Journal of Clinical Monitoring and Computing
https://doi.org/10.1007/s10877-026-01488-y

ORIGINAL RESEARCH

Transition from mandatory to discretionary nurse-measured MEWS assessment in a continuously monitored surgical ward: Diagnostic performance of wearable-derived early warning scores

EE Jerry[1,2] · MA Afonso[2] · JC De Vries[2] · A Akkermans[3] · S Turco[2] · AR Bouwman[2,3] · SW Nienhuijs[1,4]

Received: 27 January 2026 / Accepted: 11 August 2026

Abstract

Early detection of clinical deterioration on surgical wards is vital to prevent adverse outcomes. Conventional ward monitoring using the Modified Early Warning Score (MEWS) relies on intermittent nurse-measured vital signs, but its role may change when continuous wearable monitoring is already available. This study evaluated the transition from mandatory to discretionary nurse-measured MEWS assessment in a continuously monitored surgical ward and compared the diagnostic performance of two wearable-derived Early Warning Scores (EWS) with conventional MEWS. In this single-center cohort study conducted between October 2024 and June 2025, postoperative patients were continuously monitored using a wearable sensor measuring heart rate, respiratory rate, and oxygen saturation. Continuous wearable monitoring was available throughout the entire study period. During Phase 1, nurse-measured MEWS assessments were mandatory according to local protocol. During Phase 2, nurse-measured MEWS assessments were performed at staff discretion. Two wearable-derived algorithms were evaluated: the Continuous Remote EWS (CREWS), based on a previously developed algorithm, and Remote 3 EWS (R3EWS), a modified version of the hospital's standard MEWS using three parameters. The primary outcome was diagnostic accuracy for clinical deterioration, defined as postoperative complications graded Clavien–Dindo≥II. A total of 544 admissions (516 patients) were included, of which 98 admissions (18%) experienced clinical deterioration (comprising 154 clinical deterioration events). Following the transition from mandatory to discretionary nurse-measured MEWS assessment, intermittent nurse-measured vital signs were performed less frequently, while MEWS≥3 occurred more often, resulting in higher MEWS sensitivity and lower specificity. In the separate diagnostic performance analysis, R3EWS showed slightly higher AUROC values (0.61) compared with MEWS (0.58) and CREWS (0.59), although differences were modest. Wearable-derived alarm frequency increased 6–8 h before clinical deterioration. In a surgical ward where continuous wearable monitoring was already available, transition from mandatory to discretionary nurse-measured MEWS assessment was associated with fewer intermittent observations, higher MEWS sensitivity, and lower specificity. Wearable-derived EWS algorithms showed comparable but modest discriminative performance relative to conventional MEWS. The increased number of alerts highlights the need for prospective validation and clear protocols for integrating continuous monitoring data into ward escalation workflows.

Keywords Continuous monitoring · Wearable sensor · Early warning score · Vital parameters · Surgical ward

✉ EE Jerry
ephrahim.jerry@catharinaziekenhuis.nl

[1] Department of Surgery, Catharina Hospital Eindhoven, Michelangelolaan 2, Eindhoven 5623 EJ, The Netherlands

[2] Department of Electrical Engineering, Eindhoven University of Technology, Groene Loper 19, Eindhoven 5612 AP, The Netherlands

[3] Department of Anesthesiology, Intensive Care and Pain Medicine, Catharina Hospital Eindhoven, Eindhoven 5623 EJ, The Netherlands

[4] NUTRIM (School for Nutrition and Translational Research in Metabolism), Universiteitssingel 40, Maastricht 6229 ER, The Netherlands

Published online: 29 August 2026

1 Introduction

Early detection of clinical deterioration on surgical wards is crucial to prevent adverse outcomes such as unplanned intensive care (ICU) admissions, cardiac arrest, or death. Conventional ward monitoring is based on the Modified Early Warning Score (MEWS) assessment [1], which relies on intermittent nurse-measured vital signs performed two to three times daily by nursing staff. Although MEWS is widely implemented, previous studies have reported only modest predictive performance for clinical deterioration, indicating limited sensitivity in detecting subtle but clinically relevant changes in a patient's condition [2–5–4].

Continuous monitoring using wearable technology has emerged as a promising strategy to complement intermittent ward observations [6, 7]. By providing repeated or continuous measurements of vital parameters, wearable sensors may support earlier recognition of clinical deterioration and provide additional information between scheduled nurse-measured observations. However, the availability of continuous monitoring also raises an important workflow question: how should conventional nurse-measured MEWS assessments be adapted when continuous vital sign data are already available to patient-care staff?

Despite these promising results, large-scale adoption of wearable monitoring in surgical care remains limited. Barriers include workflow integration, trust in device accuracy, fear of a high rate of false alarms, and costs, as well as a lack of clear protocols for handling continuous data [8–10]. Implementation studies have shown that successful integration depends on usability, staff acceptance, and embedding monitoring into daily routines [11–13]. Most previous devices, however, were patch-based and limited to measuring heart and respiratory rate, without incorporating other vital signs of MEWS, such as oxygen saturation. In this context, the clinical effect of continuous monitoring cannot be evaluated solely by comparing algorithms; it also depends on how monitoring data are embedded into existing nursing routines and escalation protocols.

Recent developments include upper-arm wearable sensors based on photoplethysmography (PPG), enabling continuous monitoring of multiple physiological parameters, including oxygen saturation (SpO_2). The addition of SpO_2 may provide complementary information, particularly in surgical patients at risk of respiratory complications, although the incremental diagnostic value remains to be established [14, 15]. The reusable armband form factor may facilitate longer monitoring periods compared with adhesive patch devices, but its diagnostic performance and role within routine ward monitoring protocols require further evaluation.

In the present study, continuous wearable monitoring was available and visible to clinical staff throughout the entire study period. During this period, the ward monitoring protocol changed from mandatory intermittent nurse-measured MEWS assessments to MEWS assessments performed at staff discretion. Therefore, the first aim of this study was to evaluate the impact of this transition on the frequency and diagnostic performance of nurse-measured MEWS assessments. The second aim was to compare the diagnostic performance of two wearable-derived early warning scores, CREWS and R3EWS, with conventional nurse-measured MEWS for detecting clinical deterioration on a surgical ward.

We hypothesized that wearable-derived EWS algorithms would demonstrate comparable diagnostic performance to conventional MEWS, while allowing more frequent assessment of physiological trends.

2 Methods

2.1 Study design and setting

A retrospective analysis of a prospective, single-center observational study was conducted at the Department of Surgery, Catharina Hospital Eindhoven, the Netherlands. The study population included all patients admitted to the general surgical ward between October 2024 and June 2025. This ward admits patients from non-malignant surgical specialties, including general surgery, urology, and gynaecology.

The study period was divided into two phases. During Phase 1 (October 2024 to March 2025), continuous wearable monitoring was available and visible to clinical staff, while routine nurse-measured vital signs and corresponding MEWS assessments remained mandatory according to local protocol. During Phase 2 (March 2025 to June 2025), continuous wearable monitoring remained available and visible to clinical staff, but routine nurse-measured vital signs and MEWS assessments were no longer mandatory and were performed at staff discretion based on clinical judgement. Thus, the principal difference between phases concerned the protocol for nurse-measured MEWS assessment rather than the availability of continuous monitoring.

Continuous wearable monitoring had previously been introduced on the ward as part of the REQUEST implementation study and remained available throughout both study phases [10].

The duration of each phase was determined by real-world clinical implementation rather than a predefined study design.

Importantly, no automated alerting or notification system was implemented during either phase. Continuous monitoring data and derived scores were displayed in the electronic health record (EHR) during both study phases, where nursing staff reviewed these values during routine clinical workflow. When an elevated score (≥3) was identified, nurses performed a manual vital sign assessment with corresponding MEWS calculation or contacted the responsible physician based on clinical judgement.

The study was approved by the Medical Research Ethics Committee United (reference number W24.065).

2.2 Study population and data collection

All admitted patients were screened for availability of wearable monitoring data. Patients for whom at least one valid data sample from each vital sign was available from the wearable device were included in the analysis. For admissions during Phase 1, patients were additionally required to have at least one set of intermittent nurse-measured vital signs in the electronic health records (EHR). Exclusion criteria were technical device failure or the inability to provide valid wearable data. We analysed admissions rather than patients, as multiple admissions from the same patient were included when all required data were available. Continuous monitoring data were obtained from the hospital system, including heart rate, respiratory rate and SpO_2, as measured by the wearable device. In addition, all routine intermittent nurse-measured vital signs and calculations of EWS scores from medical professionals were extracted from the EHR.

Continuous monitoring data were recorded in the electronic health record (EHR) and were visible to clinical staff during both phases; no blinding was applied.

Intermittent nurse-measured vital signs consisted of bedside measurements of physiological parameters (including heart rate, respiratory rate, blood pressure, temperature, and oxygen saturation). MEWS scores were calculated based on these measurements when documented in the EHR; however, not all measurement sets resulted in a recorded MEWS score.

To capture clinical outcomes and clinical deterioration, all patient records were manually reviewed. Complications were initially identified and recorded by one researcher and subsequently verified by a second independent researcher. Any discrepancies in classification or timing were resolved through consensus. The timing of clinical deterioration was defined as the moment of clinical confirmation of the complication, based on documentation in the medical record (e.g., diagnostic confirmation, initiation of treatment, or documented clinical diagnosis). For the purpose of this study, clinical deterioration was defined as postoperative complications graded as Clavien–Dindo ≥ II [16].

Table 1 CREWS computation method*

Vital Parameter	Score		
	0	1	2
Heart Rate (bpm)	40–110		<40 or >110
Respiration Rate (bpm)	8–20	<8 or >20	
SpO_2 (%)	≥90		<90

*Based on the work by Van Der Stam et al. (9)

2.3 Wearable device

The wearable sensor (viQtor, SmartQare, Eindhoven, the Netherlands) is a CE-marked, reusable armband worn on the upper arm. It continuously measures heart rate, respiratory rate and SpO_2 through PPG, at a sampling frequency of one sample per minute. It transmits data wirelessly to the hospital system at 1-minute-wide intervals; 4-hour medians are computed from these vitals and displayed within the electronic health record (EHR) and reviewed by nursing staff during routine workflow (approximately twice per shift), without automated notifications or mandatory alerts. This device has been evaluated in post-operative patients, demonstrating acceptable agreement for heart rate measurements, while agreement for respiratory rate and oxygen saturation showed greater variability [17].

2.4 Remote early warning score

In this study, the term "early warning score (EWS)" refers specifically to structured scoring systems based on vital sign measurements, such as MEWS or wearable-derived scoring algorithms, and not to the measurement of vital signs themselves.

To evaluate how different early warning score algorithms performed within a ward where continuous wearable monitoring was already available, two wearable-derived EWS calculation methods were assessed.

The first method, Continuous Remote Early Warning Score (CREWS, Table 1) [18], has been previously published and is based on continuous measurements of heart rate and respiratory rate using wearable sensors. As this wearable device also included SpO2, this parameter was added into the table. The original CREWS algorithm, developed by van Ede et al., relied on these two parameters without incorporating oxygen saturation (SpO_2). In this study, as the wearable device also included SpO_2, this parameter was added to the algorithm. The original CREWS was developed to improve early detection of clinical deterioration and reduce false alarms, with a cut-off of 2.2. With the addition of SpO_2, the cut-off was set at 3 in this study.

The second method, Remote MEWS-based Continuous score (R3EWS, Table 2), is a novel approach designed as an extension of the conventional Modified Early Warning

Table 2 R3EWS computation method

Vital Parameter	Score			
	0	1	2	3
Heart Rate (bpm)	51–100	40–50 or 101–110	<40 or 111–130	>130
Respiration Rate (bpm)	9–14	15–20	<8 or 21–30	>30
SpO_2 (%)	≥90			<90

Score (MEWS), adapted for use with the wearable device. R3EWS includes, just like CREWS, only three vital parameters measured by the wearable sensor. However, it aligns the scoring process more closely with the original EWS table to improve usability and reduce alarm fatigue. This method represents a further development of the CREWS system, aiming to provide a more streamlined approach for continuous monitoring in clinical practice.

The main differences between the two methods are how they handle serious cases (with a lower penalty in the CREWS protocol, which assigns a score of 2 instead of 3) and how small deviations in vital signs are treated (with R3EWS assigning a partial score greater than 0 for small deviations, while CREWS assigns a score of 0). This creates higher scores for the R3EWS both in cases where there are small or big vital sign deviations from normality. Regardless, it was defined that both methods would provide an alarm at a cumulative EWS score equal or higher than 3, reflecting the predefined clinical escalation threshold used in local ward protocols. This threshold was selected to align with routine clinical decision-making rather than to represent a universally validated definition of clinically meaningful deterioration. Individual parameter scores did not independently trigger alerts; instead, alarms were based solely on the cumulative score. As a result, an isolated abnormal value (e.g., an elevated respiratory rate resulting in a score of 2 in R3EWS) would not trigger an alert unless additional abnormalities contributed to reaching the cumulative threshold.

2.5 Data processing

Continuous data were sampled at one-minute intervals. When multiple samples occurred within the same minute, their mean value was used. Most recent manufacturer's algorithm and daylight-saving time adjustments were harmonized to ensure temporal consistency across datasets.

2.6 Data preparation

Continuous monitoring data were analyzed from the time of admission (or 00:01 on 23rd October 2024, the start of Phase 1, whichever came later) until discharge (or 23:59 on 30th June 2025, the end of Phase 2, whichever came first). The patient's stay period was then divided into 4-hour epochs to mirror clinical practice. Median values of heart rate, respiratory rate and SpO2 were computed for each epoch (provided that at least one sample was available in those 4 h) with the present samples. If no sample was accessible for a certain vital sign for the 4-hour period, no median would be computed for that vital. In the case the patient's stay was not exactly divisible into 4-hour epochs, the last epoch was deemed as the remainder of samples. The 4-hour median vital signs were then used to calculate R3EWS and CREWS scores. As nurses would check the EWS score at the end of the 4-hour period, during both phases, the time of measurement (comparable to the time of intermittent manual measurement) was considered to be the time of the last sample in the 4-hour epoch. In the event of at least one vital sign median missing, no R3EWS or CREWS would be computed, but a missing score would still be stored, as a missing score would also occur in real-world scenarios and is important to be evaluated.

Furthermore, intermittent nurse-measured vital signs recorded after the protocol change may reflect measurements performed in response to clinical concern. Therefore, to allow consistent comparison between MEWS, CREWS, and R3EWS, only MEWS scores derived from intermittent nurse-measured vital signs from which MEWS scores were calculated during Phase 1 were compared to the scores obtained with the wearable device (during both phases). All manual intermittent nurse-measured vital signs from Phase 2 were only used for descriptive analyses. Although the individual vital signs extracted during nurse-measured vital signs were available even when nurses did not compute the MEWS score manually, we only took into account the MEWS values available in the EHR, as these also encompass the non-vital signs parameters.

2.7 Prediction of clinical deterioration

The predictive ability of each intermittent manual measurement and each ViQtor epoch was assessed within the 24 h preceding clinical deterioration. All measurements, including those occurring after clinical deterioration, were retained to reflect real-world clinical conditions, acknowledging that multiple events of clinical deterioration may occur within short intervals. An EWS≥3 within 24 h before a complication was classified as a true positive; absence of such an alarm was considered a false negative, and alarms outside this window were defined as false positives.

In addition, "complications captured" was defined as the number and proportion of clinical deterioration events for which at least one alarm (score≥3) occurred within the 24-hour window preceding the documented deterioration.

2.8 Time analysis of predictive power of EWS

Clinical deterioration is commonly predicted in the 24 h before its onset, using the deviations in vital sign parameters (24). As a verification experiment, we analyzed the percentage of 1-minute samples within each hour that triggered an EWS score alarm according to each of the EWS computation methods. This analysis was performed for the 24 h leading up to a clinical deterioration, if such data were available. If the data for the full 24-hour period were not available, the analysis was instead conducted for the time period between admission and the deterioration event.

2.9 Statistical analysis

Baseline characteristics and clinical outcomes were summarized using descriptive statistics. The number of intermittent nurse-measured vital signs, EWS scores, length of stay and demographic characteristics between Phases 1 and 2 were compared using the Mann-Whitney U-test for independent non-parameterized data. The number of intermittent nurse-measured vital signs per admission day was calculated by dividing the total number of manually recorded measurement sets by the length of stay (in days). Values below 1.0 therefore reflect patients with relatively few measurements over longer admissions (e.g., one measurement over multiple days), rather than fractional measurements.

Categorical demographic data was compared using the Chi Square Test for Independence. Sensitivity and Specificity for MEWS between Phase 1 and 2 were compared using Fisher's Test. Then, the discriminative power of EWS obtained by MEWS, R3EWS and CREWS to predict clinical deterioration within 24 h was assessed using the receiver operating characteristic (ROC) curves, with calculation of the area under the curve (AUROC) and 95% confidence intervals, using 1000 stratified bootstrap replicates. ROC analyses were performed to evaluate model performance across the full range of score values, independent of a single predefined threshold. Precision-Recall curves were also performed and analyzed as the output is expected to be imbalanced (there are few complications). The area under the precision-recall curve (AUPRC) 95% confidence intervals were also calculated using 1000 stratified bootstrap replicates. Additionally, diagnostic performance was further evaluated by calculating sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) at predefined thresholds (MEWS$\geq$3; R3EWS$\geq$3 and CREWS$\geq$3). AUROCs were compared using DeLong's test for paired data between the R3EWS and the CREWS and using the Mann-Whitney U-Test between the MEWS and both continuous monitoring tests, as the number of available intermittent nurse-measured vital signs for the manual and continuous measurements were different. AUPRC were compared using the Bootstrap Permutation Test and the Mann-Whitney U-test. Finally, optimal thresholds for the CREWS and R3EWS were obtained by matching the percentage of true positives obtained with the MEWS during the phase with mandatory intermittent nurse-measured vital signs. Because continuous wearable monitoring was available throughout both phases, comparisons between Phase 1 and Phase 2 should be interpreted as reflecting the combined effects of a change in MEWS assessment protocol and the presence of continuous monitoring within routine care. A p-value of 0.05 was considered to be significant. All analyses were performed using MATLAB (R2024b; The MathWorks, Inc. Natick, MA, USA).

3 Results

In total, 621 admissions were extracted from the EHR. Of these, 544 admissions corresponding to 516 unique patients were included in the final analysis, based on the availability of both wearable monitoring data and manual measurements. The baseline characteristics, as well as clinical deterioration outcomes for all phases considered are available in Table 3. Patients were admitted under 14 different specialties, most commonly general surgery (48%), urology (28%), and gynecology (8%). Clinical deterioration occurred in 28% of admissions, most frequently classified as Clavien–Dindo grade II.

3.1 Comparison between clinical deterioration and no deterioration

In the whole analysis, patients who experienced clinical deterioration had a longer median length of stay (9, IQR [6–10, 19–24]days) than patients who did not have clinical deterioration (2, IQR [1–19] days, $p < 0.001$). This trend was the same regardless of the phase of analysis.

3.2 Comparison between MEWS assessment protocols

Furthermore, the same characteristics were compared between the 2 phases. There was no significant difference in age. In total, 1,069 intermittent nurse-measured vital signs were recorded during Phase 1 and 956 during Phase 2 with the overall difference in frequency between phases being statistically significant ($p < 0.01$). The median length of stay was lower in Phase 2 (3 [2–21] days for Phase 1 and 3 [1–19] days for Phase 2, $p < 0.001$).

Following the transition from mandatory to discretionary nurse-measured MEWS assessment, intermittent

Table 3 Baseline characteristics

Variable	All admissions	Phase 1	Phase 2	*P*-value Phase 1 vs. Phase 2
Number of unique admissions	544	181	363	
Age, years (median [IQR range])	65 [52–77]	66 [52–77]	65 [51–76]	0.367
Female sex, n (%)	255 (46.9%)	86 (47.5%)	164 (46.3%)	0.795
Number of surgeries, n (% admissions with surgery)	426 (65.4%)	146(83.6%)	280 (64.5%)	
Length of stay, days (median [IQR range])	3 [1–6]	3[2–7]	3 [1–5]	<0.001
– Deteriorating patients	9[5–15]	7 [3–11]	9 [5–21]	0.074
– Non-deteriorating patients	2[1–5]	3 [2 - 5.75]	2 [1–4]	<0.001
Clinical deterioration events, n (%)	154 (28.3%)	49 (27.1%)	94 (25.9%)	0.067
– Clavien–Dindo II	87 (56.5%)	32 (65.3%)	48 (51.1%)	
– Clavien–Dindo IIIa	47 (30.5%)	9 (18.4%)	35 (37.2%)	
– Clavien–Dindo IIIb	18 (11.7%)	8 (16.3%)	9 (9.6%)	
– Clavien–Dindo IV	2 (1.3%)	0 (0%)	2 (2.1%)	
Nurse-measured vital signs per admission day (median [IQR range])	0.75 [0.5–1.25]	1 [0.67–1.5]	0.5 [0.29–1]	<0.01
Nurse-measured vital signs triggering MEWS≥3, n (%)	262 (12.9%)	63 (5.9%)	199 (20.8%)	<0.001
Sensitivity of MEWS (≥3)	0.250	0.122	0.378	<0.001
Specificity of MEWS (≥3)	0.877	0.947	0.809	<0.001

nurse-measured vital signs were performed less frequently in Phase 2 than in Phase 1 ($p<0.01$). The proportion of nurse-measured vital sign sets triggering a MEWS≥3 was higher during Phase 2 than Phase 1 (20.8% vs. 5.9%), corresponding to higher MEWS sensitivity (0.38 vs. 0.12, $p<0.001$) and lower specificity (0.81 vs. 0.95, $p<0.001$). These findings should be interpreted in relation to the change in MEWS assessment protocol, as continuous wearable monitoring was available during both phases.

3.3 Diagnostic performance of EWS methods

In a separate analysis, the diagnostic accuracy of nurse-measured MEWS, R3EWS, and CREWS for predicting clinical deterioration within 24 h was assessed and is summarized in Table 4.

Receiver operating characteristic (ROC) curves showed that R3EWS showed a slightly higher AUROC (AUROC 0.61, 95% CI [0.58–0.63]), values compared to MEWS (0.58, 95% CI [0.52–0.64], $p<0.01$) and CREWS (0.58, 95% CI [0.56–0.60], $p<0.05$) (Figs. 2 and 3). When stratifying by phases and analyzing Phase 1, AUROC for R3EWS is significantly larger than that of CREWS (0.58 with 95% CI [0.54–0.63] and 0.54 with 95% CI [0.50–0.58], $p<0.01$). During Phase 2, AUROC values for R3EWS (0.62 (95% CI [0.59–0.65])) were numerically higher when compared to CREWS (0.59 (95% CI [0.56–0.62], $p<0.01$)).

Precision-recall curves (PRC) demonstrated that MEWS had a higher AUPRC (0.12, 95% CI[0.09–0.17], $p>0.05$) than CREWS (0.09, 95%[0.07–0.10]) and R3EWS (0.08, 95%[0.07–0.10]).

At the predefined threshold of ≥3, R3EWS captured a higher proportion of clinical deterioration events within 24 h (18%) than MEWS (13%) or CREWS (9%), but this was accompanied by a higher total number of alerts. The R3EWS also achieved higher sensitivity (0.16) with acceptable specificity (0.93).

The CREWS algorithm produced fewer alarms and showed the highest specificity (0.98) but the lowest sensitivity (0.05), identifying fewer true clinical deterioration events.

The total number of alarms generated by R3EWS and CREWS was substantially higher compared to MEWS.

3.4 Wearable-derived EWS performance across MEWS assessment protocol phases

In order to understand how each phase and the change in respective hospital protocols influenced the overall performance of CREWS and R3EWS, a performance analysis of both models was made for each phase (Tables 7 and 8, Supplementary Material). In Phase 1, the use of wearable-derived 4-hour EWS epochs provided 121% more available EWS data than nurse-measured MEWS assessments, increasing from 1069 nurse-measured vital sign sets to 2372 wearable-derived measurements. Regarding the continuous methods performance, both methods performed worse in terms of sensitivity (0.117 for R3EWS and 0.031 for CREWS) than the MEWS (0.122). Nonetheless, during Phase 1, R3EWS was able to predict the same number of clinical deteriorations as MEWS (6 (12.5%)).

Table 4 Statistics on nurse-measured vital signs and predictive power of the different protocols

Variable	MEWS *		R3EWS		CREWS	
Cut-off	<3	≥3	<3	≥3	<3	≥3
Total Number of alarms	1006	63	7994	667	8487	174
True Positive Alarms	79	11	391	72	439	24
*Complications captured***	-	6 (12.5%)	-	27 (17.6%)	-	14 (9.2%)
Complications with no data 24 h prior		11 (25%)	-	58 (37.9%)		58 (37.9%)
Complications with data but no alarm	-	30 (62.5%)	-	68 (44.5%)		81 (52.9%)
Performance						
AUROC [95% CI]	0.58 [0.52–0.64]		0.61 [0.58–0.63]		0.58 [0.56–0.60]	
AUPRC [95%CI]	0.12 [0.09–0.17]		0.08 [0.07–0.10]		0.09 [0.07–0.10]	
Sensitivity	0.122		0.155		0.051	
Specificity	0.947		0.927		0.982	
PPV	0.175		0.107		0.138	
NPV	0.921		0.951		0.948	

*The MEWS period corresponds only to the period where nurse-measured vital signs were mandatory or Phase 1

** Complications captured (number and proportion of clinical deterioration events preceded by ≥1 alarm within 24 h)

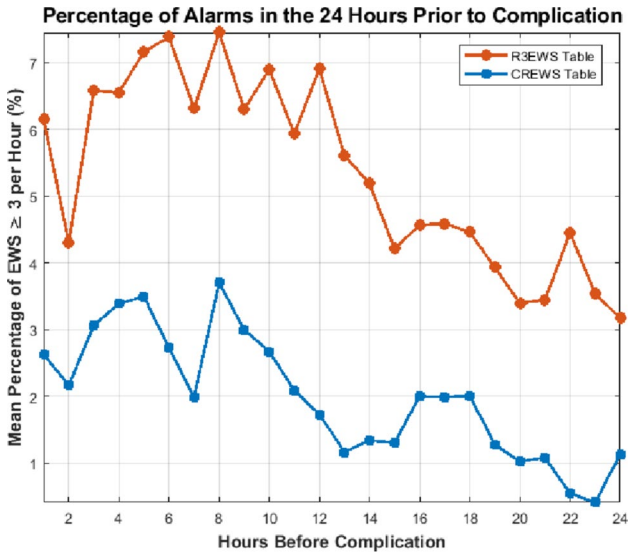

Fig. 1 Percentage of alarms in the 24 h prior to a complication

During Phase 2, when nurse-measured MEWS assessments were discretionary, the percentage of alarming measurements was higher for R3EWS. This was not the case for the CREWS, in which the percentage of alarming measurements was lower (2.1%) than MEWS in Phase 1. In percentage, the clinical deteriorations predicted with the R3EWS were almost twice as many as those anticipated with the MEWS (20.2% vs. 12.5%), whereas for the CREWS, the percentage was comparable (10.6% vs. 12.5%). The sensitivity performance was similar across the two periods of analysis, where the R3EWS (0.172) outperformed both intermittent nurse-measured vital signs derived data (0.120) and the CREWS (0.061) method.

3.5 Temporal trends preceding clinical deterioration

Temporal trends preceding clinical deterioration were explored in an exploratory analysis of alarm patterns. In the 24 h preceding clinical deterioration, the proportion of EWS≥3 alarms increased steadily, with distinct peaks approximately 8 and 6 h before the event (Fig. 1).

As MEWS≥3 served as the clinical reference threshold, an exploratory analysis was conducted to determine the optimal cutoff values for continuous algorithms. When matching the proportion of true positive alarms to MEWS (11%), the optimal threshold was approximately 2.8 for R3EWS and around 1.0 for CREWS.

4 Discussion

This study evaluated two related but distinct questions in a surgical ward where continuous wearable monitoring was available throughout the study period. First, we assessed the transition from mandatory nurse-measured MEWS assessments to MEWS assessments performed at staff discretion. This protocol change was associated with a lower frequency of intermittent nurse-measured vital signs, a higher proportion of MEWS≥3 measurements, higher MEWS sensitivity, and lower MEWS specificity. Second, we compared the diagnostic performance of two wearable-derived early warning scores, CREWS and R3EWS, with conventional nurse-measured MEWS. Overall discriminative performance was modest for all three methods, with R3EWS showing slightly higher AUROC and sensitivity than CREWS, but at the cost of more alerts.

The between-phase differences should primarily be interpreted in relation to the change in MEWS assessment

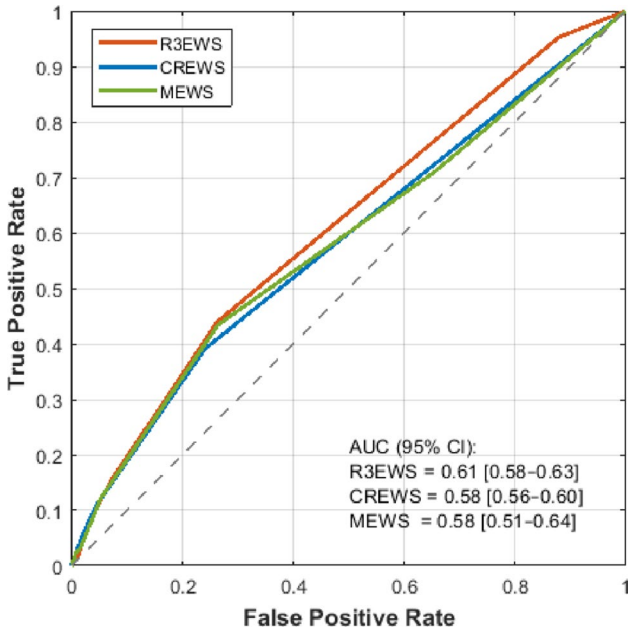

Fig. 2 ROC curve for each of the 3 protocols

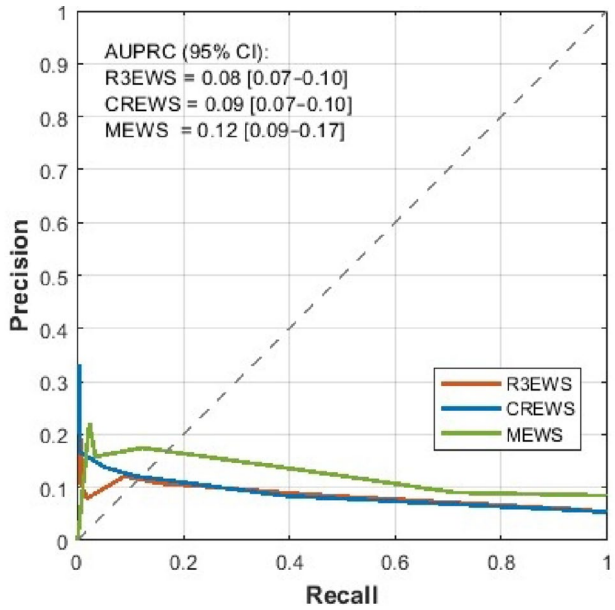

Fig. 3 Precision-Recall curve (PRC for each of the 3 protocols

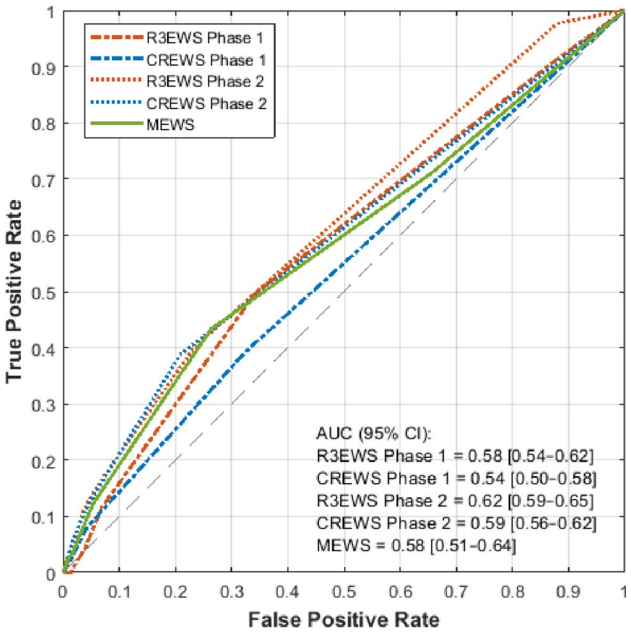

Fig. 4 ROC curves for the 3 EWS computation methods and separated by the 2 periods (with and without mandatory manual measurements)

protocol rather than the introduction of continuous monitoring, since wearable-derived data were available and visible to staff during both phases. The transition from scheduled mandatory observations to discretionary assessment likely changed the clinical context in which nurse-measured MEWS was performed. In Phase 2, manual measurements may have been more likely to occur in response to clinical concern, which may explain the higher proportion of MEWS≥3 measurements, increased sensitivity, and reduced specificity. Therefore, the observed reduction in intermittent observations cannot be attributed solely to the wearable monitoring system.

In the separate diagnostic performance analysis, R3EWS demonstrated slightly higher sensitivity and AUROC than CREWS, although differences were small and overall discrimination remained modest. Fig. 2, 3, 4. The higher sensitivity of R3EWS was accompanied by a greater number of alerts, reflecting the lower threshold for smaller deviations in vital signs. In contrast, CREWS generated fewer alerts and achieved higher specificity, but captured fewer clinical deterioration events. These findings illustrate the trade-off between sensitivity and alert burden that is inherent to continuous early warning score algorithms [6, 14, 22].

Wearable-derived alarms increased in the hours preceding clinical deterioration, with peaks approximately 6–8 h before the event. This exploratory temporal pattern suggests that continuous physiological data may contain clinically relevant information before documented deterioration [11, 12]. However, because this analysis was retrospective and did not include a matched control group without deterioration, these findings should be interpreted as hypothesis-generating.

From a clinical perspective, these findings suggest that continuous wearable monitoring may support a more selective approach to nurse-measured MEWS assessment, but does not eliminate the need for clear escalation protocols. When nurse-measured MEWS assessments are performed at staff discretion, the diagnostic characteristics of MEWS change because measurements may become more clinically targeted. Future implementation should therefore define

when manual reassessment is required, how wearable-derived scores should be interpreted, and what alert threshold is acceptable within local workflow [8–10].

In the discussion of our results, we observed that when matching the proportion of true positive alarms to MEWS (11%), the optimal threshold was approximately 2.8 for R3EWS and around 1.0 for CREWS. These findings suggest that continuous scoring systems may require lower alarm thresholds than conventional nurse-measured vital signs–based EWS to achieve similar sensitivity. This difference may be related to more continuous monitoring but also because of the reduced number of parameters measured in continuous systems, compared to the more comprehensive data collected through intermittent nurse-measured vital signs.

A strength of this study is the inclusion of both ROC and Precision–Recall analyses. Given the relatively low prevalence of clinical deterioration events, AUPRC provided additional insight into model performance under class-imbalanced conditions.

This study has several limitations. The main limitation of this study is the before-and-after design. Because continuous wearable monitoring was available throughout both phases, this study cannot disentangle the independent effect of continuous monitoring from the effect of changing the nurse-measured MEWS assessment protocol. The observed between-phase differences should therefore not be interpreted as being caused by the implementation of continuous monitoring alone, but rather as reflecting a protocol change within an already continuously monitored ward. It was conducted in a single center and on one surgical ward, which limits generalizability. Missing data and non-compliance, particularly early in implementation, may have reduced sensitivity. Post-event EWS score calculations were not excluded; although these alarms do not influence sensitivity estimates, they may have inflated the total number of alarms and slightly reduced specificity. However, this choice reflects a more realistic scenario of what occurs in the ward, where medical professionals have access to information even after clinical deterioration occurs, and decide for themselves whether there is cause for concern. The relatively small number of clinical deterioration events also limited statistical power for subgroup analyses. Furthermore, both the R3EWS and the CREWS were applied retrospectively; prospective validation is required to confirm their predictive performance in real-time clinical practice. In addition, the modest AUROC values observed in this study indicate limited discriminative ability of all evaluated methods. Also, no interrupted time series analysis was done and no differentiation was performed in relation to time of year, despite this having an influence on ward staff and hospital capacity. In addition, the use of a single predefined threshold (EWS$\geq$3) may oversimplify the relationship between score values and clinical deterioration. Although this threshold reflects routine clinical practice, it may not represent the optimal cutoff for prediction. Importantly, ROC analyses performed in this study evaluate performance across the full range of score values, partially addressing this limitation.

Lastly, the timing of clinical deterioration was defined based on the moment of clinical confirmation rather than the onset of physiological changes. This may introduce temporal bias, as physiological deterioration may precede formal diagnosis. Furthermore, the temporal analysis of alarm patterns preceding clinical deterioration was exploratory and did not include a matched control group without clinical deterioration. As a result, potential circadian effects or time-dependent biases cannot be excluded.

Future research should focus on prospective, multicenter evaluations of continuous early warning algorithms embedded within clinical workflows, particularly in settings where such systems have already become standard of care. This will provide large datasets that reflect real-world usage. A stepped-wedge design could be considered to account for the time required for staff to become fully acquainted with the system, ensuring a more accurate assessment of its impact on clinical outcomes. Additionally, future studies should further explore optimal threshold selection and dynamic risk stratification across the full range of early warning scores.

In conclusion, in a surgical ward where continuous wearable monitoring was already available, transition from mandatory to discretionary nurse-measured MEWS assessment was associated with fewer intermittent observations, higher MEWS sensitivity, and lower MEWS specificity. In a separate diagnostic performance analysis, wearable-derived EWS algorithms showed modest discriminative performance comparable to conventional MEWS. R3EWS captured more clinical deterioration events than CREWS, but generated more alerts. These findings highlight the need for prospective validation of wearable-derived EWS algorithms and clear protocols for integrating continuous monitoring data into ward escalation workflows.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10877-026-01488-y.

Author contributions E.E. and M.D. contributed equally as first authors. They jointly developed the concept, collected the data, and performed the analyses. A.R.B. and S.W.N. conceived the project, supervised the study, and were involved in the overall progress of the research. S.P. and T.M. provided feedback, participated in data collection, and contributed to the completion of the study.All authors reviewed the manuscript. Authors contributed equally.

Funding No external funding was received for this research.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Leidraad-Vitaal bedreigde patiënt. (n.d.). https://richtlijnendatabase.nl/richtlijn/leidraad_-_vitaal_bedreigde_patient/signaleren_van_de_vitaal_bedreigde_patient.html
2. Ludikhuize J, de Jonge E, Goossens A, Netelenbos T, de Rooij S. Identification of deteriorating patients on general wards: Measurement of vital parameters and the Modified Early Warning Score. J Crit Care. 2012;27(4):e4247–42413. https://doi.org/10.1016/j.jcrc.2012.02.004.
3. 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. BMJ. 2020;369:m1501. https://doi.org/10.1136/bmj.m1501.
4. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation. 2013;84(4):465–70. https://doi.org/10.1016/j.resuscitation.2012.12.016.
5. Hall KK, Shoemaker-Hunt S, Hoffman L, Richard S, Gall E, Schoyer E, Costar D, Gale B, Schiff G, Mardon R. Failure to rescue. In *Making Healthcare Safer III: A Critical Analysis of Existing and Emerging Patient Safety Practices*. Agency for Healthcare Research and Quality. 2020. https://www.ncbi.nlm.nih.gov/books/NBK555513/
6. Eddahchouri Y, Pouwels R, Koeneman M, Touw HR, van Goor H, Bredie SJH, et al. Effect of continuous wireless vital sign monitoring on unplanned ICU admissions and rapid response team calls: A before-and-after study. Anaesthesia. 2022;128(5):857–63. https://doi.org/10.1111/anae.15694.
7. Dykes PC, Lipsitz S, Salvucci SM, Yoon C, Bates DW, et al. Reducing ICU utilization, length of stay, and cost by optimizing the clinical use of continuous monitoring system technology in the hospital. Am J Med. 2021;135(3):337–41. https://doi.org/10.1016/j.amjmed.2021.11.022.
8. Leenen JPL, van den Hoogen A, Patijn GA, Schoonhoven L, Kalkman CJ. Nurses' experiences with continuous vital sign monitoring on the general surgical ward: A qualitative study based on the Behaviour Change Wheel. BMC Nurs. 2022;21(1):60. https://doi.org/10.1186/s12912-022-00859-8.
9. Gijsbers HJH, van der Stam JA, Scheerhoorn J, Nienhuijs SW. Upscaling telemonitoring in Dutch hospitals: A baseline measurement. Int J Med Informatics. 2023;175:105085. https://doi.org/10.1016/j.ijmedinf.2023.105085.
10. Jerry EE, Bouwman ARA, van Esch EMG, Richardson R, Nienhuijs SW. Remote Monitoring With a Reusable Device Upon Implementation on a Surgical Department (REQUEST Trial). World J Surg. 2026;50(3):676–83. https://doi.org/10.1002/wjs.70261.
11. Taenzer AH, Pyke JB, McGrath SP, Blike GT. Impact of pulse oximetry surveillance on rescue events and ICU transfers. Anesthesiology. 2010;112(2):282–7. https://doi.org/10.1097/ALN.0b013e3181ca7a9b.
12. Elvekjaer M, Rasmussen S, Grønbæk KK, Porsbjerg CM, Jensen J-U, Haahr-Raunkjær C, et al. Clinical impact of vital sign abnormalities in patients admitted with acute exacerbation of chronic obstructive pulmonary disease: An observational study using continuous wireless monitoring. Intern Emerg Med. 2022;17(6):1689–98. https://doi.org/10.1007/s11739-022-02963-9.
13. Stehlik J, Schmalfuss C, Bozkurt B, Nativi-Nicolau J, Wohlfahrt P, Wegerich S, et al. Continuous wearable monitoring analytics predict heart failure hospitalization. Circulation: Heart Fail. 2020;13(3):e006513. https://doi.org/10.1161/CIRCHEARTFAILURE.119.006513.
14. Breteler MJM, KleinJan EJ, Dohmen DAJ, Leenen LPH, van Hillegersberg R, Ruurda JP, et al. Vital signs monitoring with wearable sensors in high-risk surgical patients: A clinical validation study. Anesthesiology. 2020;132(3):424–39. https://doi.org/10.1097/ALN.0000000000003029.
15. Taenzer AH, Blike GT, McGrath SP. Defining the ideal early warning score for continuous monitoring of general ward patients. J Crit Care. 2011;26(1):9–10. https://doi.org/10.1016/j.jcrc.2010.08.004.
16. Dindo D, et al. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. Ann Surg. 2004;240:205e13.
17. Reijmers N, van Kootwijk A, de Waal EEC. Accuracy of vital sign monitoring using a photoplethysmography upper arm wearable device in postoperative non-cardiac surgery patients: a prospective observational clinical validation study. J Clin Monit Comput. 2025. https://doi.org/10.1007/s10877-025-01358-z.
18. Van Ede ES, Scheerhoorn J, Bonomi AG, Buise MP, Bouwman RA, Nienhuijs SW. Continuous remote monitoring in post–bariatric surgery patients: development of an early warning protocol. Surg Obes Relat Dis. 2022;18(11):1298–303. https://doi.org/10.1016/j.soard.2022.06.018.
19. Joshi M, Ashrafian H, Aufegger L, Khan S, Arora S, Cooke G, Darzi A. Wearable sensors to improve detection of patient deterioration. Expert Rev Med Dev. 2019;16(2):145–54. https://doi.org/10.1080/17434440.2019.1567821.
20. Weenk M, Bredie SJH, Koeneman M, Hesselink G, van Goor H, van der Belt TH. Continuous monitoring of vital signs in the general ward using wearable devices: Randomized controlled trial. J Med Internet Res. 2020;22(6):e15471. https://doi.org/10.2196/15471.
21. Bowles T, Trentino KM, Lloyd A, Trentino L, Jones G, Murray K, Thompson A, Halpin S, Waterer G. Outcomes in patients receiving continuous monitoring of vital signs on general wards: A systematic review and meta-analysis of randomised controlled trials. Digit health. 2024;10:20552076241288826. https://doi.org/10.1177/20552076241288826.
22. Patel V, Orchanian-Cheff A, Wu R. Evaluating the validity and utility of wearable technology for continuously monitoring patients. JMIR mHealth uHealth. 2021;9(8):e17411. https://doi.org/10.2196/17411.

23. van der Stam JA, Mestrom EHJ, Nienhuijs SW, de Hingh IHJT, Boer A-K, van Riel NAW, de Groot KTJ, Verhaegh W, Scharnhorst V, Bouwman RA. A wearable patch based remote early warning score (REWS) in major abdominal cancer surgery patients. Eur J Surg Oncol. 2023;49(1):278–84. https://doi.org/10.1016/j.ejso.2022.08.034.
24. Areia C, Biggs C, Santos M, Thurley N, Gerry S, Tarassenko L, Watkinson P, Vollam S. The impact of wearable continuous vital sign monitoring on deterioration detection and clinical outcomes in hospitalised patients: a systematic review and meta-analysis. Crit Care (London England). 2021;25(1):351. https://doi.org/10.1186/s13054-021-03766-4.