OPEN

A Trend-Based Early Warning Score Can Be Implemented in a Hospital Electronic Medical Record to Effectively Predict Inpatient Deterioration

David Bell, MS, MBBS[1,2]
John Baker, BCompSc[1]
Chris Williams, BSc[1]
Levi Bassin, BSc, MBBS, PhD[1,2]

OBJECTIVES: To determine whether a statistically derived, trend-based, deterioration index is superior to other early warning scores at predicting adverse events and whether it can be integrated into an electronic medical record to enable real-time alerts.

DESIGN: Forty-three variables and their trends from cases and controls were used to develop a logistic model and deterioration index to predict patient deterioration greater than or equal to 1 hour prior to an adverse event.

SETTING: Two large Australian teaching hospitals.

PATIENTS: Cases were considered as patients who suffered adverse events (unexpected death, unplanned ICU transfer, urgent surgery, and rapid-response alert) between August 1, 2016, and April 1, 2019.

INTERVENTIONS: The logistic model and deterioration index were tested on historical data and then integrated into an electronic medical record for a 6-month prospective "silent" validation.

MEASUREMENTS AND MAIN RESULTS: Data were acquired from 258,732 admissions. There were 8,002 adverse events. The addition of vital sign and laboratory trend values to the logistic model increased the area under the curve from 0.84 to 0.89 and the sensitivity to predict an adverse event 1–48 hours prior from 0.35 to 0.41. A 48-hour simulation showed that the logistic model had a higher area under the curve than the Modified Early Warning Score and National Early Warning Score (0.87 vs 0.74 vs 0.71). During the silently run prospective trial, the sensitivity of the deterioration index to detect adverse event any time prior to the adverse event was 0.474, 0.369 1 hour prior, and 0.327 4 hours prior, with a specificity of 0.972.

CONCLUSIONS: A deterioration prediction model was developed using patient demographics, ward-based observations, laboratory values, and their trends. The model's outputs were converted to a deterioration index that was successfully integrated into a live hospital electronic medical record. The sensitivity and specificity of the tool to detect inpatient deterioration were superior to traditional early warning scores.

KEY WORDS: early warning score; implementation; medical informatics; patient deterioration; patient monitoring

Despite widespread adoption of Early Warning Score (EWS) systems, such as the Modified Early Warning Score (MEWS) (1), the National Early Warning Score (NEWS) (2) and the Between the Flags (BTF)

DOI: 10.1097/CCM.0000000000005064

protocols (3), delayed detection of inpatient deterioration still occurs in close to 50% of cases (4). Current escalation of care triggers identify patients who are already unwell, but they do a poor job at identifying patients at risk of deterioration.

Most EWS are triggered by discrete values or aggregate scores, which are reflective of a clinical picture at a single point in time (1–3). Furthermore, unlike a clinician, they do not account for patient demographics, laboratory values, or trends in values (5). For example, BTF, a system widely implemented in Australian hospitals, is a two-tiered system in which specific observation values trigger either a yellow flag or red flag. A yellow flag requires nursing staff to inform a doctor of the result, whereas a red flag requires immediate medical review. An example of a yellow flag is a respiration rate greater than 25. An example of a red flag is a respiration rate greater than 30.

We hypothesized that it was possible to develop a predictive model with demographic data, vital signs, laboratory values, and their trends and use this model to create a deterioration index (DI). This would be updated in real time to flag deterioration and facilitate earlier intervention. Our aim was to predict adverse events (AEs) at least 1 hour prior to the events themselves, with a low false positive rate to minimize alert fatigue. We further sought to implement a system that would not require manual data entry and would therefore integrate seamlessly in an acute inpatient setting.

METHODS

Data Collection

Following multisite ethics approval (2019/ETH00557), de-identified data were extracted for patients admitted to two Australian teaching hospitals between August 1, 2016, and April 1, 2019. All patients over 16 were included excepting those in the ICU, palliative care, or the operating theatre environment. Static variables collected included sex, age, and whether or not the patient had undergone surgery. Time-varying variables collected included vital signs (heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, oxygen saturation, supplemental oxygen, and conscious state) and laboratory values (hemoglobin, WBC count, urea, and estimated glomerular filtration rate). AEs were defined as unplanned inpatient death, medical emergency team (MET) call, unplanned admission to ICU, or unplanned return to the operating theatre. MET calls triggered by bradycardia, hypopnea, or hypertension were excluded as outcomes.

MEWS, NEWS, and Between the Flags

MEWS and NEWS were calculated to serve as a comparison to the developed model. Cutoff points to trigger alerts were set at 4 and 7, respectively, as this resulted in specificity as close to 0.98 as practical for effective comparison (0.972 and 0.986, respectively). The results for MEWS and NEWS are described in **Supplement e1** (http://links.lww.com/CCM/G457). Yellow flag alerts were also calculated to compare the predictive capability of the BTF system.

Data Engineering and Statistical Analysis

All data engineering, model generation and statistical analysis were performed in R Studio (RStudio Team, Integrated Development for R. Boston, MA; http://www.rstudio.com/) (6). For cases, the time was said to be equal to 0 at the time of the AE, and for controls, time was said to be 0 at a randomly selected timepoint. Ward and laboratory measurements were labeled in hours prior to the reference time stamp 0. For example, a blood pressure measured 1.5 hours prior to an AE was time stamped at $T_{-1.5}$. Measurements occurring up to 7 days prior an AE (T_{-168}) were recorded.

Data were then grouped into intervals based on hours prior to T_0: 0–1, 1–4, 4–8, 8–12, 12–18, 18–24, 24–36, 36–48, 48–72, 72–96, 96–120, 120–144, and 144–168. If multiple measurements were recorded in the same time interval, the value at time point closest to T_0 was used as the representative value for that interval. If no measurement was recorded in the interval, the most recently recorded preceding value was persisted forward until a new value was recorded. Where no measurement was recorded, and no preceding value was available (e.g., in the case of a patient admitted 5 d prior to an AE and therefore with no value at T_{-168}), cells were populated with the median value matched for age (≥ 65 or < 65) and sex. Baseline values were defined by the earliest available measurement for each patient. Three trends were calculated by comparing each variable value to the baseline value (trend $[TR]_{BL}$), the previously measured value (TR_1), and the value measured two episodes prior (TR_2).

Logistic Regression Analysis

Logistic regression was performed with the outcome variable defined by an AE. Sixty-five variables were modified to produce a linear relationship between the variable and the log-odds of an AE. This is described in more detail in **Supplement e2** (http://links.lww.com/CCM/G457). Data were then divided into training and test sets in a 2:1 ratio using a random number generator. Each variable was assessed for statistical significance. Those that were not statistically significant were discarded, leaving 43 variables. The model was then revalidated in a similar fashion. Statistical significance was set at a *p* value of less than or equal to 0.05. The area under the curve (AUC) was calculated for the receiver operating characteristic (ROC) curve and displayed with the 95% CIs.

Selecting the Ideal Logistic Model

Three different logistic models were built to optimize for deterioration prediction at three different time points prior to the AE. Model 1 used all data up to and including data recorded at T_0, model 2 used all data up to and including T_{-1}, and model 3 used all data up to and including T_{-4}. Each model was then tested at three different time points prior to an AE ($T_0/T_{-1}/T_{-4}$) to assess which model would be the most effective in practice (**Supplement e3**, http://links.lww.com/CCM/G457). Model 2 (using T_{-1} data) was selected as the optimal "the logistic model," as it had the best balance of early and late prediction based on the AUC and sensitivities. The coefficients for the final model are included in **Supplement e4** (http://links.lww.com/CCM/G457).

Simulation

To mimic a clinical scenario, the model was run at each time bracket in sequence from T_{-48} to T_0 (0–1, 1–4, 4–8, 8–12, 12–18, 18–24, 24–36, and 36–48 hr) on the holdout dataset. For each time bracket, a simulation was run with 100 different alert triggers (probabilities of AE) from 0 to 1 in 0.01 increments. The absolute number of alerts was recorded for each trigger at each time bracket to yield the sensitivity and specificity. The ROC was then obtained by plotting the false positive rate (1–specificity) against the true positive rate (sensitivity). This method of simulation was applied to the MEWS and NEWS systems to allow for comparison (**Fig. 1**). The simulation was also run for the BTF yellow flag alert to determine a comparison sensitivity and specificity.

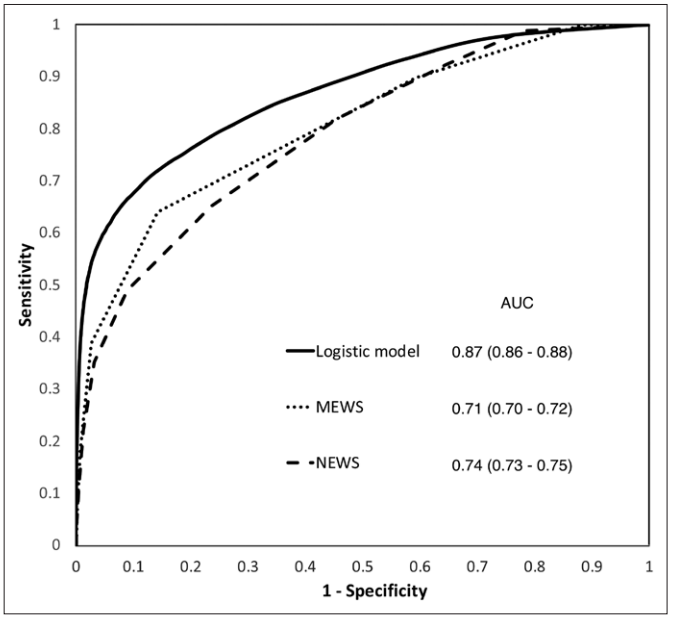

Figure 1. The receiver operating characteristic curve comparing the logistic model to the Modified Early Warning Score (MEWS) and National Early Warning Score (NEWS) scores in the 48 hr prior to an adverse event. Simulations were run on a holdout sample of 86,989 patients. AUC = area under the curve.

Deterioration Index

A DI was created to convert the probability of deterioration to a number between 1 and 10, linearize the positive predictive value (PPV), and facilitate easy adjustment of model sensitivity and specificity following implementation. The PPV was linearized to create an easily interpretable and clinically relevant score. In order to create the DI, the "P(AE)" was set to the power of an adjustable exponent, the deterioration coefficient (DC) and then multiplied by 10.

Figure 2 demonstrates the relationship between the DI and the total true and false positive rate, as well as the PPV.

Silent Trial

A prospective "silent trial" was run on all ward-based inpatients over 16, not in ICU or on a palliative ward, between October 2019 to April 2020 at one of the involved hospitals (hospital 1). In this period, the logistic model was integrated with the electronic medical record (EMR) (SanCare; Adventist HealthCare, Wahroonga, NSW, Australia) and alerts were triggered after real-time analysis, but they were not sent to clinical staff.

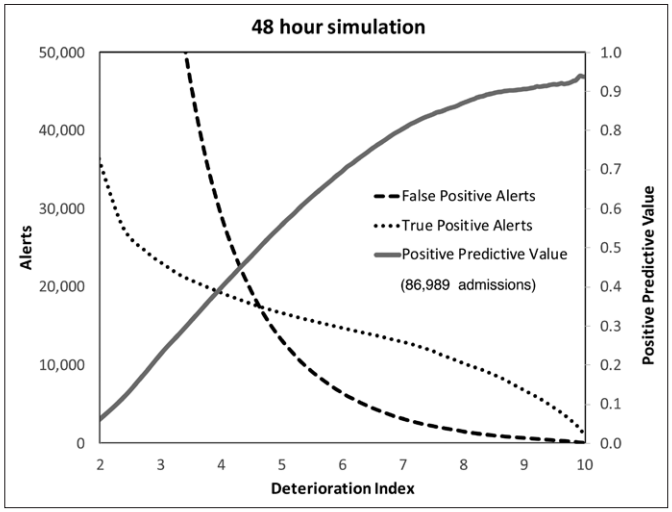

Figure 2. Deterioration index versus positive predictive value, true positive, and false positive alerts for the 48 hr prior to an adverse event. The simulation was performed on a holdout set of 86,989 admissions.

RESULTS

The authors hypothesized that rapid responses triggered only by bradycardia, hypopnea, and hypertension were frequently not preceded by a gradual deterioration in trends. As a result, MET calls triggered by single episodes of hypertension, bradycardia, or hypopnea in isolation were not included as AEs. Notably, only patients who suffered a MET call but for whom all other values during the MET call were within normal limits were excluded. No patients who had an admission to ICU, return to theatre, or death were excluded. Furthermore, the authors wish to stress that these patients were still included in the silent trial analysis, but we only excluded for the purposes of model generation and validation. Five-hundred thirty MET calls from 473 patients were excluded from model development and testing on this basis and 3,462 MET calls from 3,201 patients that were included. As summarized in **Supplement e5** (http://links.lww.com/CCM/G457), MET calls for isolated hypertension, bradycardia, or hypopnea were indeed far less likely to result in death, unplanned ICU admission, or unplanned surgery.

Data consisted of 258,732 admissions in total between August 1, 2016, and April 1, 2019. Eight-thousand two eligible AEs occurred across 5,885 individual patient admission. Each AE was considered as a single case. Two-hundred fifty-two–thousand eight-hundred forty-seven admissions occurred without an AE, and these were set as controls. Cases consisted of 1,313 deaths (16.4%), 2,911 unplanned admissions to ICU (36.4%), 3,462 MET calls (43.3%), and 316 transfers for unplanned surgery (3.9%). The median age for cases was 75, and for controls, 64. Controls were more likely to be female (51.2% vs 48.2%). Patient demographics are described in Supplement e5 (http://links.lww.com/CCM/G457).

Addition of Patient Demographics, Laboratory Results, and Trends

The model's predictive capability was significantly enhanced by the addition of patient demographics, laboratory results, and trends. **Table 1** demonstrates the incremental predictive value these parameters add when tested at T_{-1}. Notably, vital sign trends were a more powerful predictor of an AE (AUC of 0.83 ± 0.01) than vital signs alone (0.79 ± 0.01).

TABLE 1.

The Impact of Additional Variables on the Logistic Model's Ability to Predict Adverse Events More Than 1 Hour Prior

Model Characteristics	Area Under the Curve	Sensitivity	Specificity
Current vital signs only	0.79 (0.78–0.8)	0.21	0.99
+ Demographics	0.83 (0.82–0.84)	0.31	0.99
+ Laboratory values	0.85 (0.84–0.86)	0.35	0.99
+ Vital sign trends	0.88 (0.87–0.89)	0.40	0.99
+ Laboratory values trends	0.89 (0.88–0.89)	0.41	0.99
Vital sign and laboratory trends only	0.83 (0.82–0.84)	0.31	0.99

TABLE 2.
The Predictive Capacity of Each System in a 48-Hour Simulation Prior to an Adverse Event

Deterioration Predictor	Area Under the Curve	Sensitivity	Specificity	Positive Predictive Value
Logistic model	0.87 (0.86–0.88)	0.43	0.98	0.65
National Early Warning Score	0.74 (0.73–0.75)	0.22	0.987	0.35
Modified Early Warning Score	0.71 (0.70–0.72)	0.39	0.972	0.31
Yellow flag	Not available	0.36	0.69	0.03

Simulation

To mimic a clinical scenario, the model was run through a simulation. The logistic model, MEWS and NEWS were all corrected for a similar specificity (0.98). The results of the simulation (**Table 2**) demonstrate that both the sensitivity and AUC for the logistic model are superior to those of MEWS and NEWS. The yellow flag alerts were noted to have the lowest specificity (0.72) and incur the greatest number of false positives. During this simulation, the logistic model also had a higher PPV at least 4 hours prior to an event (0.6) than NEWS (0.293), MEWS (0.255), or yellow flag alerts (0.071) and is described in more detail in **Supplement e6** (http://links.lww.com/CCM/G457).

Alert Timing

Figure 3 in **supplement e7** (http://links.lww.com/CCM/G457) shows the sensitivity for the 48-hour simulation and compares that to the sensitivity based on only the most recent 24 hours prior to an AE and then only the most recent 12 hours prior to an AE. The sensitivities were 0.44, 0.43, and 0.41, respectively, implying that imputed data has little effect on the results and that there are very few patients generating alerts at 48 hours out that did not continue to generate ongoing alerts as they approached an AE.

Figure 4 in **supplement e8** (http://links.lww.com/CCM/G457) shows how the sensitivity of alerts varied approaching T = 0. The mean alert time prior to each event was calculated through the simulation by presuming alerts occurred at the latest possible time within each bucket. For example, if an alert was first sent for a patient in the 1–4 bucket, it was presumed that an alert was sent 1 hour prior to an event. The mean first alert time for an amber alert (defined by a DI > 6) was 21.98 hours prior to an AE. The mean first alert time for a red alert (defined by a DI > 8) was 17.05 hours prior to an AE. The mean first alert time for a MEWS of 4 and NEWS of 7 was 8.73 and 8.24 hours prior to an AE, respectively.

Silent Trial

The software was trialed silently, in real time, over a 6-month period. During the prospective 6-month silent trial period at hospital 1, there were 450 AEs from 28,533 admissions (1.58%). Forty-three deaths were excluded from the analysis due to documented evidence of palliation. A total of 407 outcomes were therefore included in the analysis, identified by patient death (103) and all MET calls (304). Reliably timed data on unplanned admissions to ICU was not available.

Over this period, a total of 1,106 patients triggered an "Amber alert," and 639 patients triggered a "Red alert." In total, 1,343 patients triggered at least one alert of any kind. One-hundred ninety-three of 407 patients who experienced an AE triggered an alert some time before that outcome. A total of 150 alerts were first triggered more than 1 hour prior to an AE and a total of 133 alerts were first triggered more than 4 hours prior to an event. Forty-three alerts were first triggered between 0 and 1 hours prior to an AE. One-hundred four patients triggered an alert at the same time as the outcome with no alert prior. One-hundred ten patients experienced an outcome without triggering an alert any time prior to or at the time of the outcome.

There was a total of 120 false positive red alerts and 657 false positive amber alerts. At 4 hours prior to an AE, the sensitivity for any alert was 0.327, and at 1 hour prior, it was 0.369. The specificity for a red alert was 0.996 and for an amber alert 0.977. **Table 3** summarizes the major silent trial results.

TABLE 3.
Alert Sensitivities and Specificities for the Real Time Silent Trial of the Deterioration Index

Alert Type	Sensitivity	Specificity
Amber alert (DI 6–7)	0.388	0.977
Red alert (DI 8–10)	0.268	0.996
Any alert (DI ≥ 6) ≥ T_0	0.754	0.972
Any alert (DI ≥ 6) > T_0	0.474	0.972
Any alert (DI ≥ 6) > T_{-1}	0.369	0.972
Any alert (DI ≥ 6) > T_{-4}	0.327	0.972

DI = deterioration index, T = the reference time of an adverse event/control (hr) prior to the event.

DISCUSSION

Creating a deterioration tool using trend data is a more complex undertaking than developing a model with single point in time variables alone. It requires significantly more data and more complex software systems to store and manipulate the results. We have shown that the addition of demographics, laboratory values, and trends increased the AUC and sensitivity for detecting an AE from 0.79 and 0.21, respectively, in a stepwise fashion to 0.89 and 0.41 over a model with vital signs alone.

Implementation practice is just as important as effective model generation. Collaboration with nursing staff to best understand how the output of the logistic model (a probability between 0 and 1) could be most effectively translated into a meaningful alert that would augment, rather than interfere with patient workflow, drove most of the implementation decisions in this article. Feedback from clinical staff drove the decision to implement a two-tiered (amber and red) alert system.

Very deterioration prediction models have been externally validated into a real-time production EMR, as presented in this article. Following integration into the production EMR, an extended period of prospective, silent, validation, was vital to ensure a seamless transition into sending live alerts and to validate that the results generated on historical data were also generated prospectively. The DI has now been launched in a live clinical trial, with alerts being displayed in the EMR and sent to clinical staff.

This study has limitations relating to generalizability. The model was developed using data from two Australian hospitals and may not be generalizable to other jurisdictions and patient populations. Furthermore, the silent trial took place at a private hospital, with a lower burden of acuity and therefore lower event rate. Finally, implementation of this model requires an EMR that can stream data in real time, which is not yet a universal capability. This study also has limitations relating to deficiencies in the data used. Patients were deemed to be not palliative if they were not residing on the palliative ward, meaning that some palliative patients may have been included. Another complicating factor relates to the fact that the time of death entered into the EMR occurs inconsistently and may have occurred after the fact resulting in a misleading positive predictive bias. The authors hope this is limited by presenting predictions that occurred at least 1 hour prior to the AE. There are further limitations related to study design. The use of PPV may generate a favorable bias. That is to say, should this system be successful at reducing AEs in clinical practice, the PPV should actually be lower than in the simulation or silent trial as AEs should be prevented. Also, MEWS and NEWS were used as comparators, but the authors note that they were built to detect outcomes that differ slightly than those used to defined AEs in this study. Finally, although the predictive capability of the model appears promising, this study does not analyze the effect this has on clinical outcomes. This will be the primary focus of the live trial. The product described in this text is still investigational and not U.S. Food and Drug Administration approved.

CONCLUSIONS

A model to predict inpatient deterioration was developed using routinely measured clinical variables and their associated trends. The model's outputs were converted to a DI that was successfully integrated into a live EMR environment. The sensitivity and specificity of the tool to detect inpatient deterioration were shown to be superior to traditional EWSs. Notably, the addition of demographic data, laboratory values, and trends improved the predictive capability of the model.

1 *Department of Clinical Informatics, Sydney Adventist Hospital, Sydney, NSW, Australia.*
2 *Department of Cardiothoracic Surgery, Royal North Shore Hospital, Sydney, NSW, Australia.*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Drs. Bell and Bassin disclosed that they are the co-owners of the Australian company Ainsoff Pty, which sold the model developed in this article that forms the basis for an early warning system. Dr. Bell disclosed the off-label product use of a real-time early warning system using trend analysis. Drs. Baker and Williams disclosed that they are employees of Adventist HealthCare trustee for Sydney Adventist Hospital.

For information regarding this article, E-mail: davidjbell218@gmail.com

REFERENCES

1. Subbe CP, Kruger M, Rutherford P, et al: Validation of a modified early warning score in medical admissions. *QJM* 2001; 94:521–526
2. Royal College of Physicians: National Early Warning Score (NEWS): Standardising the Assessment of Acute-Illness Severity in the NHS. Report of a Working Party. London, United Kingdom, Royal College of Physicians, 2012
3. Clinical Excellence Commission NSW: Recognition and Management of Patients Who Are Clinically Deteriorating (PD2013_049). 2013. Available at: https://www1.health.nsw.gov.au/pds/ActivePDSDocuments/PD2020_018.pdf. Accessed October 1, 2020
4. Jones D: The epidemiology of adult rapid response team patients in Australia. *Anaesth Intensive Care* 2014; 42:213–219
5. Churpek MM, Adhikari R, Edelson DP: The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation* 2016; 102:1–5
6. R Studio Team: R Studio. 2015. Available at: http://www.rstudio.com/. Accessed October 1, 2020