Physiological Measurement

IPEM
Institute of Physics and Engineering in Medicine

PAPER • OPEN ACCESS

Technical considerations for evaluating clinical prediction indices: a case study for predicting code blue events with MEWS

To cite this article: Kais Gadhoumi *et al* 2021 *Physiol. Meas.* **42** 055005

IOP Publishing Physiol. Meas. 42 (2021) 055005 https://doi.org/10.1088/1361-6579/abfbb9

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
17 December 2020

REVISED
25 March 2021

ACCEPTED FOR PUBLICATION
26 April 2021

PUBLISHED
17 June 2021

PAPER

Technical considerations for evaluating clinical prediction indices: a case study for predicting code blue events with MEWS

Kais Gadhoumi[1], Alex Beltran[2], Christopher G Scully[3], Ran Xiao[1], David O Nahmias[3] and Xiao Hu[1]

[1] School of Nursing, Duke University, Durham, NC, United States of America
[2] Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, CA, United States of America
[3] Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, MD, United States of America

E-mail: xiao.hu@duke.edu

Keywords: early warning score, clinical deterioration, predictive value of tests, vital signs, clinical alarms, performance evaluation

Supplementary material for this article is available online

Abstract

Objective. There have been many efforts to develop tools predictive of health deterioration in hospitalized patients, but comprehensive evaluation of their predictive ability is often lacking to guide implementation in clinical practice. In this work, we propose new techniques and metrics for evaluating the performance of predictive alert algorithms and illustrate the advantage of capturing the timeliness and the clinical burden of alerts through the example of the modified early warning score (MEWS) applied to the prediction of in-hospital code blue events. *Approach*. Different implementations of MEWS were calculated from available physiological parameter measurements collected from the electronic health records of ICU adult patients. The performance of MEWS was evaluated using conventional and a set of non-conventional metrics and approaches that take into account the timeliness and practicality of alarms as well as the false alarm burden. *Main results*. MEWS calculated using the worst-case measurement (i.e. values scoring 3 points in the MEWS definition) over 2 h intervals significantly reduced the false alarm rate by over 50% (from 0.19/h to 0.08/h) while maintaining similar sensitivity levels as MEWS calculated from raw measurements (~80%). By considering a prediction horizon of 12 h preceding a code blue event, a significant improvement in the specificity (~60%), the precision (~155%), and the work-up to detection ratio (~50%) could be achieved, at the cost of a relatively marginal decrease in sensitivity (~10%). *Significance*. Performance aspects pertaining to the timeliness and burden of alarms can aid in understanding the potential utility of a predictive alarm algorithm in clinical settings.

1. Introduction

Early warning systems (EWS) are tools that warn about physiological instabilities in patients at risk of deterioration. They can play a crucial role in healthcare by enabling early and rapid intervention to help prevent in-hospital all-cause catastrophic events. Emerging tools are based on complex algorithms that use statistical and machine learning techniques to identify precursors of adverse events in single or multimodal physiological variables available at the bedside. However, the lack of comprehensive validation of many such tools to inform their implementations is one of the reasons that are responsible for their limited adoption in clinical practice (Damen *et al* 2016, Linnen *et al* 2019).

Methodological validation of EWSs is key to understanding their performance to predict clinical events. Conventional metrics typically used to evaluate the performance of EWSs are often limited to measures of sensitivity, specificity and discriminability (through concordance statistics), and often they are not evaluated in a way that considers the dynamic nature of a score from continuous vital sign data. These metrics are useful to

Table 1. Modified early warning score.

MEWS	3	2	1	0	1	2	3
Systolic blood pressure (mmHg)	<70	71–80	81–100	101–199		⩾200	
Heart rate (bpm)		<40	41–50	51–100	101–110	111–129	⩾130
Respiratory rate (bpm)		<9		9–14	15–20	21–29	⩾30
Temperature (°C)		<35		35–38.4		⩾38.5	
AVPU score[a]				Alert	Reacting to voice	Reacting to pain	Unresponsive

[a] AVPU: A, alert; V, reacting to voice; P: reacting to pain; U, unresponsive.

evaluate the classification power of an EWS but may not describe important aspects of an EWS. Questions about the frequency of high scores and the burden from score-based alarms, and about the extent, the pattern, and the time frame over which EWS changes are observed before an adverse event cannot be answered by examining conventional metrics alone. For example, an EWS that often warns about an impending cardiac arrest too early (e.g. a week) or too late (e.g. seconds) may have limited clinical utility. Yet, depending on the study design, such warnings could be considered true predictions. Accuracy and timeliness of EWS changes are important characteristics of performance to help understand the potential clinical utility.

In this study, we consider performance metrics for clinical prediction indices and examine the types of information they can provide. To focus on developing the approaches for validating EWS that are agnostic to algorithms driving EWS, we conduct a case study where a simple modified early warning score (MEWS) was implemented to predict code blue events in ICU patients. MEWS was proposed as a screening tool to identify inpatients at risk for deterioration and to trigger early evaluation and transfer to step-down or intensive care (Subbe *et al* 2001). Views on the clinical usefulness of MEWS remain rather controversial despite its common use (Gao *et al* 2007, Alam *et al* 2014). The goal of this study is not to prove or disprove the predictive power of MEWS in ICU patients, since MEWS was designed and has generally been evaluated for identifying patient deterioration in broad hospital populations and not a predictive score for cardiac arrest in the ICU (Morgan and Wright 2007). Rather, our goal is to explore analysis methods for these types of scores, and we use MEWS as an example due to its simplicity to calculate and frequent appearance in research studies evaluating its performance.

2. Materials and methods

2.1. Data

Demographics and vital signs measurements were extracted from the electronic health record (EHR) of patients hospitalized between 1 March, 2013 and 31 December, 2017 at the University of California San Francisco (UCSF) Medical Center. Patients aged 18 years and older who were admitted in the intensive care unit (ICU) without a do-not-resuscitate order were included. For each patient, we collected the age, gender, and measurements of heart rate (HR), systolic blood pressure (SBP), respiratory rate (RR), temperature (Temp), and Glasgow Coma Score (GCS). The study was approved for research investigation by the UCSF institutional review board.

Patients were split into case and control groups. Case patients ($n = 283$ of 3410; 8.3%) were defined as those with at least one in-hospital code blue event, including cardiopulmonary arrest (182 cases), acute respiratory compromise (34 cases) and other medical emergencies (67 cases), as documented and confirmed by the code blue committee of the medical center. In order to control for any increased risk of another code blue event following a first occurrence, only data recorded between the time of ICU admission and the time of the first code blue event were retained for analysis. Control patients ($n = 3127$ of 3410; 91.7%) were defined as those who did not experience a code blue event during their stay. Data recorded between the time of ICU admission and discharge of control patients were extracted for analysis. The median length of ICU stay was 86.3 h in the case group (IQR = 245.9 h) and 160.9 h in the control group (IQR = 193.2 h).

Vital signs were available on an irregular time interval. On average, a new vital sign (HR, SBP, RR, Temp) was measured every 0.6 h in the case group and every 1.7 h in the control group. GCS was less frequently measured than vital signs.

2.2. MEWS calculation

MEWS is derived using a set of point assignment rules applied to physiological parameter measurements as shown in table 1 (Subbe *et al* 2001) (some institutions may implement modified versions of these rules and/or use other/different physiological parameters such as urine output and oxygen saturation). The AVPU (A, alert; V, reacting to voice; P, reacting to pain; U, unresponsive) scale used in MEWS corresponds to distinct GCS

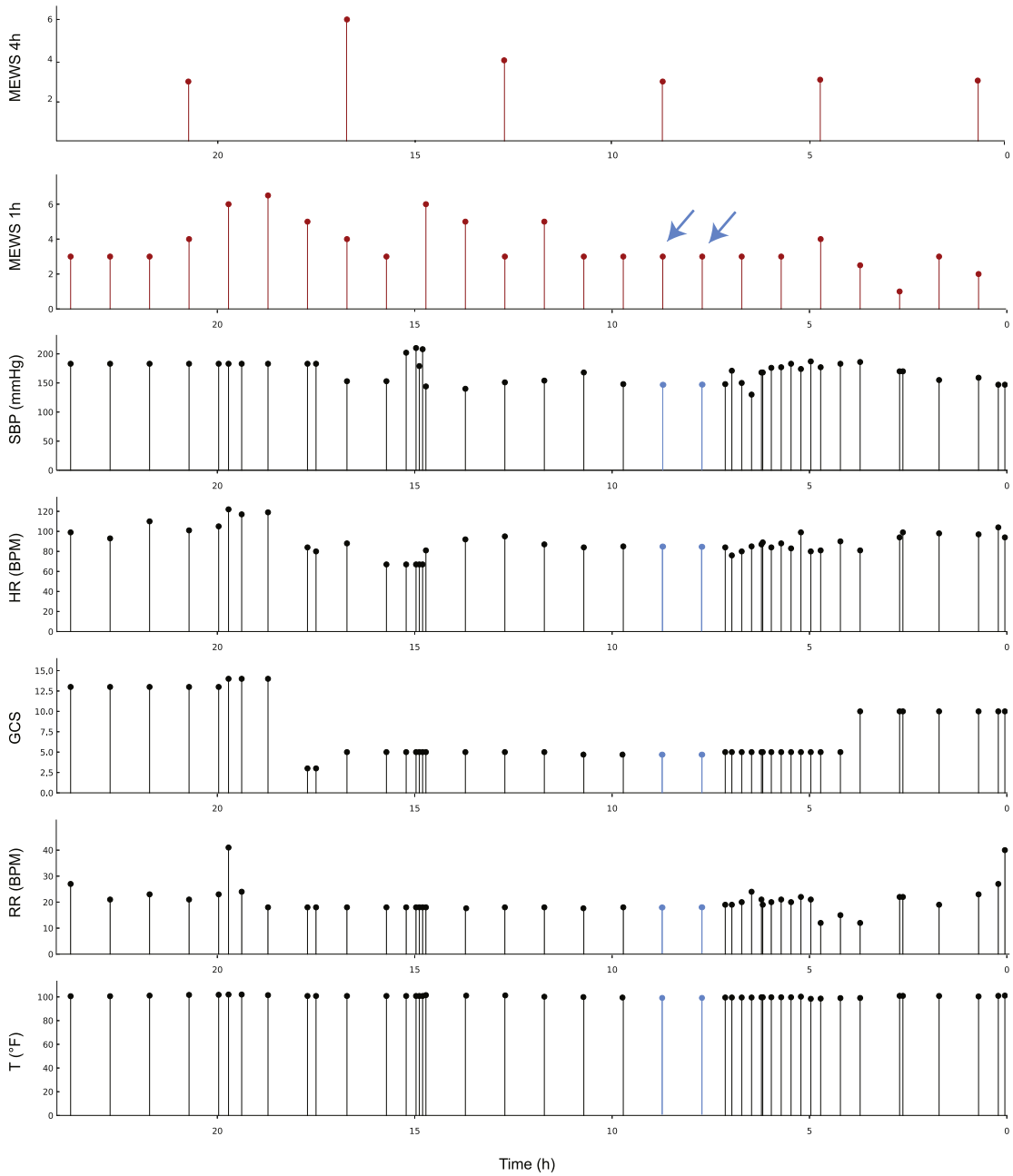

Figure 1. Derivation of a regularly sampled MEWS from physiological parameters. First and second top panels illustrate MEWS values calculated in one patient's 24 h recording, using $T_{MEWS} = 4$ h and $T_{MEWS} = 1$ h, respectively. MEWS was calculated using the median of physiological measurements within T_{MEWS}. Missing measurements were imputed to calculate MEWS at 1 h intervals (blue arrows indicate two MEWS values calculated from imputed physiological values, derived by carrying forward the last available measurement, indicated in blue lines).

ranges with overlap between some ranges (Kelly *et al* 2004, Romanelli and Farrell 2020). To derive a one-to-one correspondence between both scales, we adopted the following mapping: Alert = GCS 14–15; Reacting to voice = GCS 10–13; Reacting to pain = GCS 4–9; Unresponsive = GCS 3.

When EHR is used as the source of physiological parameter measurements, evaluating MEWS at regular intervals may be challenging since only a subset of measurements might be available at any sampling time point. Two approaches can generally be used to calculate MEWS. In a first approach, scores are calculated at a prescribed regular time interval and missing measurements are imputed. In a second approach, scores are calculated only when a new measurement of one or more parameters is available.

The details of how MEWS is calculated from irregularly sampled physiological parameters and how missing values are imputed is seldom described in the literature (Subbe *et al* 2001, 2003, Churpek *et al* 2012, Cooksley *et al* 2012, Fullerton *et al* 2012, Drower *et al* 2013, van Rooijen *et al* 2013, Bulut *et al* 2014, Kim *et al* 2015, Mathukia *et al* 2015, Kruisselbrink *et al* 2016, Jayasundera *et al* 2018, Al-Kalaldeh *et al* 2019). A straightforward

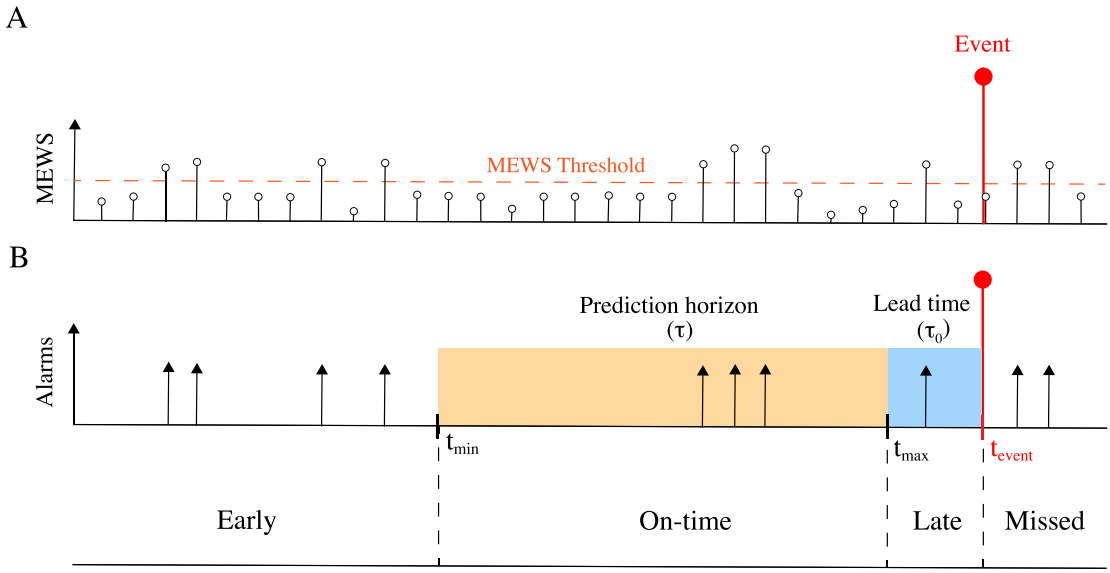

Figure 2. Definition of MEWS alarms, prediction horizon $\tau = [t_{\min}\, t_{\max}]$ and lead time $\tau_0 = [t_{\max}\, t_{event}]$. An alarm is raised each time MEWS exceeds a threshold (A). t_{event} is the time of a clinical event, e.g. a cardiac arrest. Alarms are considered early, on-time, late, and missed if they respectively occur before the start of prediction horizon, within the prediction horizon, within the lead time, and after the onset of the clinical event. Early, late and missed alarms are false alarms (B).

Table 2. Definition of a confusion matrix in a patient-level evaluation of MEWS.

	Clinical event occurred	No clinical event occurred
One or more alarms triggered	True positive	False positive
No alarm triggered	False negative	True negative

approach is to calculate a new MEWS value every time a physiological parameter gets refreshed, carrying forward the values of each parameter until a new value is available. This results in an irregularly sampled MEWS, referred to hereafter as $MEWS_{base}$ (this is a common approach of calculating MEWS). Another approach is to calculate MEWS at prescribed regular time intervals, T_{MEWS}, using a statistical summary of the available measurements for a physiological parameter within a time interval. We calculated MEWS and evaluated its performance for two different lengths of T_{MEWS} (2 and 12 h) and two different statistics: the median ($MEWS_{median}$) and the worst value of the available physiological measurements within T_{MEWS} ($MEWS_{worst}$). When no new measurements are available within T_{MEWS}, MEWS can be calculated using imputed physiological parameter values. Missing data in each physiological parameter were imputed by carrying forward the last recorded value until replaced with a new measurement. Figure 1 illustrates this approach. We implemented these variations of MEWS calculation to test if the proposed validation approaches could reveal whether and how these MEWS implementations would lead to different performances.

MEWS was calculated in the case and control groups according to the rules in table 1. Except for GCS which may not have been recorded for every patient, MEWS was not calculated (and patient was excluded from the analysis) when no measurements were available for a given vital sign in a patient's data.

2.3. Performance evaluation

We describe below two approaches for evaluating clinical prediction indices, using MEWS as a convenient example score to illustrate the concepts and the metrics introduced.

2.3.1. Patient-level evaluation

In many studies evaluating implementations of MEWS, the number of MEWS threshold crossings—given a fixed threshold value—were collectively evaluated with regards to correct prediction of an event (Lee *et al* 2008, Cooksley *et al* 2012, Fullerton *et al* 2012, Bulut *et al* 2014). That is, true predictions (or true positives) are clinical events for which MEWS crossed a threshold, regardless of how many times this occurred. When MEWS crosses a threshold in a patient who did not experience clinical deterioration during his/her hospitalization period it leads to a false positive, regardless of the number of times the threshold was crossed. Clinical deterioration events for which MEWS did not cross a threshold are missed predictions (or false negatives). Finally, true negatives are

Table 3. Summary of proposed metrics to evaluate a predictive alarm system.

Metric	Definition	Rationale	Formula
Time-dependent sensitivity S^{τ,τ_0}	Proportion of events preceded by at least one alarm within a prediction horizon	This metric quantifies the proportions of predicted events given a fixed time window within which MEWS scores above a threshold are considered true alarms, and outside of which they are considered false alarms	$S^{\tau,\tau_0} = \frac{N_{predicted\ events}}{N_{events}}$ $N_{predicted\ events}$: Number of predicted events N_{events}: Total Number of events
False positive ratio (FPR)	Proportion of the expected number of control patients in whom alarms are triggered within a window of a length equal to the prediction horizon	This metric estimates the proportion of control patients who are likely to trigger a false alarm. This estimate is based on randomized occurrences of a time window within which MEWS scores above a threshold are deemed false alarms	$\hat{\mu}_{FPR} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} T_{ij}}{N.M}$ N: Number of control patients M: Number of trials to estimate FPR (e.g. $M = 1000$) T_{ij}: jth randomly selected time window from the control data of the ith control patient $T_{ij} = \begin{cases} 1, & \textit{an alarm is triggered within } T_{ij} \\ 0, & \textit{otherwise} \end{cases}$
Alarm rate (r)	Number of alarms per unit time	This is a measure of the burden of alarms, including true and false alarms	$r = \frac{N_{alarms}}{T_{recording}}$ N_{alarms}: Number of alarms in a patient recording data $T_{recording}$: Recording data duration
Alarm proportion (ρ)	Proportion of score samples triggering an alarm	This is measure of the alarm burden. It quantifies the relative number of score samples triggering an alarm	$\rho = \frac{N_{MEWS > k}}{N_{MEWS \geqslant 0}}$ $N_{MEWS > k}$: Number of MEWS samples above thrshold k $N_{MEWS \geqslant 0}$: Number of MEWS samples
False alarm rate (r^0)	Number of false alarms per unit time	Measures the burden of false alarms by estimating their (hourly) frequency. A high rate may lead to alarm fatigue	$r^0(case) = \frac{N_{falseAlarms}}{T}$ $N_{falseAlarms}$: Number of false alarms $T = \begin{cases} T_{recording} - \tau: \textit{ for case patients} \\ T_{recording}: \textit{ for control patients} \end{cases}$
False alarm proportion (ρ^0)	Proportion of score samples triggering a false alarm	Measures the burden of false alarms by estimating their (relative) quantity. A high proportion may lead to alarm fatigue	$\rho^0 = \frac{N_{falseAlarms}}{N_{MEWS \geqslant 0}}$

Table 3. (Continued.)

Metric	Definition	Rationale	Formula
Work-up to detection ratio (*WDR*)	Ratio of the numbers of true and false predictions to the number of true predictions events	Measures the number of patients who will receive a workup before one additional adverse outcome can be prevented	$WDR = \frac{N_{case} + \hat{\mu}_{FP}}{N_{case}}$ N_{case}: Number of case patients in whom at least one alarm was triggered within a prediction horizon $\hat{\mu}_{FP}$: Estimated number of false positives
Time profile of alarm proportion	Normalized summary histogram of alarms with respect to number of alarms	An estimate of the probability of an alarm triggered early, on-time, late, or missing an event	N/A
Time profile of alarms per event	Normalized summary histogram of alarms with respect to number of events	Time profile of alarm proportion with respect to events	N/A

patients who did not experience clinical deterioration by the end of their hospitalization and in whom MEWS did not cross a threshed (table 2).

In a system where MEWS threshold crossings translate into MEWS alarms, the above definitions become limited as they do not quantify the timeliness and burden of these alarms. We hereafter refer to the above evaluation scheme as *patient-level* evaluation and propose an *event-level* evaluation by considering the timeliness and the temporal distribution of MEWS alarms to quantify the burden and practicality of notifications. MEWS alarms can be defined as events triggered and cleared following a set of rules. Here, and to simplify subsequent analysis, we define a MEWS alarm an instantaneous notification occurring when MEWS exceeds a given threshold value (figure 2(A)).

2.3.2. Prediction horizon and lead time

The definition of a temporal window preceding the time of an event allows performance to be evaluated in respect to an actionable timeframe. MEWS alarms have different clinical implications when they occur 'too early' or 'too late'. We therefore define two temporal windows, the *prediction horizon* and the *lead time*, to quantify the timeliness of alarms. A prediction horizon is a time window preceding the time of event (i.e. [$t_{\min}$ $t_{\max}$] where $t_{\min} < t_{\max} < t_{event}$), within which alarms are deemed actionable, that is providing sufficient time for caregivers to intervene (figure 2(B)). For code blue events, the length of a prediction horizon $\tau = t_{\max} - t_{\min}$ could be defined according to typical ICU nurse's shift (e.g. 12 h).

$t_{\min}$ is the earliest time (with respect to t_{event}) an alarm can contribute to true predictions. By varying $t_{\min}$ and setting $t_{\max} = t_{\min} + \tau$, we can create a realistic and clinically meaningful characterization of the ability of MEWS to predict code blue events in the ICU. Applying such definitions of $t_{\min}$ and $t_{\max}$ is a general approach and allows performance characterization at a lead time before the onset of event, defined by $t_{\max}$, where the lead time is the interval of time [$t_{\max}$ t_{event}]. The duration of the lead time τ_0 serves as the minimum time window required for an intervention to be effective. Any alarm occurring within the lead time is a late alarm. It is considered either a 'short-notice' or a redundant alert if it is preceded by one or more alarms triggered within the prediction horizon. Alarms generated after event onset (t_{event}) are deemed missed alarms. Early, late, and missed alarms are all false alarms.

2.3.3. Event-level evaluation

Given the definitions of prediction horizon and lead time, we derive a series of time-dependent metrics to evaluate the performance of MEWS and the burden of alarms. Table 3 summarizes these metrics and their definitions.

2.3.3.1. Time-dependent sensitivity

Alarms triggered within the prediction horizon are defined as true alarms and alarms triggered outside of the prediction horizon (before or after) as false alarms. To quantify the sensitivity of MEWS given these definitions, we consider an event to be successfully predicted if it was preceded by at least one true alarm (Notice that a distinction is made between a true alarm and a true positive. A 'positive' refers a code blue event just like it was

defined in the patient-level evaluation.) This sensitivity depends on the choice of τ and τ_0 and its formally defined as the proportion of clinical events for which at least one alarm was triggered within the prediction horizon:

$$S^{\tau,\tau_0} = \frac{N_{predicted\ events}}{N_{events}}, \tag{2.1}$$

where $N_{predicted\ events}$ is the number of predicted clinical events and N_{events} is the total number of clinical events.

2.3.3.2. Time-dependent specificity

Event-level specificity metrics can be derived from case and control records. We first propose to estimate the false positive ratio (FPR) in the control group, defined as the ratio of the expected number of control patients in whom alarms are triggered within a window of a length equal to the prediction horizon, to the total number of control patients N (Bai *et al* 2015). Let $\hat{\mu}_{FPR}$ be the expected the estimated value of FPR. $\hat{\mu}_{FPR}$ can be calculated using a bootstrapping approach. For each control patient i ($1 \leqslant i \leqslant$ N), we randomly sample a window of length τ over its whole monitoring time. This is repeated a sufficiently large number of time M (e.g. $M = 1000$). The windows are then temporally sorted and indexed using their temporal order (the jth window is preceded in time by the $(j-1)$th window). We then calculate the number of windows in which one or more alarms occurred (triggered windows) to estimate the expected value of whether the patient counts as a false positive.

Let $T_{ij} = 1$ ($1 \leqslant i \leqslant N; 1 \leqslant j \leqslant M$) if the jth window selected for the ith control patient gets triggered, and $T_{ij} = 0$ otherwise. The expected number of control patients 'triggering' a jth window is:

$$\hat{\mu}_{FP}^{j} = \sum_{i=1}^{N} T_{ij} \tag{2.2}$$

and its standard deviation is:

$$\hat{\sigma}_{FP}^{j} = \sqrt{\frac{\sum_{i=1}^{N}(T_{ij} - \hat{\mu}_{FP}^{j})^2}{N-1}}. \tag{2.3}$$

The estimated value $\hat{\mu}_{FPR}$ and its standard deviation can then be calculated as:

$$\hat{\mu}_{FPR} = \frac{\sum_{j=1}^{M} \hat{\mu}_{FP}^{j}}{M} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} T_{ij}}{N.M}, \tag{2.4}$$

$$\hat{\sigma}_{FPR} = \sqrt{\frac{\sum_{j=1}^{M}(\hat{\sigma}_{FP}^{j})^2}{M}}. \tag{2.5}$$

Other metrics (accuracy, positive and negative predictive values and F1 score) can be calculated using the same bootstrapping approach (see supplementary material S1 (available online at stacks.iop.org/PMEA/42/055005/mmedia)). To quantify the specificity of MEWS in the case groups, we propose to calculate the proportion and rate of alarms and false alarms which also quantify the alarm burden.

2.3.3.3. Alarm burden

The number of alarms generated during an ICU stay in case patients can be summarized by visualizing the distribution of the alarms that occur 'early', 'late', and 'on-time', and alarms that were 'missed', with respect to the prediction horizon and time of event. To quantify these distributions, alarms from all case patients are summed up within each of the four attributed windows to create a *time profile or alarms*. This profile may be interpreted as the probability distribution of generating an alarm early, on-time, late, or missing an event. The number of alarms can further be normalized by the total number of alarms or by the total number of events resulting in a *time profile of alarms per event* and a *time profile of alarm proportions*, respectively (Scully and Daluwatte 2017).

Additionally, we calculate the rate of alarms and false alarms (r and r^0, resp.), and the proportion of alarms and false alarms (ρ and ρ^0, resp.) to measure the burden of alarms and false alarms. r and r^0 measures the number of alarms and false alarms, respectively, per unit time per patient.

2.3.3.3.1. Rate and proportion of alarms

For a given patient in the case or control group, r and ρ are given by:

$$r = \frac{N_{alarms}}{T_{recording}}, \tag{2.6}$$

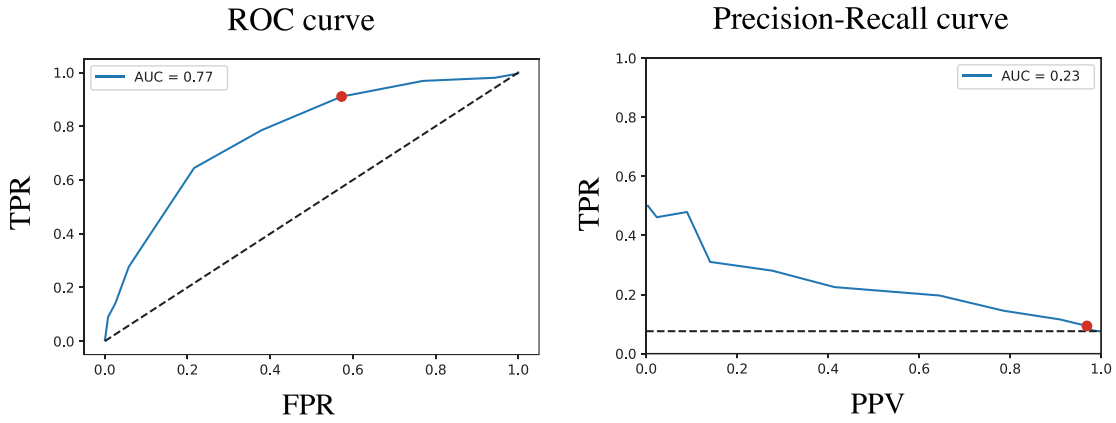

Figure 3. ROC curve (left) and precision-recall (right) curve of MEWS performance evaluated for 14 discrete threshold levels ($k = 0$–13) using the patient-level evaluation approach. The area under the curve (AUC) is a measure of MEWS performance. Dashed lines represent chance level. Red dots indicate the performance of MEWS at threshold $k = 4$.

$$\rho = \frac{N_{alarms}}{N_{MEWS}} = \frac{N_{MEWS > k}}{N_{MEWS \geqslant 0}},\ 0 \leqslant k \leqslant 13, \tag{2.7}$$

where N_{alarms} is the number of MEWS above a threshold k and N_{MEWS} is the number of MEWS samples (equivalently, the number of MEWS above threshold $k = 0$) in the patient recording.

2.3.3.3.2. Rate and proportion of false alarms

A false alarm is an alarm that occurs outside of the prediction horizon by the previous definitions. Hence, the false alarm rate for a given patient in the case group is given by:

$$r^0(case) = \frac{N_{falseAlarmsCase}}{T_{recordingCase} - \tau}. \tag{2.8}$$

With $N_{falseAlarmsCase}$ is the total number of false alarms, and $T_{recordingCase}$ the duration of the recording. Here, the duration of the prediction horizon is subtracted from the total duration of recording since a false alarm cannot occur within a prediction horizon by definition.

To calculate r^0 for a control patient, we simply derive the average rate of alarms triggered for a control patient:

$$r^0(control) = \frac{N_{falseAlarmsControl}}{T_{recordingControl}}, \tag{2.9}$$

where $N_{falseAlarmsControl}$ is the total number of false alarms and $T_{recordingControl}$ the duration of the recording in a given control patient.

By substituting $N_{falseAlarmsCase}$ and $N_{falseAlarmsControl}$ in the numerator of equation (2.7) we get the proportion of false alarms $\rho^0(case)$ and $\rho^0(control)$ in the case and control group, respectively.

2.3.3.4. Work-up to detection ratio (WDR)

Furthermore, we introduce the *WDR* ratio to evaluate the effectiveness of MEWS. *WDR* is defined as the ratio of the number of case patients N_{case} in whom at least one alarm was triggered within a prediction horizon (equivalently the number of predicted events) and the number of control patients in whom a random window of length τ is triggered (i.e. the estimated number of false positives, $\hat{\mu}_{FP}$), to the number of predicted events:

$$WDR = \frac{N_{case} + \hat{\mu}_{FP}}{N_{case}}. \tag{2.10}$$

Conceptually, *WDR* is similar to the number needed to alert, defined as the number of alerts that need to be reviewed to detect one potential adverse event (Moore *et al* 2009). WDR measures the number of required workups to prevent one additional adverse outcome (equivalently, the number of case patients to be treated for one of them to benefit from the treatment compared with a control patient).

Using (2.4), and (2.5), *WDR* and its standard deviation can be estimated as:

$$\hat{\mu}_{WDR} = 1 + \frac{\sum_{i=1}^{N} \hat{\mu}_i}{N_{case}}, \tag{2.11}$$

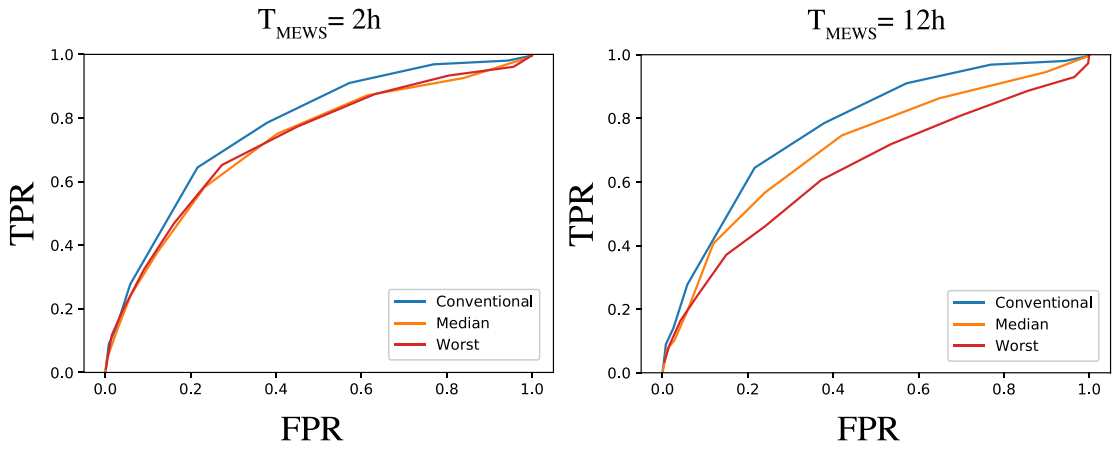

Figure 4. Performance of MEWS calculated at irregular time points ($MEWS_{base}$) and at regular time intervals $T_{MEWS} = 2$ and 12 h using the median ($MEWS_{median}$) and the worst value of physiological measurements within T_{MEWS} ($MEWS_{worst}$). MEWS threshold was set to $k = 4$.

Table 4. Average sampling rate (measurements per hour) of physiological parameters and proportion of missing values in the case and control groups.

		GCS	Temp	SBP	RR	HR
Case	Average sampling rate (1/h)	0.3	1	1.2	2	2.2
	Proportion of imputed values	0.9	0.5	0.5	0.1	0
Control	Average sampling rate (1/h)	0.1	0.4	0.5	0.7	0.8
	Proportion of imputed values	0.9	0.6	0.4	0.1	0.1

Table 5. Performance of MEWS using patient-level and event-level evaluations. MEWS threshold was set to $k = 4$.

Evaluation	TPR	FPR (std)	NPV (std)	PPV (std)	ACC (std)	F1 (std)	WDR (std)
Patient-level	0.97	0.77	0.99	0.09	0.29	0.17	10.7
Event-level ($\tau = 12$ h, $\tau_0 = 0, M = 1000$)	0.87	0.30 (0.14)	0.98 (0.00)	0.23 (0.12)	0.71 (0.13)	0.35 (0.13)	5.3 (1.94)

Note. TPR: True Positive Rate (Sensitivity); FPR: False Positive Rate; NPV: Negative Predictive Value; PPV: Positive Predictive Value; ACC: Accuracy. WDR: Workup-to-detection ratio.

$$\hat{\sigma}_{WDR} = \frac{\sqrt{\sum_{i=1}^{N} \hat{\sigma}_i^{\,2}}}{N_{case}}. \tag{2.12}$$

3. Results

27 of 283 case patients (9.5%) and 6 of 3127 control patients (0.2%) were excluded from the analysis because MEWS could not be calculated (due to missing of one or more vital signs) and/or because the duration of the recording was less than the prediction horizon $\tau = 12$ h for the sensitivity to be evaluated in case patients and FPR to be estimated in control patients. This led to 7% prevalence of code blue events.

$MEWS_{base}$ was evaluated using the traditional metrics (AUC, accuracy, sensitivity, specificity, and precision). We then evaluated the impact of calculating MEWS from summary statistics of physiological measurements in comparison with $MEWS_{base}$. Finally, we evaluate MEWS using the proposed metrics. The prediction horizon τ was set to 12 h—typical length of nursing shift—and the lead time τ_0 was varied between 0 and 6 h at 10 min increments. Case patients for whom the length of ICU stay was less than the sum of lead time and seizure prediction horizon (i.e. length of stay < 12 h $+ \tau_0$) were excluded from the relevant analysis since true predictions could not be evaluated. The number of excluded patients increased linearly from 0 (for $\tau_0 = 0$) to 38 (for $\tau_0 = 6$ h). MEWS was evaluated for threshold values k varying between 0 and 13. Table 4 describes the frequency of vital signs and GCS measurements and the proportion of imputed values in the dataset.

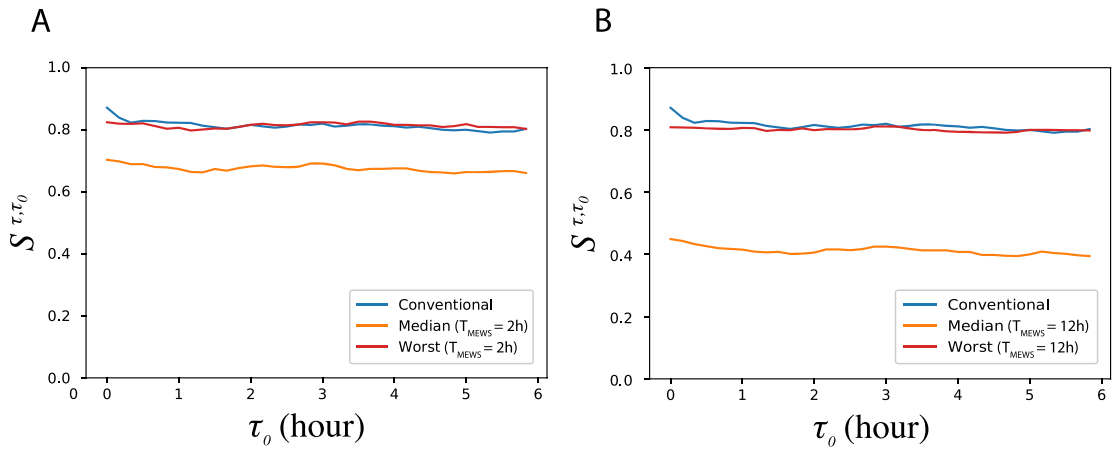

Figure 5. Sensitivity S^{τ,τ_0} of MEWS evaluated for a prediction horizon $\tau = 12$ h and lead time τ_0 varied between 0 and 6 h at 10 min increments. MEWS was calculated at irregular time points (conventional or $MEWS_{base}$) and at regular time intervals $T_{MEWS} = 2$ h (A) and 12 h (B) using the median ($MEWS_{median}$) and the worst value of physiological measurements within T_{MEWS} ($MEWS_{worst}$). MEWS threshold was set to $k = 4$.

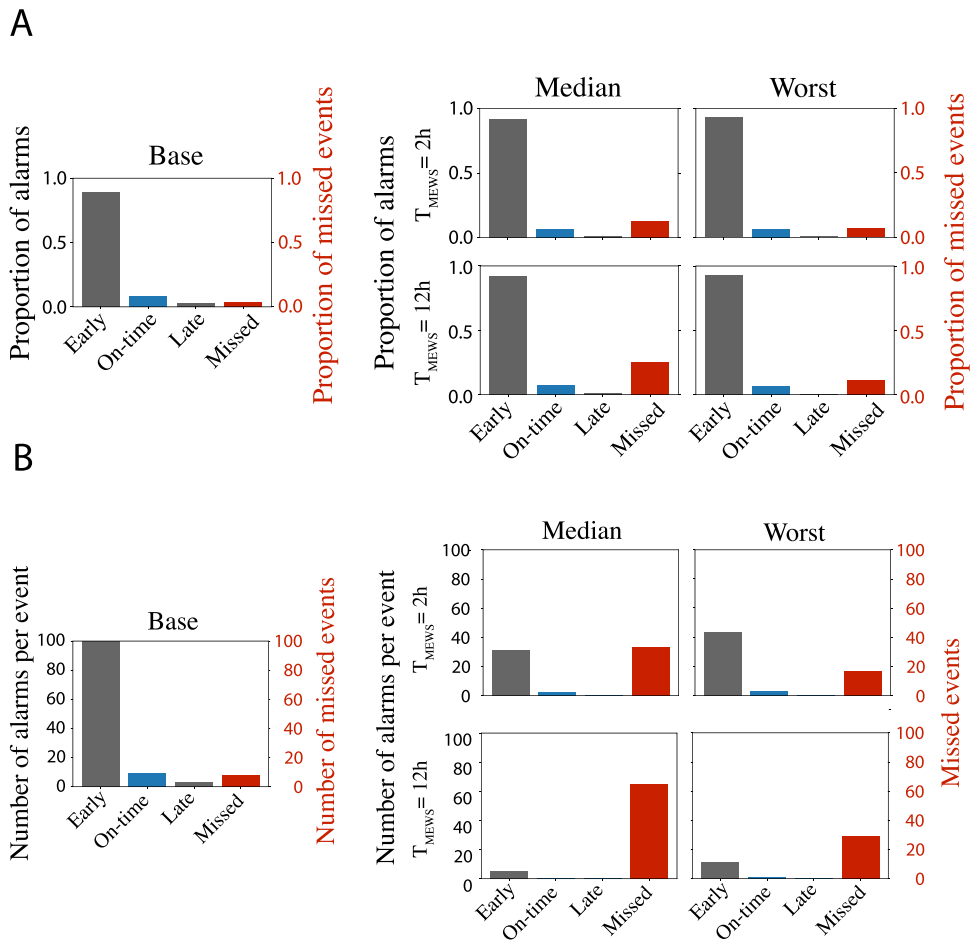

Figure 6. Time profile of alarms and alarms per event calculated in case patients for 3 different calculation methods of MEWS and 2 values of T_{MEWS}. (A) Proportion of early, on-time, and late alarms and missed events. (B) Number of early, on-time and late alarms per event and number of missed events. MEWS threshold was set to $k = 4$. Prediction horizon was set to $\tau = 12$ h and lead time was set to $\tau_0 = 1$ h.

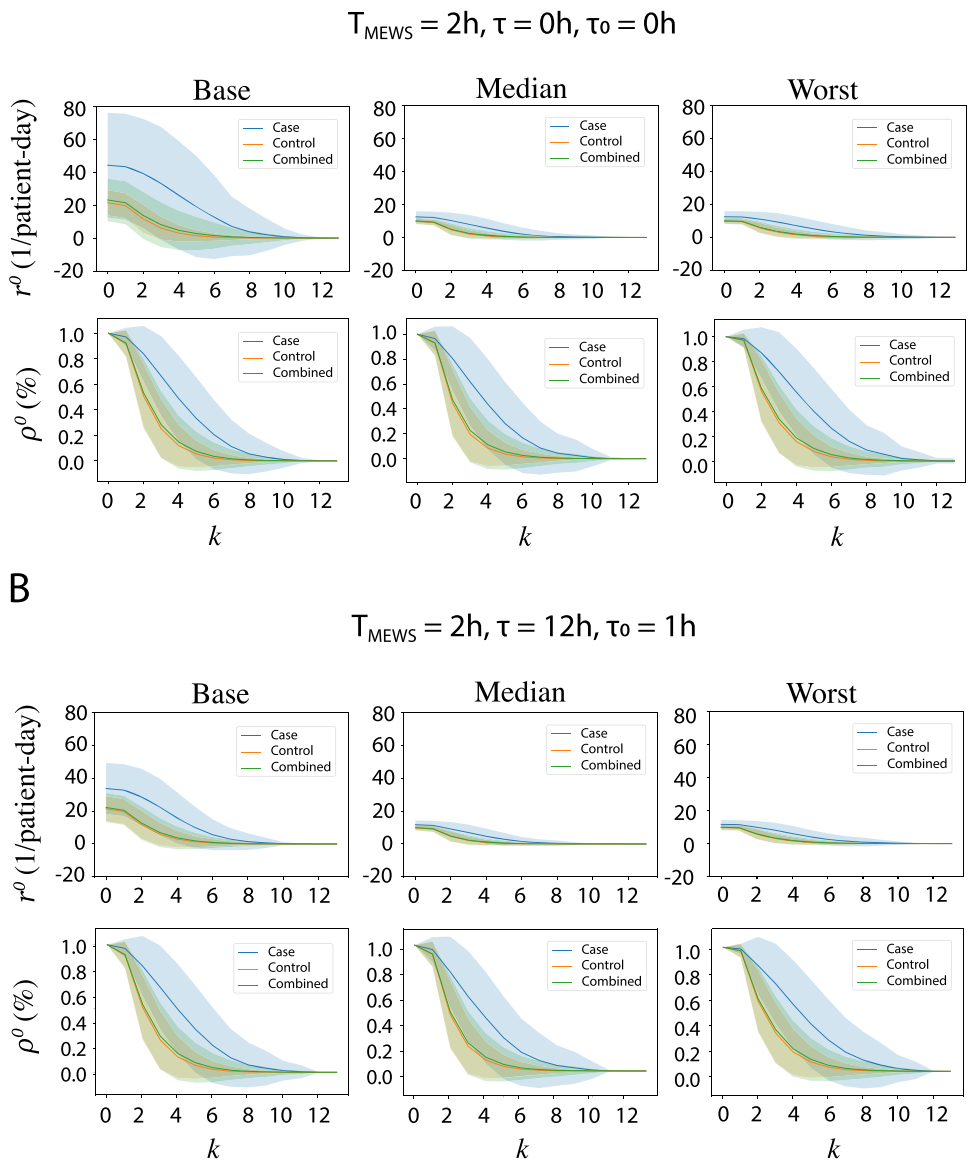

Figure 7. False alarm rate and proportion of false alarms in the case and control groups, and in the combined population for different MEWS thresholds k, T_{MEWS} = 2 h, and two different value sets of the prediction horizon and the lead time: (A) τ = 0 h, τ_0 = 0 h and (B) τ = 12 h, τ_0 = 1 h. Shadows indicate one standard deviation of the mean.

3.1. Patient-level and event-level evaluation of $MEWS_{base}$

Table 5 shows the performance of $MEWS_{base}$ evaluated using classic metrics applied to the patient-level and event-level approaches (see sections 2.3.1 and 2.3.3, resp.). MEWS threshold k was set to 4, a commonly used threshold value (Subbe *et al* 2001, Burch *et al* 2008).

Using a patient-level evaluation, the high sensitivity of MEWS (TPR = 0.97) was significantly compromised by low precision (0.09) and low specificity (FPR = 0.77). Figure 3 depicts the performance of MEWS for each of the 14 possible threshold values using a Receiver operating characteristic (ROC) curve and the precision-recall curve (PRC).

Using an event-level approach with a prediction horizon τ = 12 h and a lead time τ_0 = 0 h in the case group, and 1000 random windows in the control group, a substantial improvement in the precision (~155%), the FPR (~60%) and in the WDR (~50%) could be achieved in the expense of a relatively marginal decrease in sensitivity (~10%).

The performance of MEWS calculated at regular time intervals was lower than that of the $MEWS_{base}$ (measured by AUC) regardless of the length of T_{MEWS} and the statistic used (figure 4). $MEWS_{median}$ was less affected by the length of T_{MEWS} than $MEWS_{worst}$, which performance degraded using a T_{MEWS} of 12 h compared to a T_{MEWS} of 2 h.

Table 6. Number of MEWS samples, alarm rate (r) and alarm proportion (ρ), in the case and control group. For different MEWS calculation methods. Average numbers shown. MEWS threshold was set to $k = 4$.

	$MEWS_{Base}$				$MEWS_{Median}$		$MEWS_{Worst}$	
	Case	Control			Case	Control	Case	Control
Number of MEWS samples (1/patient-day)	51.98	21.50	T_{MEWS} (h)	2	14.17	9.61	14.17	9.61
				12	5.70	2.10	5.70	2.10
r (1/patient-day)	13.68	4.08	T_{MEWS} (h)	2	4.08	1.2	5.52	2.16
				12	0.72	0.24	1.44	0.72
ρ (%)	0.46	0.19	T_{MEWS} (h)	2	0.37	0.13	0.52	0.22
				12	0.33	0.11	0.72	0.35

3.2. Time-dependent sensitivity

Setting MEWS threshold to $k = 4$ and the prediction horizon to $\tau = 12$ h, there was no significant change in the sensitivity when the lead time was varied between 0 and 6 h (see figure 5). The sensitivity of $MEWS_{median}$ was significantly lower than that of $MEWS_{worst}$ and $MEWS_{base}$ (t-test, $p < 0.05$) when T_{MEWS} was set to 2 or 12 h, and for different lead time values (see section 2.3.3.1). $MEWS_{base}$ and $MEWS_{worst}$ showed comparable sensitivities for either value of T_{MEWS}.

3.3. Alarm burden

3.3.1. Alarm time profile

Figure 6 illustrates the distribution of the proportion of alarms and alarms per event (see section 2.3.3.3) across four time periods. Regardless of the length of T_{MEWS} and how MEWS is calculated, most alarms were triggered early, before the start of the pre-defined prediction horizon. The proportion of missed events is relatively lower with $MEWS_{base}$ than with $MEWS_{Median}$ or $MEWS_{worst}$, for both T_{MEWS} values. Longer T_{MEWS} results into higher proportion of missed events and a lower rate of early and on-time alarms per event for both $MEWS_{Median}$ and $MEWS_{worst}$.

3.3.2. False alarm rate and proportion of false alarms (see section 2.3.3.3.2)

We set the values of prediction horizon τ and lead time τ_0 to 0 in order to evaluate the effect of calculating MEWS with different methods on the burden of false alarms. The average false alarm rate and average proportion of false alarms across patients were generally highest for $MEWS_{base}$ and decayed exponentially with increasing values of k (figure 7(A)). At $k = 4$ and $T_{MEWS} = 2$ h, the average combined (case and control) proportion of false alarm ρ^o and the average combined false alarm rate r^0 were significantly higher (t-tests, $p < 0.05$) for $MEWS_{base}$ (19% and 0.19/h, resp.) compared to $MEWS_{median}$ (12% and 0.05/h resp.). $MEWS_{worst}$ resulted in significantly lower r^0 (0.08/h) but not ρ^o (18%) compared with $MEWS_{base}$.

At $\tau = 12$ h and $\tau_0 = 1$ h, both r^0 and ρ^o decayed similarly as for $\tau = \tau_0 = 0$ h (figure 7(B)). For $k \geqslant 4$ any improvement in r^0 and ρ^o is arguably marginal. Highest values of r^0 and ρ^o were observed for $MEWS_{base}$ overall.

3.3.3. Rate and proportion of alarms (see section 2.3.3.3.1)

At $\tau = 0$ and $\tau_0 = 0$, the average alarm rate decayed exponentially with MEWS threshold as was the case of false alarm rate. $MEWS_{base}$ generates the highest alarm rate at any MEWS threshold value compared with $MEWS_{median}$, and $MEWS_{worst}$, which both led to comparable alarm rates for all values of k and T_{MEWS} values and larger alarm rates for longer T_{MEWS} values. Average alarm rates were expectedly higher in the case group than in the control group.

The proportion of MEWS values above a given threshold (alarm proportion) was relatively comparable for all MEWS calculation methods and T_{MEWS} values. Compared with conventional MEWS, the alarm rate decreased by ~70% (from 13.68/patient-day to 4.08/patient-day) in the case group for $T_{MEWS} = 2$ h, and by 95% for $T_{MEWS} = 12$ h (from 13.68/patient-day to 0.37/patient-day). This translates into a reduction in the hourly alarm rate from ~1 alarm every 1.7 h to ~1 alarm every 5.8 h, and to 1 alarm every 33 h, for $T_{MEWS} = 2$ h and 12 h, respectively. Table 6 provides a summary of alarm proportions and alarm rates for $k = 4$.

4. Discussion

The clinical importance of predictive alarm systems as risk management tools to help prevent in-hospital patient deterioration has been discussed in numerous systematic reviews (Robert *et al* 1999, McNeill and Bryden 2013,

Vincent *et al* 2018, Kramer *et al* 2019, Gerry *et al* 2020). Many of these reviews highlighted the methodological weaknesses and questioned the true predictive power of these systems. Methodological and statistical evaluations that consider among other aspects the temporal dynamic of the measure profile (in the case of continuous monitoring), the timeliness of alarms, and frequency of false alarms can aid in this evaluation. In this study, we investigated the performance a widely adapted EWS, MEWS, using a new set of metrics that capture two critical aspects of any predictive alarm algorithm, namely the timeliness and the burden of alarms. We adjusted the definition of sensitivity to take into account two clinically important time intervals, the prediction horizon and the lead time. These parameters can help provide performance information that may be more informative on clinical utility, since an alarm is arguably useful if it's *actionable* (occurs within a nursing shift) and *intervenable* (provides a minimum time for an intervention to be effective). This proposed definition of sensitivity informs whether a continuous predictive alarm algorithm can practically identify patients at risk of deterioration and alert caregivers early enough to intervene. The characteristic curve of the sensitivity versus lead time can guide the choice of operational lead time values.

This study aims at proposing a general framework of methods and metrics for evaluating the performance of early warning scoring and predictive algorithms. The intended framework is sought to engage researchers in testing the predictive power of a proposed algorithm under different conditions and alarm protocols to emulate the realm of real-word clinical setting. Parameters such as the prediction horizon and the lead time can be applied universally to test the performance of any predictive alarm algorithm operating on continuous data. Moreover, a variety of alarm triggering strategies can be envisaged and tested to assess if they influence the clinical burden.

4.1. Limitations

This study has limitations. While patients were selected consecutively to avoid a selection bias, some case patients had significantly shorter recordings than others. Some metrics (e.g. proportion of early alarms) may have been biased for these patients leading to lower values than what could be observed if recordings had the same length. Estimating specificity-metrics (proportion of false alarms, false alarm rate) over recordings of equal length rather would have introduced a patient selection bias. To overcome these limitations, an alternative approach would be to normalize the number of false alarms by the duration of time preceding the prediction horizon (e.g. 10 false alarms triggered within 1 h would contribute the same proportion of early alarms as 100 false alarms triggered within a 10 h timeframe).

Constraining the timeframe in which an alarm is counted a true alarm to a prescribed horizon while relaxing the period of time alarms are labeled as false inherently biases the proportion of false to true alarms. The false alarm rate may not be affected if we assume that the probability of false alarm occurrence is independent of time-to-event. Such an assumption may not be (always) true though. For example, a MEWS system in which the threshold is adjusted as information about the patients' health status matures with the length of stay may result in fewer (hourly) false alarms generated towards the end of stay than false alarms generated early after the admission using the 'default' threshold value.

Performance results depend on alarm definition and protocol. Had we grouped multiple threshold crossings to form one alarm, for example by applying a lockout period after each alarm, within which no alarm would be generated, the rate and proportion of alarms obtained (see table 6) would have likely been reduced since less alarms would have been generated. Such reduction could in fact be significant and lessens the alarm burden. Studies which applied a lockout period demonstrated a decrease of 55% in alerts from a predive model for cardiothoracic ICU patients (King *et al* 2012). A recent study demonstrated that the number of alerts could be reduced by 90% when a 15 min lockout period was applied to a prediction algorithm for ICU hypotension events (Yoon *et al* 2020).

EHR data are sparse with irregular time interval, it may be challenging to design rules to group alarms that result in appreciable difference. However, if MEWS was generated from continuous vital signs (i.e. from bedside monitors), the number of MEWS alarms generated would have differed significantly in both alarm approaches and would have led to significant difference in sensitivity and specificity results.

Finally, data imputation through forward carrying of the last available physiological measurement were not constrained to expire after an amount of time in this study. It is worth noting that this practice is safe for imputing missing physiological parameter values that are slow changing in nature, such as temperature, or GCS in the case of MEWS, but may not reflect the true underlying physiological state in the case of other parameters. Limiting the time an imputed value may be carried forward can limit the effect of misrepresenting the physiological state but may result in missing samples (in the case of regular sampling) if no measurements are available after expiration of the imputed physiological parameter value.

A critical step in the development of a predictive algorithm is to statistically validate its performance. That is to demonstrate it has an above-chance predictive power. A key question is whether an adverse event can be

predicted and whether an algorithm, which is presumed to perform better than chance, does not have predictive power but simply has not yet been tested against appropriate null hypotheses (Andrzejak *et al* 2009). Different statistical approaches have been proposed to answer this question. These include analytical methods (e.g. comparison with naïve predictors) and Monte Carlo based methods (e.g. alarm surrogates) (Schelter *et al* 2006, Snyder *et al* 2008, Feldwisch-Drentrup *et al* 2011).

4.2. On the performance of MEWS

The proposed metrics were applied to measure and evaluate the performance of different implementations of MEWS to predict code blue events. Here, we discuss new insights as provided by the proposed performance metrics relating to the potential clinical performance of MEWS that were missing in prior literature (Fairclough *et al* 2009, Heitz *et al* 2010, Tirotta *et al* 2017, Akgun *et al* 2018, Xie *et al* 2018, Jiang *et al* 2019, Ahn *et al* 2020, Aygun *et al* 2020, Balshi *et al* 2020, Gerry *et al* 2020, Kumar *et al* 2020).

Although we found that the sensitivity did not significantly change with lead time values up to 6 h (see section 3.2), it remains to be investigated if higher sampling rates of physiological measurements (e.g. continuous vital sign recordings from patient monitors) can uncover trending changes in MEWS values that may lead to higher sensitivities at short lead times. Additionally, measurements were sampled on average 2.5 times more frequently in patients of the case group than of the control group to presumably capture sudden physiological changes that are more likely to occur in the sicker patient. Higher sampling rates in the control group may have led to different distributions of MEWS alarms and therefore influenced the calculated performance results. The effect of the sampling rate on the calculated metrics may be less prominent for $MEWS_{median}$ than for $MEWS_{worst}$ and conventional MEWS since $MEWS_{median}$ is the least affected by a change in data fluctuation that may be caused by increasing the number of samples within T_{MEWS}.

Despite a high sensitivity (97%) observed using the widely adopted patient-level evaluation approach, the false positive rate and the precision of the conventional MEWS were remarkably poor at a threshold value of 4 (77% and 9%, resp; see section 3.1). One of ten patients who triggered an alarm had a code blue outcome (WDR $\sim$ 10), while almost 8 of 10 patients were falsely identified as potentially high-risk patients (FPR $=$ 77%). Judging whether these numbers are acceptable amounts to mainly evaluating the clinical burden associated with a false alert for a particular application. Additionally, the rate of alarms was highest for conventional MEWS. By calculating MEWS over 12 h window, the alarm rate dramatically decreased. For example, with $MEWS_{median}$, one alarm would be triggered every 33 h, instead of every 1.7 h in the case of conventional MEWS. A substantial reduction in the frequency of alarms in the ICU minimizes alarm fatigue (Blum and Tremper 2010).

A direct comparison of these results with the performance of MEWS reported in previous studies may not be feasible since the datasets, the frequency of measurements, and the outcomes analyzed vary among MEWS studies. Noticeably, studies that evaluated the effect of MEWS implementation on cardiac and cardiopulmonary arrests reported mixed performance results (Subbe *et al* 2003, Jones *et al* 2011, Moon *et al* 2011, Churpek *et al* 2012). It's worth emphasizing that MEWS was not designed to be a predictor of code blue events nor it is the standard of care to use MEWS in ICU settings. It's a screening tool to identify inpatients at risk for deterioration and to trigger early evaluation and transfer to step-down or intensive care. The low precision and high false positive rate in ICU population we analyzed do not necessarily indicate a deficiency of MEWS. We simply used MEWS as un exemplar of a clinical prediction index to examine characterization and performance evaluation techniques, and not to examine the performance of MEWS as a predictor of code blue events.

By adopting an event-level evaluation of MEWS where only alarms triggered within a defined nursing shift period of 12 h are considered true alarms, performance levels changed significantly compared to those obtained under patient-level evaluation. Notably, the number of workups required to detect a cardiac arrest dropped significantly by close to 50%, from 10.7 to 5.3 (see section 3.1). Less patients to evaluate translates to less disruption of the clinical workflow and care cost saving. This improvement in performance is due to reducing the probability of false positives by limiting false alarms to those occurring within nursing shifts in a control recording as defined in the event-level evaluation (see section 2.3.3.2). Many control patients may have sporadic false alarms, which under the proposed event-level evaluation are less likely to trigger randomly sampled windows and lead to false positives. It is important to note that the proposed definition of WDR does not consider the burden of alerts leading to a workup. One may define an alarm-level WDR that measures how many false alarms get triggered before a cardiac arrest gets detected (true alarm), which can be calculated as the total number of true and false alarms to the number of true alarms. An efficient predictive system is one that demonstrates a practically low level of alarms per outcome.

The prediction horizon is a parameter that can be selected based on the clinical relevance of an application. The choice of 12 h nursing shift as a prediction horizon for the evaluation of MEWS in this study was motivated by the need to select an actionable timeframe within which MEWS alerts should lead to intervention to demonstrate the proposed performance metrics. We used 12 h here because it is the typical duration of a nursing

shift in US hospital settings, including critical care (Stimpfel *et al* 2019). In an ICU unit, two teams of nurses—one per shift—ensure around the clock monitoring. Assuming MEWS threshold is adjustable by the attending nurse, MEWS alerts could be evaluated relative to the nursing shift where they were triggered. A nurse covering the morning shift may set the threshold to different value than the nurse from the previous night shift.

Interestingly, conventional MEWS led to similar sensitivity as with MEWS calculated from worst case values (see section 3.2). The alarm burden, however, was less with $MEWS_{worst}$ than with conventional MEWS (see section 3.3). Despite a slightly lower proportion of on-time alarms and higher missed alarms, $MEWS_{worst}$ may be regarded as better choice than conventional MEWS as it maintains a high level of sensitivity and alleviates the alarm burden. The probability distribution of generating alarms did not substantially differ between MEWS implementations tested here, indicating that different MEWS calculation methods change the time distribution of alerts but not their predictive power.

5. Conclusion

Comprehensive evaluation of predictive alarm algorithms to establish their true predictive power and clinical usefulness is lacking. This study proposes approaches and technical considerations to quantify the performance and practicality of predictive alarm algorithms in predicting adverse clinical events. Using MEWS as a case study, revisited classic measures of performance and tools that capture and illustrate the burden of alarms are presented and compared to conventional approaches. The proposed approach addresses the incompleteness and limitation of classic measures to incorporate key clinical considerations and suggests measures and methods that capture the clinical burden and the timeliness of alarms to guide the judgment about the practicality and utility of a candidate predictive alarm algorithm.

Work performed at the Department of Physiological Nursing, School of Nursing, UCSF, San Francisco, CA and at the Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, MD.

Conflicts of interest and source of funding

None declared

Reprints will not be ordered.

Funding

This work was supported in part by the National Institutes of Health (grant R01HL128679) and by the Center of Excellence in Regulatory Science and Innovation (CERSI) grant to University of California, San Francisco (UCSF) and Stanford University from the US Food and Drug Administration (U01FD005978). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the HHS or FDA.

ORCID iDs

Kais Gadhoumi https://orcid.org/0000-0003-4148-6118
Christopher G Scully https://orcid.org/0000-0001-8244-0832
Ran Xiao https://orcid.org/0000-0002-3689-1680
David O Nahmias https://orcid.org/0000-0001-6159-1172
Xiao Hu https://orcid.org/0000-0001-9478-5571

References

Ahn J H, Jung Y K, Lee J-R, Oh Y N, Oh D K, Huh J W, Lim C-M, Koh Y and Hong S-B 2020 Predictive powers of the modified early warning score and the national early warning score in general ward patients who activated the medical emergency team *PLoS One* **15** e0233078

Akgun F S, Ertan C and Yucel N 2018 The prognastic efficiencies of modified early warning score and mainz emergency evaluation score for emergency department patients *Niger. J. Clin. Pract.* **21** 1590–5

Alam N, Hobbelink E L, van Tienhoven A J, van de Ven P M, Jansma E P and Nanayakkara P W B 2014 The impact of the use of the early warning score (EWS) on patient outcomes: a systematic review *Resuscitation* **85** 587–94

Al-Kalaldeh M, Suleiman K, Abu-Shahroor L and Al-Mawajdah H 2019 The impact of introducing the modified early warning score 'MEWS' on emergency nurses' perceived role and self-efficacy: a quasi-experimental study *Int. Emerg. Nurs.* **45** 25–30

Andrzejak R G, Chicharro D, Elger C E and Mormann F 2009 Seizure prediction: any better than chance? *Clin. Neurophysiol.* **120** 1465–78

Aygun H, Eraybar S, Ozdemir F and Armagan E 2020 Predictive value of modified early warning scoring system for identifying critical patients with malignancy in emergency department *Arch. Iran Med.* **23** 536–41

Bai Y, Do D H, Harris P R, Schindler D, Boyle N G, Drew B J and Hu X 2015 Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction *J. Biomed. Inform.* **53** 81–92
Balshi A N *et al* 2020 Modified early warning score as a predictor of intensive care unit readmission within 48 h: a retrospective observational study *Rev. Bras. Ter. Intensiva.* **32** 301–7
Blum J M and Tremper K K 2010 Alarms in the intensive care unit: too much of a good thing is dangerous: is it time to add some intelligence to alarms? *Crit. Care Med.* **38** 702–3
Bulut M, Cebicci H, Sigirli D, Sak A, Durmus O, Top A A, Kaya S and Uz K 2014 The comparison of modified early warning score with rapid emergency medicine score: a prospective multicentre observational cohort study on medical and surgical patients presenting to emergency department *Emerg. Med. J.* **31** 476–81
Burch V C, Tarr G and Morroni C 2008 Modified early warning score predicts the need for hospital admission and inhospital mortality *Emerg. Med. J.* **25** 674–8
Churpek M M, Yuen T C, Huber M T, Park S Y, Hall J B and Edelson D P 2012 Predicting cardiac arrest on the wards: a nested case-control study *Chest* **141** 1170–6
Cooksley T, Kitlowski E and Haji-Michael P 2012 Effectiveness of modified early warning score in predicting outcomes in oncology patients *QJM* **105** 1083–8
Damen J A *et al* 2016 Prediction models for cardiovascular disease risk in the general population: systematic review *Br. Med. J.* **353** i2416
Drower D, McKeany R, Jogia P and Jull A 2013 Evaluating the impact of implementing an early warning score system on incidence of in-hospital cardiac arrest *N.Z. Med. J.* **126** 26–34
Fairclough E, Cairns E, Hamilton J and Kelly C 2009 Evaluation of a modified early warning system for acute medical admissions and comparison with C-reactive protein/albumin ratio as a predictor of patient outcome *Clin. Med.* **9** 30–3
Feldwisch-Drentrup H, Schulze-Bonhage A, Timmer J and Schelter B 2011 Statistical validation of event predictors: a comparative study based on the field of seizure prediction *Phys. Rev.* E **83** 066704
Fullerton J N, Price C L, Silvey N E, Brace S J and Perkins G D 2012 Is the modified early warning score (MEWS) superior to clinician judgement in detecting critical illness in the pre-hospital environment? *Resuscitation* **83** 557–62
Gao H *et al* 2007 Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward *Intensive Care Med.* **33** 667–79
Gerry S, Bonnici T, Birks J, Kirtley S, Virdee P S, Watkinson P J and Collins G S 2020 Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology *Br. Med. J.* **369** m1501
Heitz C R, Gaillard J P, Blumstein H, Case D, Messick C and Miller C D 2010 Performance of the maximum modified early warning score to predict the need for higher care utilization among admitted emergency department patients *J. Hosp. Med.* **5** E46–52
Jayasundera R, Neilly M, Smith T O and Myint P K 2018 Are early warning scores useful predictors for mortality and morbidity in hospitalised acutely unwell older patients? A systematic review *J. Clin. Med.* **7** 309
Jiang X, Jiang P and Mao Y 2019 Performance of modified early warning score (MEWS) and circulation, respiration, abdomen, motor, and speech (CRAMS) score in trauma severity and in-hospital mortality prediction in multiple trauma patients: a comparison study *Peer J.* **7** e7227
Jones S, Mullally M, Ingleby S, Buist M, Bailey M and Eddleston J M 2011 Bedside electronic capture of clinical observations and automated clinical alerts to improve compliance with an early warning score protocol *Crit. Care Resusc.* **13** 83–8
Kelly C A, Upex A and Bateman D N 2004 Comparison of consciousness level assessment in the poisoned patient using the alert/verbal/painful/unresponsive scale and the Glasgow Coma Scale *Ann. Emerg. Med.* **44** 108–13
Kim W Y, Shin Y J, Lee J M, Huh J W, Koh Y, Lim C M and Hong S B 2015 Modified early warning score changes prior to cardiac arrest in general wards *PLoS One* **10** e0130523
King A, Fortino K, Stevens N, Shah S, Fortino-Mullen M and Lee I 2012 Evaluation of a smart alarm for intensive care using clinical data *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* pp 166–9
Kramer A A, Sebat F and Lissauer M 2019 A review of early warning systems for prompt detection of patients at risk for clinical decline *J. Trauma Acute Care Surg.* **87** S67–73
Kruisselbrink R *et al* 2016 Modified early warning score (MEWS) identifies critical illness among ward patients in a resource restricted setting in Kampala, Uganda: a prospective observational study *PLoS One* **11** e0151408
Kumar A, Ghabra H, Winterbottom F, Townsend M, Boysen P and Nossaman B D 2020 The modified early warning score as a predictive tool during unplanned surgical intensive care unit admission *Ochsner J.* **20** 176–81
Lee L L, Yeung K L, Lo W Y, Lau Y S, Tang S Y and Chan J T 2008 Evaluation of a simplified therapeutic intervention scoring system (TISS-28) and the modified early warning score (MEWS) in predicting physiological deterioration during inter-facility transport *Resuscitation* **76** 47–51
Linnen D T, Escobar G J, Hu X, Scruth E, Liu V and Stephens C 2019 Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: a systematic review *J. Hosp. Med.* **14** 161–9
Mathukia C, Fan W, Vadyak K, Biege C and Krishnamurthy M 2015 Modified early warning system improves patient safety and clinical outcomes in an academic community hospital *J. Community Hosp. Intern. Med. Perspect.* **5** 26716
McNeill G and Bryden D 2013 Do either early warning systems or emergency response teams improve hospital patient survival? A systematic review *Resuscitation* **84** 1652–67
Moon A, Cosgrove J F, Lea D, Fairs A and Cressey D M 2011 An eight year audit before and after the introduction of modified early warning score (MEWS) charts, of patients admitted to a tertiary referral intensive care unit after CPR *Resuscitation* **82** 150–4
Moore C, Li J, Hung C C, Downs J and Nebeker J R 2009 Predictive value of alert triggers for identification of developing adverse drug events *J. Patient Saf.* **5** 223–8
Morgan R J and Wright M M 2007 In defence of early warning scores *Br. J. Anaesth.* **99** 747–8
Robert G, Stevens A and Gabbay J 1999 'Early warning systems' for identifying new healthcare technologies *Health Technol. Assess.* **3** 1–108
Romanelli D and Farrell M W 2021 *AVPU Score 2020 May 13* (Treasure Island, FL: StatPearls)
van Rooijen C R, de Ruijter W and van Dam B 2013 Evaluation of the threshold value for the early warning score on general wards *Neth. J. Med.* **71** 38–43
Schelter B, Winterhalder M, Maiwald T, Brandt A, Schad A, Schulze-Bonhage A and Timmer J 2006 Testing statistical significance of multivariate time series analysis techniques for epileptic seizure prediction *Chaos* **16** 013108
Scully C G and Daluwatte C 2017 Evaluating performance of early warning indices to predict physiological instabilities *J. Biomed. Inform.* **75** 14–21
Snyder D E, Echauz J, Grimes D B and Litt B 2008 The statistics of a practical seizure warning system *J. Neural. Eng.* **5** 392–401

Stimpfel A W, Fletcher J and Kovner C T 2019 A comparison of scheduling, work hours, overtime, and work preferences across four cohorts of newly licensed registered nurses *J. Adv. Nurs.* **75** 1902–10

Subbe C P, Davies R G, Williams E, Rutherford P and Gemmell L 2003 Effect of introducing the modified early warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions *Anaesthesia* **58** 797–802

Subbe C P, Kruger M, Rutherford P and Gemmel L 2001 Validation of a modified early warning score in medical admissions *QJM* **94** 521–6

Tirotta D, Gambacorta M, La Regina M, Attardo T, Lo Gullo A, Panzone F, Mazzone A, Campanini M and Dentali F 2017 Evaluation of the threshold value for the modified early warning score (MEWS) in medical septic patients: a secondary analysis of an Italian multicentric prospective cohort (SNOOPII study) *QJM* **110** 369–73

Vincent J L, Einav S, Pearse R, Jaber S, Kranke P, Overdyk F J, Whitaker D K, Gordo F, Dahan A and Hoeft A 2018 Improving detection of patient deterioration in the general hospital ward environment *Eur. J. Anaesthesiol.* **35** 325–33

Xie X, Huang W, Liu Q, Tan W, Pan L, Wang L, Zhang J, Wang Y and Zeng Y 2018 Prognostic value of modified early warning score generated in a Chinese emergency department: a prospective cohort study *BMJ Open* **8** e024120

Yoon J H, Jeanselme V, Dubrawski A, Hravnak M, Pinsky M R and Clermont G 2020 Prediction of hypotension events with physiologic vital sign signatures in the intensive care unit *Crit. Care* **24** 661