'The Score Matters': wide variations in predictive performance of 18 paediatric track and trigger systems

Susan M Chapman,[1,2,3] Jo Wray,[2,4] Kate Oulton,[2,4] Christina Pagel,[5,6] Samiran Ray,[6,7] Mark J Peters[6,7]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/archdischild-2016-311088).

[1]Great Ormond Street Hospital, London, UK
[2]UCL Great Ormond Street Institute of Child Health, London, UK
[3]Department of Children's Nursing, London South Bank University, London, UK
[4]Outcomes and Experience Research in Children's Health, Illness and Disability (ORCHID), Great Ormond Street Hospital, London, UK
[5]Clinical Operational Research Unit, University College London, London, UK
[6]Paediatric Intensive Care Unit, Great Ormond Street Hospital, London, UK
[7]Respiratory, Anaesthesia, and Critical Care Group, UCL Great Ormond Street Institute of Child Health, London, UK

Correspondence to
Dr Susan M Chapman, Great Ormond Street Hospital, Great Ormond Street, London WC1N 3JH, UK;
Sue.Chapman@gosh.nhs.uk

Received 23 April 2016
Revised 21 December 2016
Accepted 16 January 2017
Published Online First 10 March 2017

ABSTRACT

Objective To compare the predictive performance of 18 paediatric early warning systems (PEWS) in predicting critical deterioration.

Design Retrospective case-controlled study. PEWS values were calculated from existing clinical data, and the area under the receiver operator characteristic curve (AUROC) compared.

Setting UK tertiary referral children's hospital.

Patients Patients without a 'do not attempt resuscitation' order admitted between 1 January 2011 and 31 December 2012. All patients on paediatric wards who suffered a critical deterioration event were designated 'cases' and matched with a control closest in age who was present on the same ward at the same time.

Main outcome measures Respiratory and/or cardiac arrest, unplanned transfer to paediatric intensive care and/or unexpected death.

Results 12 'scoring' and 6 'trigger' systems were suitable for comparative analysis. 297 case events in 224 patients were available for analysis. 244 control patients were identified for the 311 events. Three PEWS demonstrated better overall predictive performance with an AUROC of 0.87 or greater. Comparing each system with the highest performing PEWS with Bonferroni's correction for multiple comparisons resulted in statistically significant differences for 13 systems. Trigger systems performed worse than scoring systems, occupying the six lowest places in the AUROC rankings.

Conclusions There is considerable variation in the performance of published PEWS, and as such the choice of PEWS has the potential to be clinically important. Trigger-based systems performed poorly overall, but it remains unclear what factors determine optimum performance. More complex systems did not necessarily demonstrate improved performance.

What is already known on this topic?

- Paediatric early warning systems (PEWS) are widely used to detect deterioration in hospitalised children.
- The component parameters, weighting frameworks and scoring thresholds vary between differing PEWS.
- Of the numerous PEWS in the literature and clinical practice, only a minority have been previously evaluated for their predictive performance.

What this study adds?

- There is wide variation in the performance of PEWS.
- There are no clear defining features which characterise the best performing PEWS.
- The choice of PEWS may be an important factor in improving outcome for deteriorating hospitalised children.

INTRODUCTION

Timely detection of evolving critical illness makes it easier to treat. Paediatric early warning systems (PEWS) should alert staff to deteriorating children and accelerate access to appropriate intervention.[1] Despite weak evidence,[2 3] they are widely recommended.[4–8] In 2013, 85% of UK centres caring for children were using PEWS.[9]

Early warning systems are either 'score'-based or 'trigger'-based. Score-based systems assign values to vital signs (or other parameters), describing the variance from normal. These component values are then combined into an overall score. Higher scores should indicate reduced physiological reserve and prompt an escalating series of actions, culminating in senior clinician or rapid response team (RRT) review. The simpler 'trigger'-based systems contain thresholds for parameters without combining into an overall score. Again, actions such as RRT review are often mandated. Scoring systems provide a more continuous description of the degree of abnormality in the child's physiological state compared with binary 'all or nothing' trigger systems.

The logic of standardised risk assessment is compelling, but the majority of PEWS have been developed using expert opinion alone. Comparative data are lacking on the relative performance of the 31 different published PEWS. Only a minority of these (14) have undergone *any* assessment of predictive validity.[1 10] Only one study compared the performance of multiple (3) scores.[11] Comparisons across studies are confounded by variance in the setting, methodologies and outcomes described.[2]

Some might argue that the lack of validation or performance data is a secondary issue since the implementation of *any* system is the most important step. A system provides a structure for

► http://dx.doi.org/10.1136/archdischild-2016-312136

To cite: Chapman SM, Wray J, Oulton K, *et al*. *Arch Dis Child* 2017;**102**:487–495.

communication and builds consideration of risk of deterioration into daily practice. The alternative view is that the validity and calibration of any score are essential for utility. A score consistently providing false alerts while missing critical deteriorations elsewhere carries potential for harm by triaging resources incorrectly and increasing response times through 'alarm fatigue'.[12] Systems have to balance specificity and sensitivity, and so the precision of the thresholds included may be crucial.

We undertook a study comparing the performance of 18 PEWS in predicting critical deterioration in a UK tertiary referral children's hospital. Our null hypothesis was that the scores would show equivalent areas under the receiver operating characteristic curves.

METHODS

Evaluation of predictive validity

We undertook a retrospective case–control study of patients below 19 years of age, without a 'do no attempt resuscitation' order, who were admitted to our tertiary specialist children's hospital between 1 January 2011 and 31 December 2012. All patients who suffered a respiratory and/or cardiac arrest, unplanned transfer to paediatric intensive care unit (PICU) and/or unexpected death on the ward were designated 'cases'. They were identified from local data collected for the Paediatric Intensive Care Audit Network (PICANet) database,[13] the hospital resuscitation database and cross-referenced against intensive care admission records. Case patients present on the ward for <2 hours before the event were excluded, as this was considered the minimum time for the child to be assessed, clinical signs recorded and action to be taken.

Case patients were each matched with a single control, present on the same ward at the same time. Wards were considered a proxy match for diagnostic specialty. The child closest in age to the case patient was identified. To ensure at least one set of observations could be extracted, control patients present on the ward for <24 hours were excluded, with the exception of wards classified as providing short stay/day case care where the threshold was 4 hours. Patients previously entered into the study were eligible to act as a control, provided they did not suffer a critical deterioration event within the following 48 hours. If healthcare records were unavailable or the vital sign record was missing, the patient was excluded and a new control was sought using the same procedure.

Data extraction

Clinical data were extracted from the healthcare record of case patients for a period of 48 hours before the critical deterioration event. The final hour of data before the deterioration event in the case patient was excluded to establish if the PEWS could identify critical deterioration with at least 1 hour's notice. Data from controls were extracted for the same 47-hour period. Data were extracted by a single researcher (SC) using a standardised pro forma. Vital signs were extracted as continuous variables. Respiratory effort was assessed retrospectively as mild, moderate or severe using standardised criteria.[14] Dichotomous variables were assessed using the criteria in online supplementary table S1.

At the time of the study, standard protocols were in place for recording and documenting vital signs, which nurses were informed of at induction and yearly intervals thereafter. The protocol mandated recording of a full set of vital signs within 2 hours of the start of the 12-hour shift. Elevated PEWS scores required repeat vital sign recording after 30 min. Ongoing frequency of recording was at the discretion of the bedside nurse.

Identification of PEWS

PEWS were identified through our recent systematic review.[2] We excluded a priori PEWS where vital signs were assessed subjectively or against individual patient baseline values. Components of the remaining systems were reviewed to confirm that they could be extracted from the healthcare records. Criteria for data extraction were developed for included parameters (see online supplementary table S1) together with the weighting framework for scoring systems. Minor inconsistencies such as overlapping age bandings were modified in a consistent manner to facilitate score calculation (see online supplementary table S2). Our hospital's local unpublished PEWS (Children's Early Warning Score (CEWS)) were also included (see online supplementary table S3).

PEWS score calculation

Data were electronically checked for internal consistency and manually checked for accuracy. Inconsistencies were resolved by reviewing the data extraction proforma and healthcare records.

A recording of one or more vital signs was considered an observation data set. The PEWS value for each system was calculated for each observation data set. Missing observations were presumed to be normal (score 0), consistent with clinical practice and the methodology of previous studies.[11 15 16]

Data analysis

Analysis was performed using SPSS and R (http://www.cran.r-project.org). The maximum observed value for each PEWS for each patient in the 47 hours before the event was used in the comparison. Characteristics of cases and controls were compared with the Mann-Whitney U-test for continuous variables and χ^2 for categorical variables. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), likelihood ratio positive test and likelihood ratio negative test were calculated. The area under receiver operator characteristic curves (AUROC) was calculated for each PEWS and compared with the best performing system using Delong's test for correlated curves.[17] Significance testing was adjusted for the multiple comparisons of AUROC with Bonferroni's correction, meaning values of $p<0.0025$ were considered significant.

The score that maximised sensitivity and specificity for each scoring system was identified as the optimal score.[18] The number of case and control patients who would be correctly and incorrectly identified at this threshold was calculated.

RESULTS

Characteristics of the identified PEWS

Thirty-one PEWS were identified by the systematic review.[2] Seven contained parameters requiring subjective assessment, six required knowledge of the baseline vital signs and one inadequately described the component parameters: these were excluded. The remaining systems plus our local CEWS resulted in 18 PEWS. Systems with the same name were numbered in order of publication to distinguish between them (table 1).

Twelve PEWS were 'scoring' and six were 'trigger' systems. The number of component parameters varied from 3 to 19. Some systems combined two or more variables within a single parameter, for example, oxygen therapy and saturation values. Forty variables, either alone or in combination, were identified.

Vital signs were prominent. All 18 PEWS included heart rate and respiratory rate, 13 included oxygen saturation (72%) and 11 blood pressure (61%). Temperature was a component of only seven systems (39%). Five weighting frameworks were identified across the 12 scoring systems, with 3 PEWS also

Table 1 Key characteristics and parameters

System Name, first citation	Score or trigger	Maximum score	Age ranges	Parameters (n)	Parameters (scored using weighting framework)																Additional risk factors Score 1 for each unless otherwise indicated	Weighting framework
					Vital signs						Concern		Other parameters									
					Heart rate	Respiratory rate	Oxygen saturation	Systolic BP	Capillary refill time	Temperature	Staff concern	Parent concern	Respiratory	Behaviour	Cardiovascular	Consciousness	Seizure	Respiratory distress	Airway threat	Oxygen therapy		
Bedside PEWS[19]	S	26	5	7	✓	✓	✓	✓	✓									✓		✓		0, 1, 2, 4
Bristol PEW tool[20]	T	13	1	14	✓*	✓	✓†		✓		✓					✓	✓			✓	Apnoea±bradycardia; DKA; clinically tiring or complete airway obstruction; hyperkalaemia; nebulised adrenaline; signs of shock (eg, poor perfusion, ±low BP); suspected meningococcus	Trigger
Cardiff and Vale PEWS[18]	S	8	5	8	✓	✓	✓	✓			✓					✓		✓	✓	✓		0, 1
Children's Early Warning score‡	S	21	4	6	✓	✓	✓	✓		✓						✓						0, 1,2, 3, 4
Children's Early Warning Tool[21]	S	24	4	9	✓	✓	✓	✓	✓	✓						✓		✓		✓		0, 1, 2, 3
ITAT[22]	S	8	5	4	✓	✓	✓			✓												0, 1, 2
MET activation criteria I[23]	T	9	5	9	✓	✓	✓	✓			✓					✓	✓	✓	✓		Cardiac/respiratory arrest; apnoea or cyanosis	Trigger
MET activation criteria II[24]	T	9	5	9	✓	✓	✓	✓			✓	✓				✓	✓	✓	✓		Cardiac/respiratory arrest; apnoea or cyanosis	Trigger
Modified Bristol PEWS[25]	T	15	5	16	✓*	✓*	✓†		✓		✓					✓	✓	✓			Apnoea±bradycardia; clinically tiring or complete airway obstruction; hyperkalaemia; marked increased work of breathing; nebulised adrenaline (or no improvement); pH <7.2; poor perfusion, ±low BP; large central/peripheral temperature gradient; unresolved pain on current analgesia therapy	Trigger
Modified PEWS I[26]	S	9	1	3	✓	✓			✓				✓	✓	✓					✓		0, 1, 2, 3
Modified PEWS II[27]	S	26	5	18	✓	✓	✓	✓	✓	✓						✓		✓		✓	CVL; intravenous bolus fluid or blood product within past 4 hours	Abnormal airway or positive pressure ventilation; active acquired/congenital heart disease or history of heart surgery; home oxygen; before/after any transplant; gastrostomy or jejunostomy tube; previous ICU admission; severe developmental, neurological or neuromuscular disease	0, 1, 2

Continued

Table 1 Continued

System Name, first citation	Score or trigger	Maximum score	Age ranges	Parameters (n)	Parameters (scored using weighting framework)																Additional risk factors Score 1 for each unless otherwise indicated	Weighting framework
					Vital signs						Concern		Other parameters									
					Heart rate	Respiratory rate	Oxygen saturation	Systolic BP	Capillary refill time	Temperature	Staff concern	Parent concern	Respiratory	Behaviour	Cardiovascular	Consciousness	Seizure	Respiratory distress	Airway threat	Oxygen therapy		
Modified PEWS III[28]	S	28	5	8	✓	✓	✓	✓	✓	✓								✓		✓		
NHSI PEWS[29]	S	7	4	7	✓	✓					✓	✓				✓		✓		✓	Stridor or apnoea	0, 1
PEW score I[30]	S	10	1	4	✓	✓			✓			✓	✓	✓	✓					✓		0, 1, 2, 3
PEW score II[16]	S	13	1	4	✓	✓			✓				✓	✓	✓					✓	15 min nebulisers or vomiting post-op	0, 1, 2, 3
PEW system score[15]	S	32	5	19	✓	✓	✓	✓	✓	✓						✓				✓	Bolus fluid; pulses	More than three medical specialties involved in care; abnormal airway (not tracheostomy); CVL; gastrostomy; home oxygen; medication score; previous admission to ICU; severe cerebral palsy; transplant recipient	0, 1, 2, 3
PMET triggers[31]	T	7	5	8	✓	✓	✓†	✓			✓	✓				✓	✓			✓			Trigger
THSC MET calling criteria[32]	T	7	1	7	✓	✓	✓†	✓	✓		✓					✓	✓	✓	✓	✓	Poor peripheral pulses; mottled extremities		Trigger

Indicators combined within a single parameter are presented in coloured text /✓. All studies are single-centred unless otherwise stated.
*Following one bolus of 10 mL/kg fluid.
†Separate parameters for children with and without cyanotic heart disease.
‡The Children's Early Warning Score is the local unpublished PEWS.
BP, blood pressure; CVL, central venous line; DKA, diabetic ketoacidosis; ICU, intensive care unit; ITAT, inpatient triage, assessment and treatment score; MET, Medical Emergency Team; NHSI, NHS Institute; PEWS, paediatric early warning system; PMET, Paediatric Medical Emergency Team; SVT, supraventricular tachycardia; THSC, Toronto Hospital for Sick Children.

Table 2 Patient characteristics (each patient episode)

	Cases (n=297) n (%)	Controls (n=311) n (%)	p Value
Gender			
Male	130 (43.8%)	167 (53.7)	0.018*
Female	167 (56.3)	144 (46.3)	
Age			
0 to <6 months	70 (23.6)	66 (21.2)	0.910†
6 months to <1 year	41 (13.8)	47 (15.1)	
1 year to <4 years	87 (29.3)	94 (30.2)	
4 years to <10 years	49 (16.5)	55 (17.7)	
10 years to <19 years	50 (16.8)	49 (15.8)	
Gestation below 37 weeks	60 (20.1)	48 (15.4)	0.152*
Weight, median, (IQR)	10.4 kg (1.71–87.00)	11.1 kg (2.10–94.20)	0.668†
Previous same hospital admission			
0	150 (50.5)	145 (46.6)	0.946†
1–5	66 (22.2)	92 (29.6)	
6–10	29 (9.8)	27 (8.7)	
11–20	20 (6.7)	26 (8.4)	
21–50	25 (8.4)	16 (5.2)	
>50	7 (2.4)	5 (1.6)	
Previous PICU admission (before this admission)			
0	247 (83.1)	276 (88.7)	0.061†
1	32 (10.8)	20 (6.4)	
2	15 (5.1)	4 (1.3)	
3–5	1 (0.3)	5 (1.6)	
>5	2 (0.7)	6 (1.9)	
Previous PICU admission (this episode)			
0	185 (62.3)	238 (76.5)	<0.01†
1	75 (25.2)	55 (17.7)	
2	17 (5.7)	14 (4.5)	
3–5	14 (4.7)	4 (1.3)	
>5	6 (2.0)	0 (0.0)	
Admitting specialty			
Medical	186 (62.6)	205 (65.9)	0.19*
Surgical	57 (19.2)	66 (21.2)	
Intensive care	54 (18.2)	40 (12.9)	
Type of admission			
Elective	105 (35.4)	189 (60.8)	<0.001*
Emergency	192 (64.6)	122 (39.2)	
Specialty at event			
Medical	228 (76.8)	237 (76.2)	1.0*
Surgical	69 (23.2)	74 (23.8)	
Critical deterioration event			
PICU transfer	186 (62.6)	0	N/A
Respiratory arrest	84 (28.3)	0	
Cardiac arrests	27 (9.1)	0	
Death on ward	0 (0)	0	
Reason for event			
Respiratory	176 (59.3)	0	N/A
Cardiovascular	67 (22.6)	0	
Neurological	38 (12.8)	0	
Other	16 (5.4)	0	
Length of stay in days, median (IQR)	57.1 (21.0–122.0)	35.9 (12.8–89.4)	0.001†
Length of hospital stay			
<1 day	2 (0.7)	6 (1.9)	0.021*
<7 days	22 (7.4)	39 (12.5)	
<30 days	74 (24.9)	91 (29.3)	
≥30 days	199 (670)	175 (56.3)	

Continued

incorporating additional points for risk factors. Differences between systems were often minor. The maximum scores varied from 7 to 32 (table 1).

Patient characteristics

We identified 319 critical deterioration events. In eight episodes, the patient was present on the ward for <2 hours, leaving 311 eligible critical deterioration events in 237 patients. A total of 14 case patient records were missing, leaving a case sample of 297 events in 224 patients. A total of 244 control patients were identified for the 311 events.

Overall, 13 551 observations sets were performed, 8360 on cases and 5191 on controls. The median number of observation sets per patient per day was 13 for cases and 6 for controls. Only 36.4% of observation sets contained the five vital sign parameters and assessment of consciousness required for complete recording of the local PEWS.

Case patients were more likely to be female (56.3% vs 46.3%, p=0.009), have been admitted as an emergency (64.6% vs 39.2%, p≤0.01) and have a longer hospital stay (median 57.1 vs 35.9 days, p≤0.01). Mortality was also higher for case patients at 24 hours, 30 days and hospital discharge (p≤0.001). A summary of patient characteristics is shown in table 2.

A total of 186 (62.6%) critical deterioration events were categorised as unplanned transfers to the PICU, 84 (28.3%) respiratory arrests and 27 (9.1%) cardiac arrests. Thirty-one patients remained on the ward after a cardiac or respiratory arrest. Six patients died before transfer to intensive care.

Predictive performance

Three PEWS demonstrated better performance overall (table 3). Comparing each system with the highest performing PEWS resulted in statistically significant differences for 13 systems. Overall trigger systems performed worse than scoring systems, occupying six of the lowest seven places in the AUROC rankings.

Sensitivity, specificity, PPV, NPV and positive and negative likelihood ratios for the optimal score are given in table 4. Values for trigger systems represent the breech of one or more trigger thresholds.

Trigger systems demonstrated better sensitivity (range 0.90–0.96) than scoring systems (range 0.46–0.83), but worse specificity (range 0.28–0.56 vs 0.65–0.91, respectively).

Our local PEWS performed modestly, ranked 10th overall. Comparison with the highest performing PEWS demonstrates the significantly worse predictive ability (figure 1). At the optimal score, the Cardiff and Vale PEWS would correctly identify 59 more deteriorating patients than our local PEWS, with only 4 additional false alerts.

PEWS demonstrated the ability to detect children at risk of critical deterioration a significant time before the event. Median time from optimal score[18] to event ranged from 17 hours (IQR 6.8–35.7) to 39.5 hours (IQR 17.4–46.6) for patients correctly identified by scoring systems. Longer times were demonstrated by trigger systems: 27.9 (IQR 13.7–42.4) to 39.8 hours (IQR 23.8–46.2), reflecting the increased sensitivity (table 4).

DISCUSSION

The choice of PEWS is potentially important. Effective identification of 'at risk' children is crucial, but a poorly validated system may also erode staff confidence, waste valuable resources and overburden staff with false alerts. This study found that performance varied widely. Eight PEWS were good predictors, nine were useful and two poor.[33] Score-based systems consistently outperformed trigger systems. A larger number of parameters

Table 2 Continued

	Cases (n=297) n (%)	Controls (n=311) n (%)	p Value
Outcome			
Alive at 24 hours	279 (93.9)	311 (100%)	<0.001*
Alive at 30 days	246 (82.8)	308 (99.0)	<0.001*
Alive at discharge	220 (74.1)	301 (96.8)	<0.001*

*χ^2.
†Mann–Whitney U test.
PICU, paediatric intensive care unit.

Table 3 Comparative performance

	AUROC (95% CI)	z-Score	p Value
Scoring systems			
Cardiff and Vale PEWS	0.89 (0.86 to 0.91)	N/A	N/A
Bedside PEWS	0.88 (0.85 to 0.91)	0.72	0.47
Modified PEWS III	0.87 (0.85 to 0.90)	1.58	0.11
Children's Early Warning Tool	0.85 (0.82 to 0.88)	3.21	0.001
Modified PEWS II	0.85 (0.82 to 0.88)	2.87	0.004
PEWS I	0.83 (0.80 to 0.86)	4.06	<0.001
NHSI PEWS	0.82 (0.79 to 0.86)	4.52	<0.001
PEWS system score	0.82 (0.78 to 0.85)	4.42	<0.001
PEWS II	0.79 (0.75 to 0.82)	6.00	<0.001
CEWS	0.79 (0.75 to 0.82)	7.12	<0.001
ITAT score	0.77 (0.74 to 0.81)	7.12	<0.001
Modified PEWS I	0.74 (0.70 to 0.78)	8.06	<0.001
Trigger systems			
THSC MET calling criteria	0.73 (0.69 to 0.77)	9.31	<0.001
MET activation criteria I	0.71 (0.70 to 0.75)	10.70	<0.001
MET activation criteria II	0.71 (0.70 to 0.75)	10.70	<0.001
PMET triggers I	0.71 (0.67–0.75)	10.82	<0.001
Modified Bristol PEWS	0.62 (0.58 to 0.67)	16.01	<0.001
Bristol PEWS	0.62 (0.58 to 0.67)	16.01	<0.001

Performance was assessed by calculation of the AUROC. Systems were then ranked, and performance was compared with the highest ranked PEWS (Cardiff and Vale PEWS) using Delong's test for correlated curves. z-scores represent comparison of mean values. Significance testing was adjusted for the multiple comparisons of AUROC with Bonferroni's correction, meaning values of p<0.0025 were considered significant.
AUROC, area under the receiver operator characteristic curve; CEWS, Children's Early Warning Score; ITAT, inpatient triage, assessment and treatment score; MET, Medical Emergency Team; NHSI, NHS Institute; PEWS, paediatric early warning system; PMET, Paediatric Medical Emergency Team; THSC, Toronto Hospital for Sick Children.

did not appear to improve performance, for instance, the two lowest ranked systems had 16 and 14 parameters, respectively, compared with 8 parameters of the highest ranked system.

The Cardiff and Vale PEWS, Bedside PEWS and Modified PEWS III performed better than the majority of scores, but with no significant differences between them. There were no obvious reasons why these systems outperformed the others. All three systems included heart rate and respiratory rate, oxygen saturation and blood pressure.

At the optimal score, scoring systems demonstrated poorer sensitivity, but superior specificity than trigger systems, which may reduce false alerts and build clinician confidence. Lowering the scoring thresholds improves sensitivity, creating additional opportunities to intervene and potentially improve outcome.[34] The ability to select the threshold that balances sensitivity and specificity most appropriate to the local environment gives scoring systems some advantages. However, they are more complex to use, carrying the risk of inaccurate calculation[35 36] and inappropriate response.[37 38]

The current local PEWS performed only modestly, despite being developed by local clinicians, using local data and expertise. It was considerably outperformed by systems externally validated in similar and differing populations. We have no reason to believe our situation is unique. It is likely that many other locally developed unvalidated PEWS would demonstrate similar performance if evaluated rigorously. We are considering changing to the Bedside PEWS as it has now been evaluated in similar populations, is subject to an international multicentre trial[10] and demonstrated equivalent performance with the top-ranked PEWS. This may facilitate further collaborative research in the future.

All PEWS demonstrated the ability to identify deteriorating children a number of hours before the event. Median hours from optimal score to critical deterioration event varied from 17.0 to 39.5 hours for scoring systems and 27.9 to 39.8 hours for trigger systems. This is longer than previous study findings for comparable scoring thresholds.[16] Both scoring and trigger systems can act as important 'early warning' to front-line staff of ward-based children at risk of critical deterioration, but require appropriate escalation and intervention by healthcare staff. Studies have identified that this may not always be achieved in practice.[6 39 40]

Limitations

Values for PEWS were calculated retrospectively from data extracted by a single researcher who was not blinded to the patient's outcome. Although standardised criteria were applied, there was no other verification of data and accuracy of documented vital sign values, and other observations could not be tested. Administration of a fluid bolus could not be reliably extracted affecting three of the PEWS.[15 20 27]

Data sets were frequently incomplete. Missing values were assumed to be 'normal' (score 0), but a recent study identified a greater proportion of incomplete data sets were associated with 'critical' (elevated) score compared with complete data sets.[36] Incomplete vital sign recording remains a problem in clinical practice[21 41] and may underestimate PEWS performance.

The study was conducted in a tertiary specialist children's hospital without an emergency department. Results may not be generalisable to children in other settings. Different results may also be seen for different outcomes and combinations of outcomes. Greater standardisation of reporting and consensus on pragmatic measures to evaluate PEWS and other similar interventions would facilitate meaningful comparison and collaborative research.[42]

CONCLUSION

The choice of PEWS may be important. Trigger-based systems performed poorly overall, but it remains unclear what factors determine optimum performance. More complex systems did not necessarily demonstrate improved performance. Variation in performance has important implications for effective identification of children 'at risk', staff confidence in the system and effective use of resources.

It is likely that many other hospitals have developed their own systems without rigorous evaluation of their validity.[43] The high and increasing number of both published and unpublished PEWS raises concerns that paediatrics may be following a path similar to that of adult track and trigger systems, with multiple poorly validated systems with unknown predictive power. This

Table 4 Performance at optimal score

PEWS (AUROC rank)	Optimal score/ maximum score	Case patients (n=297)		Control patients (n=311)		Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	LR +ve (95% CI)	LR −ve (95% CI)	Median hours to event (inter-quartile range)
		Score at or above threshold (TP)	Score below threshold (FN)	Score at or above threshold (FP)	Score below threshold (TN)							
Scoring systems												
Cardiff and Vale PEWS (1)	3/8	238	59	44	267	0.80 (0.75 to 0.84)	0.86 (0.81 to 0.89)	0.84 (0.80 to 0.88)	0.82 (0.77 to 0.86)	5.66 (4.28 to 7.49)	0.23 (0.18 to 0.29)	26.60 (11.57 to 42.19)
Bedside PEWS (2)	6/26	215	82	35	276	0.72 (0.67 to 0.77)	0.89 (0.85 to 0.92)	0.860 (0.81 to 0.90)	0.77 (0.72 to 0.81)	6.43 (4.67 to 8.86)	0.311 (0.26 to 0.37)	26.25 (13.94 to 43.29)
Modified PEWS III (3)	7/28	204	93	28	283	0.69 (0.63 to 0.74)	0.91 (0.87 to 0.94)	0.879 (0.83 to 0.92)	0.75 (0.71 to 0.80)	7.63 (5.31 to 10.95)	0.34 (0.29 to 0.41)	21.61 (12.42 to 40.10)
Modified PEWS II (4)	6/26	228	69	63	248	0.83 (0.78 to 0.87)	0.71 (0.66 to 0.76)	0.731 (0.68 to 0.78)	0.81 (0.76 to 0.85)	2.85 (2.38 to 3.42)	0.25 (0.19 to 0.32)	36.57 (16.57 to 46.00)
Children's Early Warning Tool (4)	4/24	245	52	90	221	0.77 (0.72 to 0.81)	0.80 (0.75 to 0.84)	0.784 (0.73 to 0.83)	0.78 (0.73 to 0.83)	3.79 (3.01 to 4.77)	0.29 (0.24 to 0.36)	37.66 (22.39 to 44.74)
PEWS I (6)	3/10	247	50	99	212	0.83 (0.78 to 0.87)	0.68 (0.63 to 0.73)	0.714 (0.66 to 0.76)	0.81 (0.76 to 0.85)	2.61 (2.20 to 3.10)	0.25 (0.19 to 0.32)	24.00 (11.23 –44.13)
NHSI PEWS (7)	2/7	247	50	108	203	0.83 (0.78 to 0.87)	0.65 (0.60 to 0.71)	0.696 (0.65 to 0.74)	0.80 (0.75 to 085)	2.40 (2.04 to 2.81)	0.26 (0.20 to 0.33)	29.90 (14.57 to 43.63)
PEWS system score (7)	9/32	207	90	78	233	0.70 (0.64 to 0.75)	0.75 (0.70 to 0.80)	0.726 (0.67 to 0.78)	0.72 (0.70 to 0.77)	2.78 (2.26 to 3.42)	0.40 (0.34 to 0.48)	39.50 (17.43 to 46.57)
PEWS II (9)	4/13	181	116	50	261	0.61 (0.55 to 0.67)	0.84 (0.79 to 0.88)	0.784 (0.72 to 0.83)	0.69 (0.64 to 0.74)	3.79 (2.89 to 4.96)	0.47 (0.40 to 0.54)	26.00 (11.75 to 41.58)
CEWS (9)	4/21	179	118	48	263	0.60 (0.54 to 0.66)	0.85 (0.80 to 0.88)	0.789 (0.73 to 0.84)	0.69 (0.64 to 0.74)	3.91 (2.96 to 5.15)	0.47 (0.41 to 0.54)	21.05 (10.38 to 40.12)
ITAT score (11)	3/8	202	95	82	229	0.68 (0.62 to 0.73)	0.74 (0.68 to 0.78)	0.711 (0.65 to 0.76)	0.71 (0.65 to 0.76)	2.58 (2.11 to 3.16)	0.43 (0.37 to 0.51)	28.95 (14.70 to 43.96)
Modified PEWS I (12)	4/9	135	162	31	280	0.46 (0.40 to 0.51)	0.90 (0.86 to 0.93)	0.813 (0.74 to 0.87)	0.63 (0.59 to 0.68)	4.56 (3.19 to 6.51)	0.61 (0.55 to 0.67)	17.00 (6.75 to 35.68)
Trigger systems												
THSC MET calling criteria (13)	1 or more triggers	267	30	138	173	0.90 (0.86 to 0.93)	0.56 (0.50 to 0.61)	0.66 (0.61 to 0.71)	0.85 (0.79 to 0.90)	2.03 (1.78 to 2.31)	0.18 (0.13 to 0.26)	27.90 (13.74 to 42.37)
MET activation criteria I (14)	1 or more triggers	276	21	158	153	0.93 (0.89 to 0.96)	0.49 (0.44 to 0.55)	0.64 (0.59 to 0.68)	0.88 (0.82 to 0.92)	1.83 (1.63 to 2.05)	0.14 (0.10 to 0.22)	33.87 (18.76 to 45.52)
MET activation criteria II (14)	1 or more triggers	276	21	158	153	0.923 (0.89 to 0.96)	0.49 (0.44 to 0.55)	0.64 (0.59 to 0.68)	0.88 (0.82 to 0.92)	1.83 (1.63 to 2.05)	0.14 (0.10 to 0.22)	33.92 (18.76 to 45.52)
PMET triggers I (14)	1 or more triggers	273	24	157	154	0.92 (0.88 to 0.95)	0.50 (0.44 to 0.55)	0.64 (0.59 to 0.68)	0.87 (0.80 to 0.68)	1.82 (1.62 to 2.04)	0.16 (0.11 to 0.24)	33.25 (16.90 to 45.42)
Modified Bristol PEWS (17)	1 or more triggers	286	11	223	88	0.96 (0.93 to 0.98)	0.28 (0.23 to 0.34)	0.56 (0.52 to 0.61)	0.90 (0.81 to 0.94)	1.34 (1.25 to 1.45)	0.13 (0.07 to 0.24)	39.83 (23.82 to 46.25)
Bristol PEWS (17)	1 or more triggers	285	12	223	88	0.96 (0.93 to 0.98)	0.28 (0.23 to 0.34)	0.56 (0.52 to 0.61)	0.88 (0.80 to 0.93)	1.34 (1.24 to 1.44)	0.14 (0.08 to 0.25)	39.73 (23.45 to 46.25)

The optimal score for trigger systems was determined as 1. The optimal score for scoring systems was determined as the cut-point which demonstrated the maximum value for the sum of the sensitivity and specificity, as described by Edwards *et al*.[18] As such, this differed between different scoring systems.
The hours to event was calculated as the number of hours between the case patient's first achieving the optimal score/trigger to the occurrence of the critical deterioration event.
AUROC, area under the receiver operator characteristic curve; CEWS, Children's Early Warning Score; FN, false negative; FP, false positive; ITAT, inpatient triage, assessment and treatment score; LR +ve, positive likelihood ratio; LR –ve, negative likelihood ratio; MET, Medical Emergency Team; NHSI, NHS Institute; NPV, negative predictive value; PEWS, paediatric early warning system; PMET, Paediatric Medical Emergency Team; PPV, positive predictive value; THSC, Toronto Hospital for Sick Children; TN, true negative; TP, true positive.

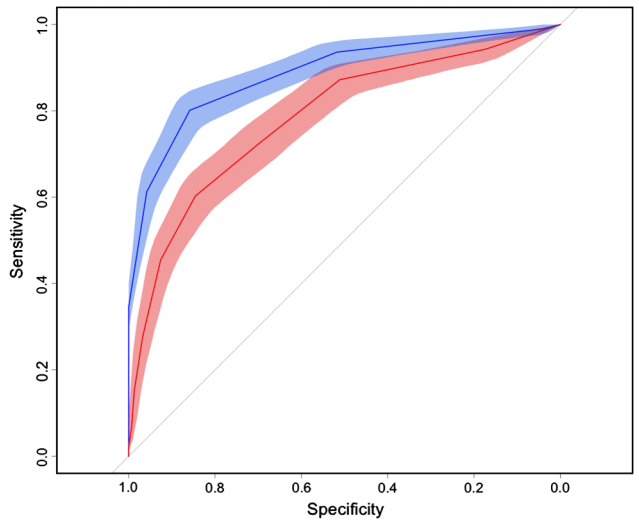

Figure 1 Comparison of AUROC of the highest performing PEWS and the local PEWS. The receiver operator characteristic curve of the local system (CEWS, AUROC 0.79) is shown in pink. The highest performing system (Cardiff and Vale PEWS, AUROC 0.89) is shown in blue. Shaded areas represent the 95% CIs for each system. AUROC, area under the receiver operator characteristic curve; CEWS, Children's Early Warning Score; PEWS, paediatric early warning system.

may explain why studies of PEWS and rapid response systems have so far failed to deliver the expected benefits.

Twitter Follow Samiran Ray @DrSamRay

Contributors SMC and MJP conceived the idea for the study. SMC, JW, KO and MJP contributed to the study design. SMC undertook the data collection. SMC, SR and MJP undertook the data analysis. SMC wrote the initial draft of the manuscript. All authors reviewed and revised the manuscript and approved the final draft.

Funding This study received no direct funding but was supported by the National Institute for Health Research, Biomedical Research Centre at Great Ormond Street Hospital for Children, NHS Foundation Trust and University College London.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Chapman SM, Grocott MPW, Franck LS. Systematic review of paediatric alert criteria for identifying hospitalised children at risk of critical deterioration. *Intensive Care Med* 2010;36:600–11.
2 Chapman SM, Wray J, Oulton K, *et al*. Systematic review of paediatric track and trigger systems for hospitalised children. *Resuscitation* 2016;109:87–109.
3 Maconochie IK, de Caen AR, Aickin R, *et al*. Part 6: pediatric basic life support and pediatric advanced life support: 2015 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science with Treatment Recommendations. *Resuscitation* 2015;95:e147–68.
4 National Confidential Enquiry into Patient Outcome and Death. Are we there yet? A review of organisational and clinical aspects of children's surgery. 2011. httpwww.ncepod.org.ukreportdownloadsSICfullreport.pdf
5 Royal College of Paediatrics and Child Health, NHS Improvement. A safe system for recognising and responding to children at risk of deterioration. 2016. http://www.rcpch.ac.uk/safer-system-children-risk-deterioration (accessed 19 Jun 2016).
6 Pearson GA. *Why children die: a pilot study*. London: CEMACH, 2008.
7 Wolfe I, Macfarlane A, Donkin A, *et al*. *Why young children die: death in infants, children and young people in the UK Part A*. Royal College of Paediatrics and Child Health, 2016.
8 Royal College of Nursing. *Standards for assessing, measuring and monitoring vital signs in infants, children and young people*. 2nd edn. Royal College of Nursing, 2011.
9 Duncan HP. Survey of early identification systems to identify inpatient children at risk of physiological deterioration. *Arch Dis Child* 2007;92:828.
10 Parshuram CS, Dryden-Palmer K, Farrell C, *et al*. Evaluating processes of care and outcomes of children in hospital (EPOCH): study protocol for a randomized controlled trial. *Trials* 2015;16:245.
11 Robson M-AJ, Cooper CL, Medicus LA, *et al*. Comparison of three acute care pediatric early warning scoring tools. *J Pediatr Nurs* 2013;28:e33–41.
12 Bonafide CP, Lin R, Zander M, *et al*. Association between exposure to nonactionable physiologic monitor alarms and response time in a children's hospital. *J Hosp Med* 2015;10:345–51.
13 Paediatric Intensive Care Audit Network National Report 2009–2011 (published September 2012): Universities of Leeds and Leicester.
14 Lakhanpaul M, MacFaul R, Werneke U, *et al*. An evidence-based guideline for children presenting with acute breathing difficulty. *Emerg Med J* 2009;26:850–3.
15 Duncan H, Hutchison J, Parshuram CS. The Pediatric Early Warning System score: a severity of illness score to predict urgent medical need in hospitalized children. *J Crit Care* 2006;21:271–8.
16 Akre M, Finkelstein M, Erickson M, *et al*. Sensitivity of the pediatric early warning score to identify patient deterioration. *Pediatrics* 2010;125:e763–9.
17 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
18 Edwards ED, Powell CVE, Mason BW, *et al*. Prospective cohort study to test the predictability of the Cardiff and Vale paediatric early warning system. *Arch Dis Child* 2009;94:602–6.
19 Parshuram CS, Hutchison J, Middaugh K. Development and initial validation of the bedside paediatric early warning system score. *Crit Care* 2009;13:R135.
20 Haines C, Perrott M, Weir P. Promoting care for acutely ill children-development and evaluation of a paediatric early warning tool. *Intensive Crit Care Nurs* 2006;22:73–81.
21 McKay H, Mitchell IA, Sinn K, *et al*. Effect of a multifaceted intervention on documentation of vital signs and staff communication regarding deteriorating paediatric patients. *J Paediatr Child Health* 2013;49:48–56.
22 Olson D, Davis NL, Milazi R, *et al*. Development of a severity of illness scoring system (inpatient triage, assessment and treatment) for resource-constrained hospitals in developing countries. *Trop Med Int Health* 2013;18:871–8.
23 Tibballs J, Kinney S, Duke T, *et al*. Reduction of paediatric in-patient cardiac arrest and death with a medical emergency team: preliminary results. *Arch Dis Child* 2005;90:1148–52.
24 Tibballs J, Kinney S. Reduction of hospital mortality and of preventable cardiac arrest and death on introduction of a pediatric medical emergency team. *Pediatr Crit Care Med* 2009;10:306–12.
25 Sefton G, McGrath C, Tume L, *et al*. What impact did a Paediatric Early Warning system have on emergency admissions to the paediatric intensive care unit? An observational cohort study. *Intensive Crit Care Nurs* 2015;31:91–9.
26 Skaletzky SM, Raszynski A, Totapally BR. Validation of a modified pediatric early warning system score: a retrospective case-control study. *Clin Pediatr (Phila)* 2012;51:431–5.
27 Bonafide CP, Roberts KE, Weirich CM, *et al*. Beyond statistical prediction: qualitative evaluation of the mechanisms by which pediatric early warning scores impact patient safety. *J Hosp Med* 2013;8:248–53.
28 Fuijkschot J, Vernhout B, Lemson J, *et al*. Validation of a Paediatric Early Warning Score: first results and implications of usage. *Eur J Pediatr* 2015;174:15–21.
29 Ennis L. Paediatric early warning scores on a children's ward: a quality improvement initiative. *Nurs Child Young People* 2014;26:25–31.
30 Demmel KM, Williams L, Flesch L. Implementation of the pediatric early warning scoring system on a pediatric hematology/oncology unit. *J Pediatr Oncol Nurs* 2010;27:229–40.
31 Kotsakis A, Lobos AT, Parshuram CS, *et al*. Implementation of a Multicenter Rapid Response System in Pediatric Academic Hospitals Is Effective. *Pediatrics* 2011;128:72–8.
32 Kukreti V, Gaiteiro R, Mohseni-Bod H. Implementation of a pediatric rapid response team: experience of the hospital for sick children in Toronto. *Indian Pediatr* 2014;51:11–15.
33 Smith GB, Prytherch DR, Schmidt PE, *et al*. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation* 2008;77:170–9.
34 Raymond TT, Bonafide CP, Praestgaard A, *et al*. Pediatric medical emergency team events and outcomes: a report of 3647 events from the American heart association's get with the guidelines-resuscitation registry. *Hosp Pediatr* 2016;6:57–64.
35 Smith AF, Oakey RJ. Incidence and significance of errors in a patient 'track and trigger' system during an epidemic of Legionnaires' disease: retrospective casenote analysis. *Anaesthesia* 2006;61:222–8.
36 Clifton DA, Clifton L, Sandu DM, *et al*. 'Errors' and omissions in paper-based early warning scores: the association with changes in vital signs-a database analysis. *BMJ Open* 2015;5:e007376.
37 Petersen JA, Mackel R, Antonsen K, *et al*. Serious adverse events in a hospital using early warning score—what went wrong? *Resuscitation* 2014;85:1699–703.

38 Odell M. Detection and management of the deteriorating ward patient: an evaluation of nursing practice. *J Clin Nurs* 2015;24:173–82.
39 Ninis N, Phillips C, Bailey L, *et al*. The role of healthcare delivery in the outcome of meningococcal disease in children: case-control study of fatal and non-fatal cases. *BMJ* 2005;330:1475.
40 Launay E, Gras-Le Guen C, Martinot A, *et al*. Suboptimal care in the initial management of children who died from severe bacterial infection: a population-based confidential inquiry. *Pediatr Crit Care Med* 2010;11:469–74.
41 Oliver A, Powell C, Edwards D, *et al*. Observations and monitoring: routine practices on the ward. *Paediatr Nurs* 2010;22:28–32.
42 Bonafide CP, Roberts KE, Priestley MA, *et al*. Development of a pragmatic measure for evaluating and optimizing rapid response systems. *Pediatrics* 2012;129: e874–81.
43 Gao H, McDonnell A, Harrison DA, *et al*. Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Med* 2007;33:667–79.