *et al.* Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ* 2020;369:m1501.
46. Zhang Z, Rousson V, Lee W-C, Ferdynus C, Chen M, Qian X, *et al.*; AME Big-Data Clinical Trial Collaborative Group. Decision curve analysis: a technical note. *Ann Transl Med* 2018;6:308.
47. Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, *et al.* Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol* 2018;74:796–804.
48. Brown M. rmda: risk model decision analysis. 2018 [accessed 2020 May 20]. Available from: https://cran.r-project.org/web/packages/rmda/readme/README.html
49. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Stat Sci* 1996;11:189–228.
50. Lundberg S, Lee S-I. A unified approach to interpreting model predictions [preprint]. arXiv; 2017 [accessed 2017 Nov 25]. Available from: https://arxiv.org/abs/1705.07874.
51. Watkinson PJ, Pimentel MAF, Clifton DA, Tarassenko L. Manual centile-based early warning scores derived from statistical distributions of observational vital-sign data. *Resuscitation* 2018;129:55–60.
52. Gardner-Thorpe J, Love N, Wrightson J, Walsh S, Keeling N. The value of Modified Early Warning Score (MEWS) in surgical in-patients: a prospective observational study. *Ann R Coll Surg Engl* 2006;88:571–575.
53. Paterson R, MacLeod DC, Thetford D, Beattie A, Graham C, Lam S, *et al.* Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. *Clin Med (Lond)* 2006;6:281–284.
54. 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;84:465–470.
55. Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs. *Crit Care Med* 2012;40:2102–2108.
56. Churpek MM, Yuen TC, Park SY, Gibbons R, Edelson DP. Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards. *Crit Care Med* 2014;42:841–848.
57. Escobar GJ, Gardner MN, Greene JD, Draper D, Kipnis P. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Med Care* 2013;51:446–453.
58. Rothman MJ, Rothman SI, Beals J IV. Development and validation of a continuous measure of patient condition using the electronic medical record. *J Biomed Inform* 2013;46:837–848.
59. Collins GS, Reitsma JB, Altman DG, Moons KGM; TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur Urol* 2015;67:1142–1151.
60. Prytherch DR, Smith GB, Schmidt PE, Featherstone PI. ViEWS: towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation* 2010;81:932–937.
61. Churpek MM, Yuen TC, Edelson DP. Predicting clinical deterioration in the hospital: the impact of outcome selection. *Resuscitation* 2013;84:564–568.
62. Sandroni C, D'Arrigo S, Antonelli M. Rapid response systems: are they really effective? *Crit Care* 2015;19:104.

63. Hogan H, Zipfel R, Neuburger J, Hutchings A, Darzi A, Black N. Avoidability of hospital deaths and association with hospital-wide mortality ratios: retrospective case record review and regression analysis. *BMJ* 2015; 351:h3239.
64. Chukwusa E, Verne J, Polato G, Taylor R, J Higginson I, Gao W. Urban and rural differences in geographical accessibility to inpatient palliative and end-of-life (PEoLC) facilities and place of death: a national population-based study in England, UK. *Int J Health Geogr* 2019;18:8.
65. Gopalan PD, Pershad S. Decision-making in ICU - a systematic review of factors considered important by ICU clinician decision makers with regard to ICU triage decisions. *J Crit Care* 2019;50:99–110.
66. 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.