Using workload indicators to evaluate patient deterioration early warning tools

Presented by Anton van der Vegt

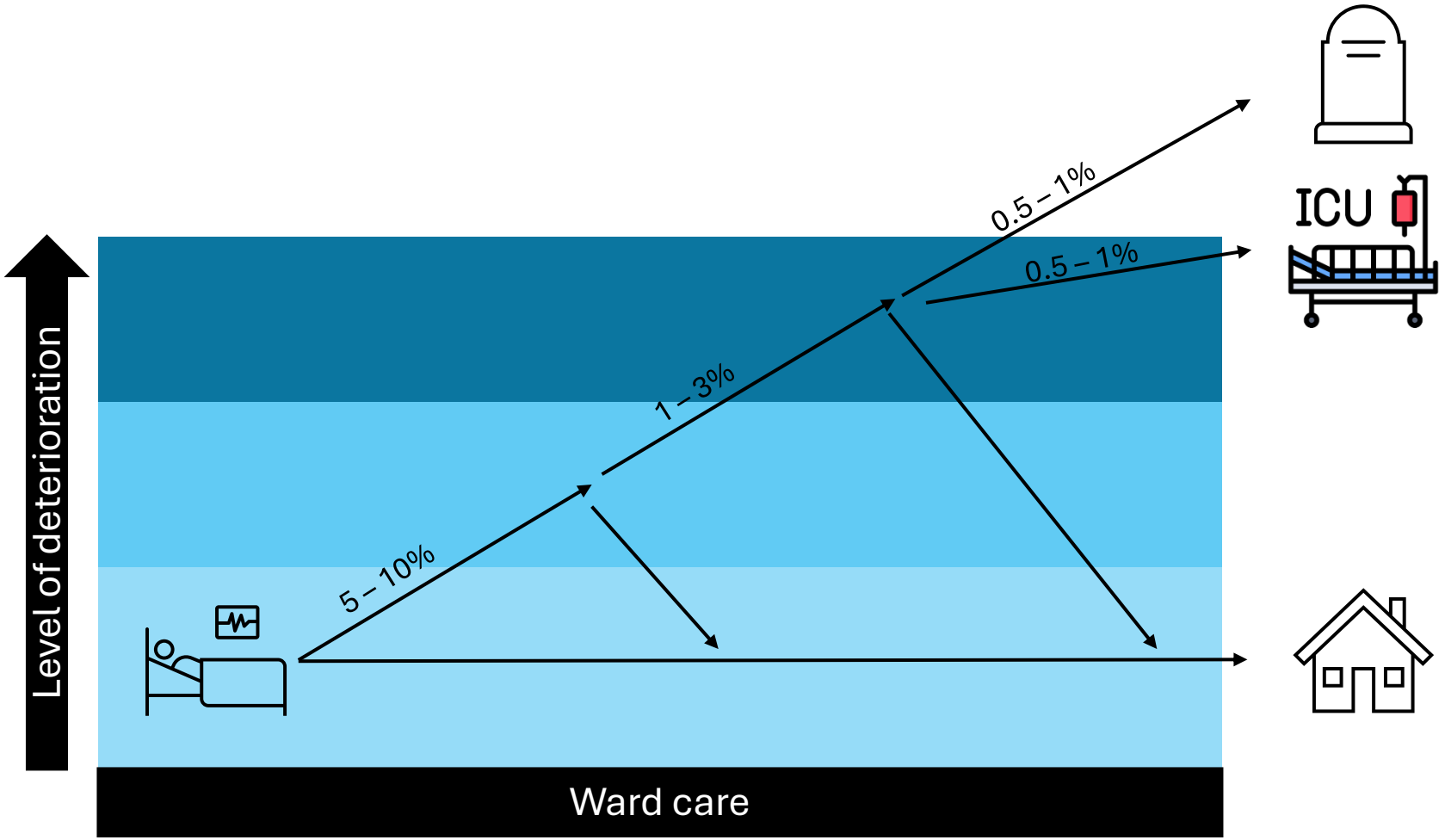
Thanks to our team:
Victoria Campbell, Imogen Mitchell, Oliver Redfern, Christian Subbe, Arthas Flabouris, Robyn Blythe, Rudolf Schnetler, Christopher Perkins, Naitik Mehta, Ian Scott

Conflict of interest

- None reported

What is patient deterioration?

'A deteriorating patient is one who moves from one clinical state to a worse clinical state which increases their individual risk of morbidity, including organ dysfunction, protracted hospital stay, disability, or death' - Jones et al, 2013



1. Jones D, Mitchell I, Hillman K, Story D. Defining clinical deterioration. Resuscitation. 2013;84(8):1029-1034. doi:10.1016/j.resuscitation.2013.01.013

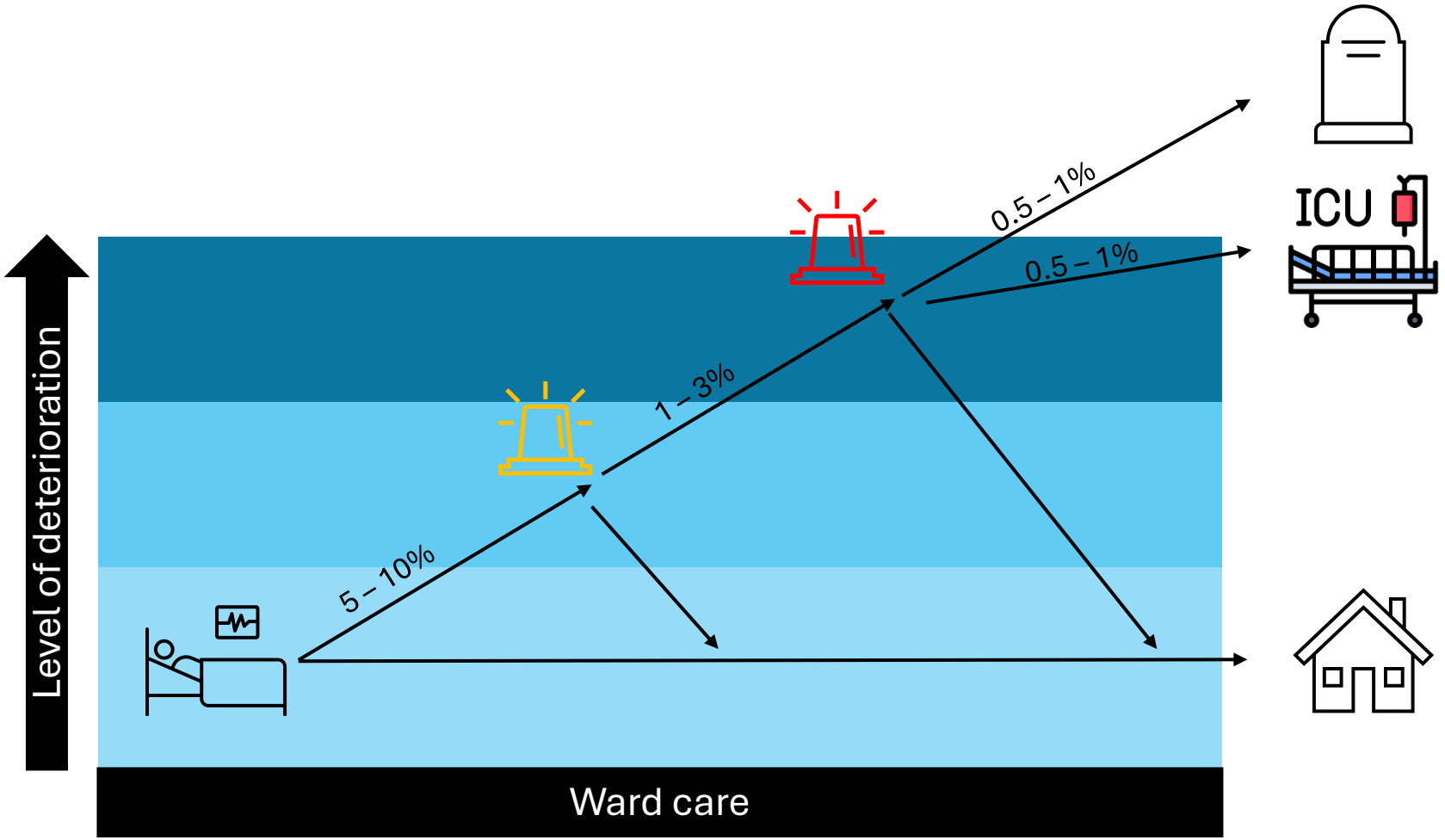
Hospital response to patient deterioration

Implement an Early Warning System:

1. Afferent arm to detect/identify deteriorating patients
2. Efferent arm to provide additional care to deteriorating patients

Early warning system: Afferent arm

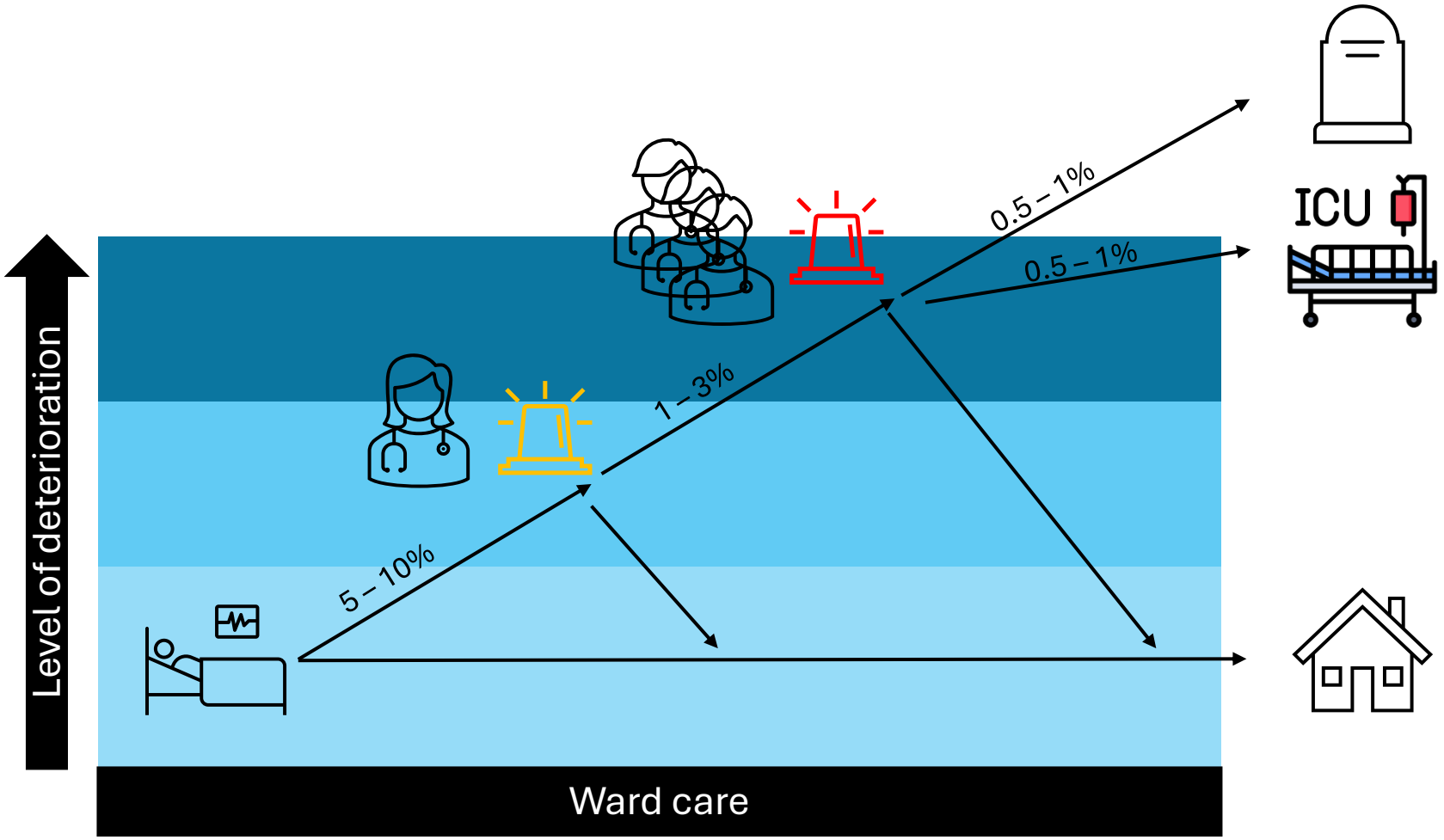
1. Early warning tools (EWTs) based on vital sign capture
2. Paper or EMR-based
3. Alert staff to patient deterioration



1. Jones D, Mitchell I, Hillman K, Story D. Defining clinical deterioration. Resuscitation. 2013;84(8):1029-1034. doi:10.1016/j.resuscitation.2013.01.013

Early warning system: efferent arm

1. Escalate care as patient deteriorates
2. Created Rapid Response teams (RRTs) or Medical Emergency teams (METs) to decide on earlier move to ICU



1. Jones D, Mitchell I, Hillman K, Story D. Defining clinical deterioration. Resuscitation. 2013;84(8):1029-1034. doi:10.1016/j.resuscitation.2013.01.013

Early Warning tools – what are they?

- 100% rule-based in Australia, NZ & UK
  - Use vitals captured every 8hr or so
  - Assign scores based on vital ranges
  - Total score gives the level of deterioration – the higher the score the more deteriorated
- We found 14 implemented AI-based EWTs internationally

Physiological Parameters	Score						
	3	2	1	0	1	2	3
Respiration Rate (per minute)	≤8		9-11	12-20		21-24	≥25
SpO2 (%)	≤91	92-93	94-95	≥96			
Air or oxygen		oxygen		air			
Systolic Blood Pressure (mmHg)	≤90	91-100	101-110	111-219			≥220
Pulse (per minute)	≤40		41-50	51-90	91-110	110-130	≥131
consciousness				A			VPU
Temperature (ºC)	≤35.5		35.1-36.0	36.1-38.0	38.1-39.0	≥39.1	

1. Royal College of Physicians L. National early warning score (NEWS): standardising the assessment of acute-illness severity in the NHS---report of a working party. In: ; 2012.
2. Van Der Vegt AH, Campbell V, Mitchell I, et al. Systematic review and longitudinal analysis of implementing Artificial Intelligence to predict clinical deterioration in adult hospitals: what is known and what remains uncertain. *Journal of the American Medical Informatics Association*. 2024;31(2):509-524. doi:10.1093/jamia/ocad220

Early Warning tools – which are best?

- Two most important things:

    1. Sensitivity: how many of the deteriorating patients does the EWT flag?

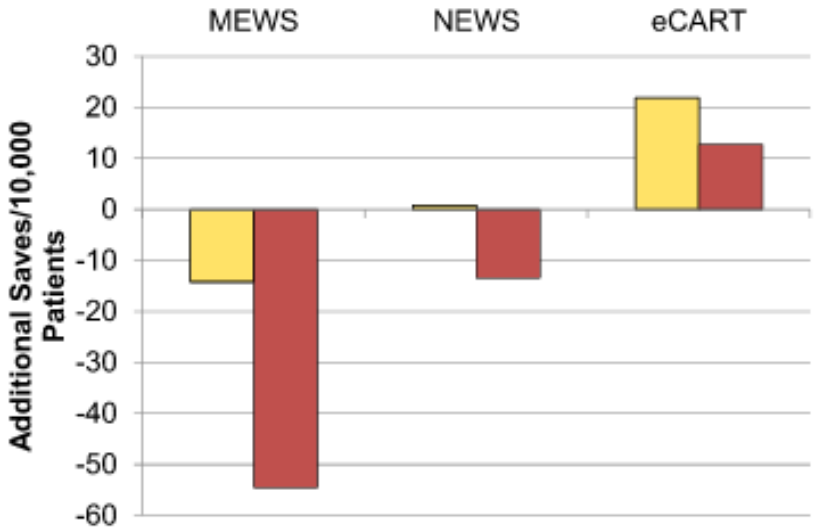
    2. False alerts: how many of the alerts correctly identify deteriorated patients
        - Usually described as precision or positive predictive value

Early Warning tools – which are best?

- Two most important metrics:

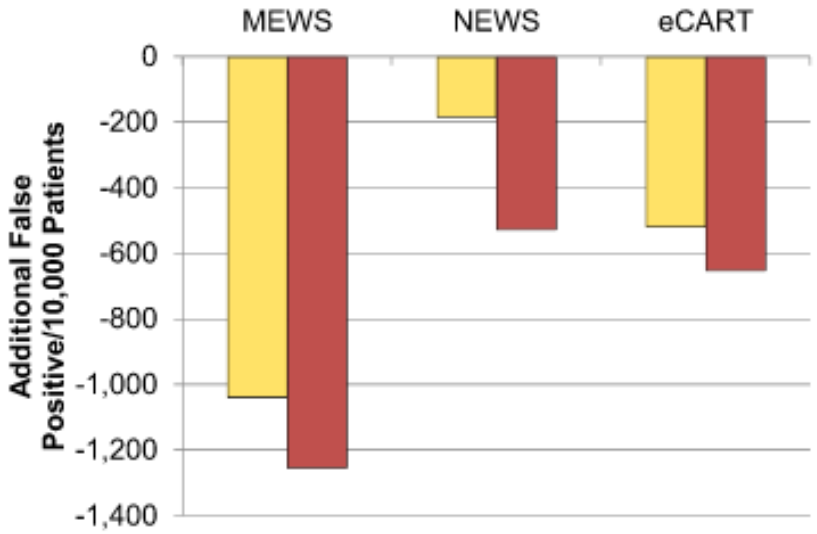
    1. Sensitivity: how many of the deteriorating patients does the EWT flag?

    2. False alerts: how many of the alerts correctly identify deteriorated patients
        - Usually described as precision or positive predictive value



Sensitivity

Fig. 1. Additional Saves Compared to BTF.

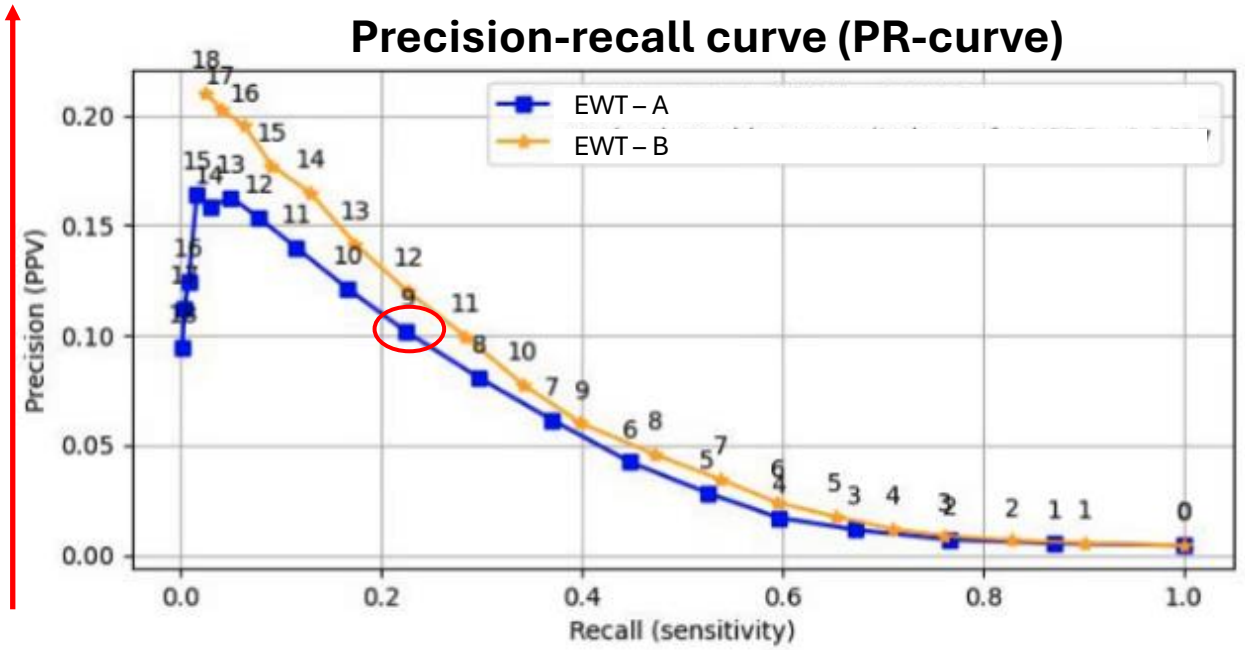


Precision – positive predictive value

1.Green M, Lander H, Snyder A, Hudson P, Churpek M, Edelson D. Comparison of the Between the Flags calling criteria to the MEWS, NEWS and the electronic Cardiac Arrest Risk Triage (eCART) score for the identification of deteriorating ward patients. *Resuscitation*. 2018;123:86-91. doi:10.1016/j.resuscitation.2017.10.028

Early Warning tools – which are best?

Higher precision = less false positives



The PR-curve plots the positive predictive value and sensitivity at each possible score of the EWT, e.g., at a score of 9, EWT-A:

- Identifies about 21% (Sensitivity=0.21) of the patients who die or going to ICU
- Has a precision of 10% (PPV = 0.10), which means that 1 in 10 alerts correctly identify patients who die or go to ICU and the other 9 are false positives

Higher sensitivity = Find more patients who have outcome (die or go to ICU)

Early Warning tools – which are best?

When implementing AI-based EWTs, alert rate (not sensitivity or false alarm rate) was used to decide where to set the alert threshold:

- *'of 19 Artificial Intelligence (AI) based EWT implementation studies, 16 reported using this approach, with* ***six studies setting the alert rates between 3 – 12 per 100 patient days****'*

Paper	Target (actual) sensitivity	Target (actual) PPV	Target (actual) specificity	Targeted Alerts per day (actual)	Primary objective / comment
Bailey et al. 2013 Kollef et al. 2014	40		97.6	1-2 alerts per nursing unit/day	Manageable number of alerts
Hackmann et al. 2011			95%		Manageable number of alerts per hospital floor per day
Lee et al., 2021				High sensitivity, low false alarm rate.	
Dziadzko et al., 2018	(63%)	(21%)			
Kang et al., 2016 Churpek et al., 2014	(60%)		95%		00 Revised and used PPV rather than sensitivity to set alert
Winslow et al., 2022				10 patients with red scores/day across the four-hospital system.	patients
Levin et al., 2022	(40%)	(13%)	(0.78%)	eight primary team and six RRT alerts / day (14 alerts total, seven per intervention unit)	Sensitivity vs mean alarm count per day
Martinez et al., 2022					clinically sustainable alert frequency, minimize clinical burden & alert fatigue
Kipnis et al., 2016 Escobar et al., 2020	(49%)			one new alert /day / 35 patients	
Lisk et al., 2020					mitigate the highest risk and not generate alert fatigue
Drummett et al., 2016	(25%)		(98%)		a level of workload that was felt to be acceptable by clinicians
O'Brien et al., 2019		10% (red)			
Pou-Prom et al., 2022 Nestor et al., 2020 Verma et al., 2021	(50%)	40%			clinicians expressed the need to minimize false alerts, and they recommended a ratio of 2 false alerts to a single true positive
Romero-Brufau et al., 2021				1 alert /day/10 patients	To reduce risk of alert fatigue

1. Van Der Vegt AH, Campbell V, Mitchell I, et al. Systematic review and longitudinal analysis of implementing Artificial Intelligence to predict clinical deterioration in adult hospitals: what is known and what remains uncertain. *Journal of the American Medical Informatics Association*. 2024;31(2):509-524. doi:10.1093/jamia/ocad220

Early Warning tools – which are best?

- Evidence is suggesting that workload capacity – the ability of your hospital's clinical staff to handle different alert levels – is an important factor in deciding which EWT to use and where to set the alert trigger points

- But current evaluation/comparison methods of sensitivity and precision does not capture this

Our study

Incorporating workload capacity into Early Warning Tool comparison and evaluation

Patient cohort

- Adult inpatients attending 11 hospitals across Queensland, all with Cerner EMR between Jan 2016 – June 2020
- 4 city, 7 regional hospitals
- Ward patients with at least one vital sign set. Average 5.3hrs per set
- Patient consent waiver: Townsville Hospital and Health Service Human Research Ethics Committee (HREC/QTHS/67897)

	Patients (N= 316,667)	Admissions (N = 683,617)	Episodes (N = 750,381)
Median age at admission (IQR)		60 (43,73)	
Male (%)	155,082 (49.0)		
Female (%)	161,585 (51.0)		
Median length of stay in hrs (IQR)		31.8 (6,96)	17.6 (2,66)
Outcomes: No. (%)			
- In-hospital death		4,954 (0.72)	4,954 (0.66)
- Transfer to ICU		3,400 (0.50)	3,717 (0.50)
ICU transfers per admission (% of transfers) - 1 transfer - 2 to 4 transfers - > 4 transfers		 3,167 (93) 227 (6.8) 6 (0.2)	
Number of vital sign sets			10,753,736

Which early warning tools we compared

- Latest standardized releases
- Australian Tools:
  - Between-the-flags (BTF)
  - Queensland Adult Deterioration Detection System (Q-ADDS)
  - Australian Capital Territory Modified Early Warning Score (MEWS)
- UK tools:
  - National Early Warning Score –2 (NEWS2)

Alert threshold	Escalation required	Q-ADDS	NEWS2	BTF	MEWS
Low-risk	Junior ward doctor review/ increase observations	4-5	5 – 6 5 – 6 +*	Yellow alert	4 - 5
Moderate-risk	Senior ward doctor review/ increase observations	6 - 7			6 - 7
MET-level	Acute or Critical care team review	≥ 8 or a single vital sign in the purple zone	≥ 7	Red alert	≥ 8 or a single vital sign in the purple zone

Evaluation outcomes

In-line with previous studies, ability of the EWT to identify patients that:

- Die in the ward
- Are transferred to ICU from the ward

	Patients (N= 316,667)	Admissions (N = 683,617)	Episodes (N = 750,381)
Median age at admission (IQR)		60 (43,73)	
Male (%)	155,082 (49.0)		
Female (%)	161,585 (51.0)		
Median length of stay in hrs (IQR)		31.8 (6,96)	17.6 (2,66)
Outcomes: No. (%)			
- In-hospital death		4,954 (0.72)	4,954 (0.66)
- Transfer to ICU		3,400 (0.50)	3,717 (0.50)
ICU transfers per admission (% of transfers) - 1 transfer - 2 to 4 transfers - > 4 transfers		 3,167 (93) 227 (6.8) 6 (0.2)	
Number of vital sign sets			10,753,736

1. Churpek MM, Yuen TC, Winslow C, et al. Multicenter development and validation of a risk stratification tool for ward patients. Am J Respir Crit Care Med. 2014;190(6):649-655. doi:10.1164/rccm.201406-1022OC

Introduction of workload capacity

Based on systematic review findings:

i. Low workload capacity

- Able to support 2 alerts per 100 patient days, such as where a medical emergency team (MET) has to cover the whole hospital;

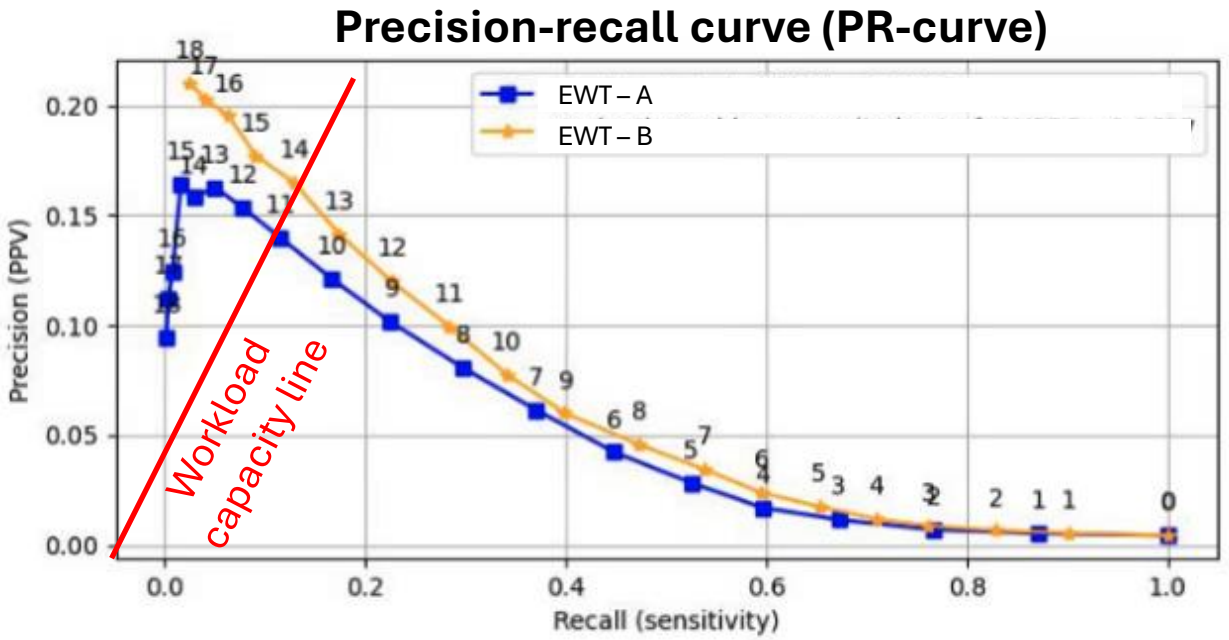
ii. Medium workload capacity

- Able to support 6 alerts per 100 patient days, such as a senior level ward doctor responsible for 20-30 patients; and

iii. High workload capacity

- Able to support 12 alerts per 100 patient days, such as ward nurse and doctor team

Introduction of workload capacity



- The workload capacity line represents a fixed alert frequency, e.g., 2 alerts per 100 patient days
- For example using the workload capacity line drawn on the graph:
  - A score of 11 for EWT–A has the same workload capacity as a score of 14 for EWT-B
  - At this same capacity, it's clear that EWT-B is superior because it has higher PPV for similar sensitivity

Evaluation methods

1. Traditional comparison
   - Table of sensitivity, positive predictive value, specificity at each alert threshold (low, medium and high)

2. Novel comparison
   - Integrate workload capacity onto precision-recall curve

Results & discussion

Junior ward doctor & nurses

Tools	Score	Admissions with alerts above threshold (%) N = 683,617	Vital sign sets above threshold (%) N = 10,753,736	Sensitivity	Specificity	Positive Predictive Value (PPV)
Low-risk threshold:						
BTF	Yellow	324,896 (48%)	2,718,484 (25.3%)	**0.696**	0.749	0.014
MEWS	≥ 4	83,700 (12%)	521,411 (4.8%)	0.456	**0.954**	**0.047**
Q-ADDS	≥ 4	91,276 (13%)	549,929 (5.1%)	0.485	0.951	**0.047**
NEWS2	≥ 5	94,773 (14%)	806,041 (7.5%)	0.563	0.927	0.037
NEWS2-plus	≥ 5	143,710 (21%)	1,149,253 (10.7%)	0.616	0.896	0.029
Moderate-risk threshold:						
MEWS	≥ 6	44,679 (6,5%)	221,394 (2.1%)	0.288	**0.981**	0.069
Q-ADDS	≥ 6	39,974 (5.8%)	218,435 (2.0%)	**0.319**	**0.981**	**0.078**
MET threshold:						
BTF	Red	70,281 (10%)	364,323 (3.4%)	0.302	0.967	0.044
MEWS	≥ 8	37,662 (5.5%)	165,932 (1.5%)	0.205	0.986	0.067
Q-ADDS	≥ 8	29,334 (4.3%)	143,588 (1.3%)	0.221	**0.988**	**0.082**
NEWS2	≥ 7	37,867 (5.5%)	285,287 (2.7%)	**0.377**	0.975	0.070

Traditional Results

- At the low risk threshold for the tools, BTF has the highest sensitivity, but lowest PPV
- MEWS and Q-adds have the highest PPV, but lowest sensitivity

Tools	Score	Admissions with alerts above threshold (%) N = 683,617	Vital sign sets above threshold (%) N = 10,753,736	Sensitivity	Specificity	Positive Predictive Value (PPV)
Low-risk threshold:						
BTF	Yellow	324,896 (48%)	2,718,484 (25.3%)	**0.696**	0.749	0.014
MEWS	≥ 4	83,700 (12%)	521,411 (4.8%)	0.456	**0.954**	**0.047**
Q-ADDS	≥ 4	91,276 (13%)	549,929 (5.1%)	0.485	0.951	**0.047**
NEWS2	≥ 5	94,773 (14%)	806,041 (7.5%)	0.563	0.927	0.037
NEWS2-plus	≥ 5	143,710 (21%)	1,149,253 (10.7%)	0.616	0.896	0.029
Moderate-risk threshold:						
MEWS	≥ 6	44,679 (6,5%)	221,394 (2.1%)	0.288	**0.981**	0.069
Q-ADDS	≥ 6	39,974 (5.8%)	218,435 (2.0%)	**0.319**	**0.981**	**0.078**
MET threshold:						
BTF	Red	70,281 (10%)	364,323 (3.4%)	0.302	0.967	0.044
MEWS	≥ 8	37,662 (5.5%)	165,932 (1.5%)	0.205	0.986	0.067
Q-ADDS	≥ 8	29,334 (4.3%)	143,588 (1.3%)	0.221	**0.988**	**0.082**
NEWS2	≥ 7	37,867 (5.5%)	285,287 (2.7%)	**0.377**	0.975	0.070

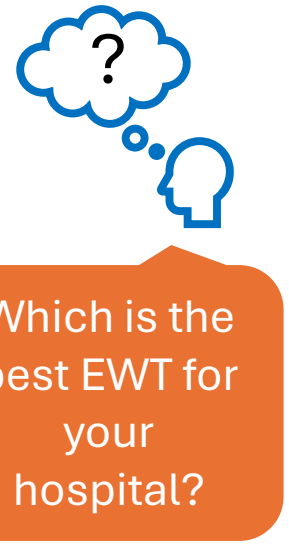
MET team

Traditional Results

- At the MET level threshold for the tools, Q-adds has the highest PPV, but NEWS2 has the greatest sensitivity

Tools	Score	Admissions with alerts above threshold (%) N = 683,617	Vital sign sets above threshold (%) N = 10,753,736	Sensitivity	Specificity	Positive Predictive Value (PPV)
Low-risk threshold:						
BTF	Yellow	324,896 (48%)	2,718,484 (25.3%)	**0.696**	0.749	0.014
MEWS	≥ 4	83,700 (12%)	521,411 (4.8%)	0.456	**0.954**	**0.047**
Q-ADDS	≥ 4	91,276 (13%)	549,929 (5.1%)	0.485	0.951	**0.047**
NEWS2	≥ 5	94,773 (14%)	806,041 (7.5%)	0.563	0.927	0.037
NEWS2-plus	≥ 5	143,710 (21%)	1,149,253 (10.7%)	0.616	0.896	0.029
Moderate-risk threshold:						
MEWS	≥ 6	44,679 (6,5%)	221,394 (2.1%)	0.288	**0.981**	0.069
Q-ADDS	≥ 6	39,974 (5.8%)	218,435 (2.0%)	**0.319**	**0.981**	**0.078**
MET threshold:						
BTF	Red	70,281 (10%)	364,323 (3.4%)	0.302	0.967	0.044
MEWS	≥ 8	37,662 (5.5%)	165,932 (1.5%)	0.205	0.986	0.067
Q-ADDS	≥ 8	29,334 (4.3%)	143,588 (1.3%)	0.221	**0.988**	**0.082**
NEWS2	≥ 7	37,867 (5.5%)	285,287 (2.7%)	**0.377**	0.975	0.070

MET team

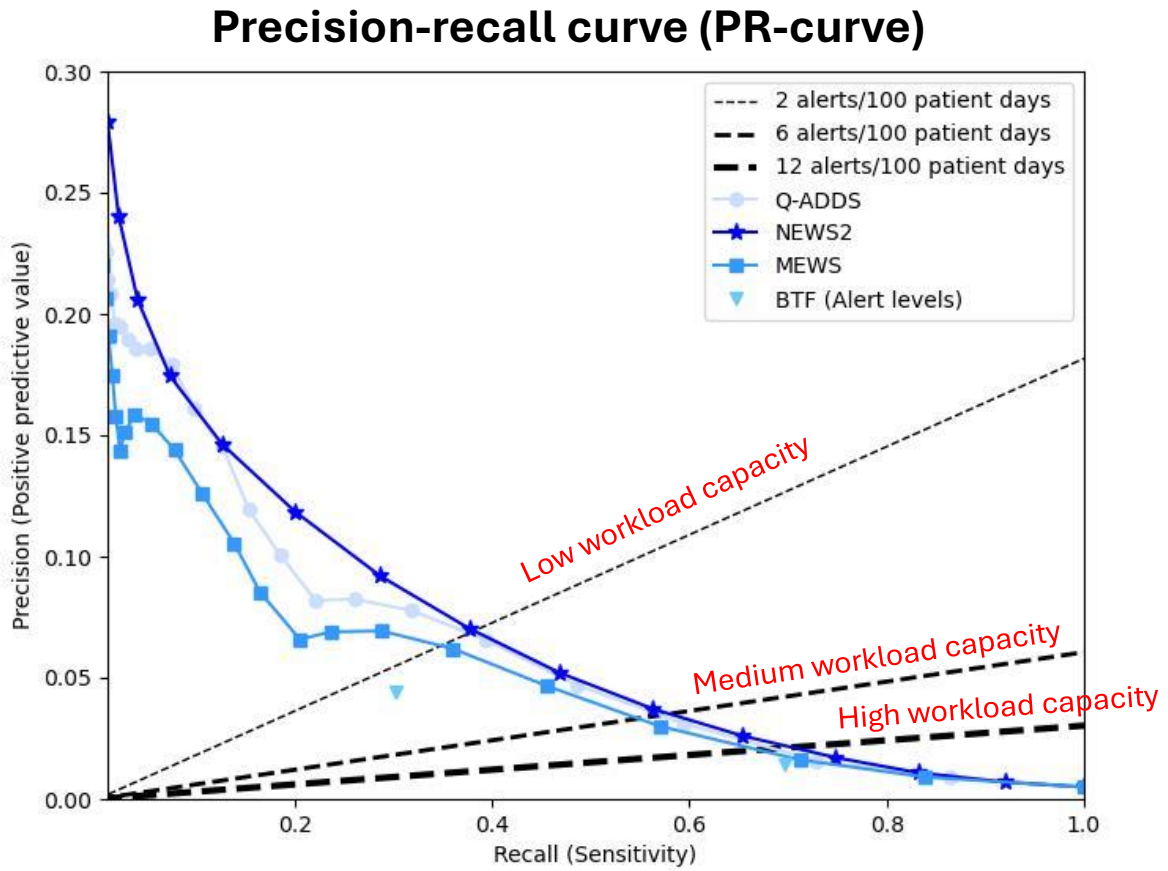


Traditional Results

- At the MET level threshold for the tools, Q-adds has the highest PPV, but NEWS2 has the greatest sensitivity

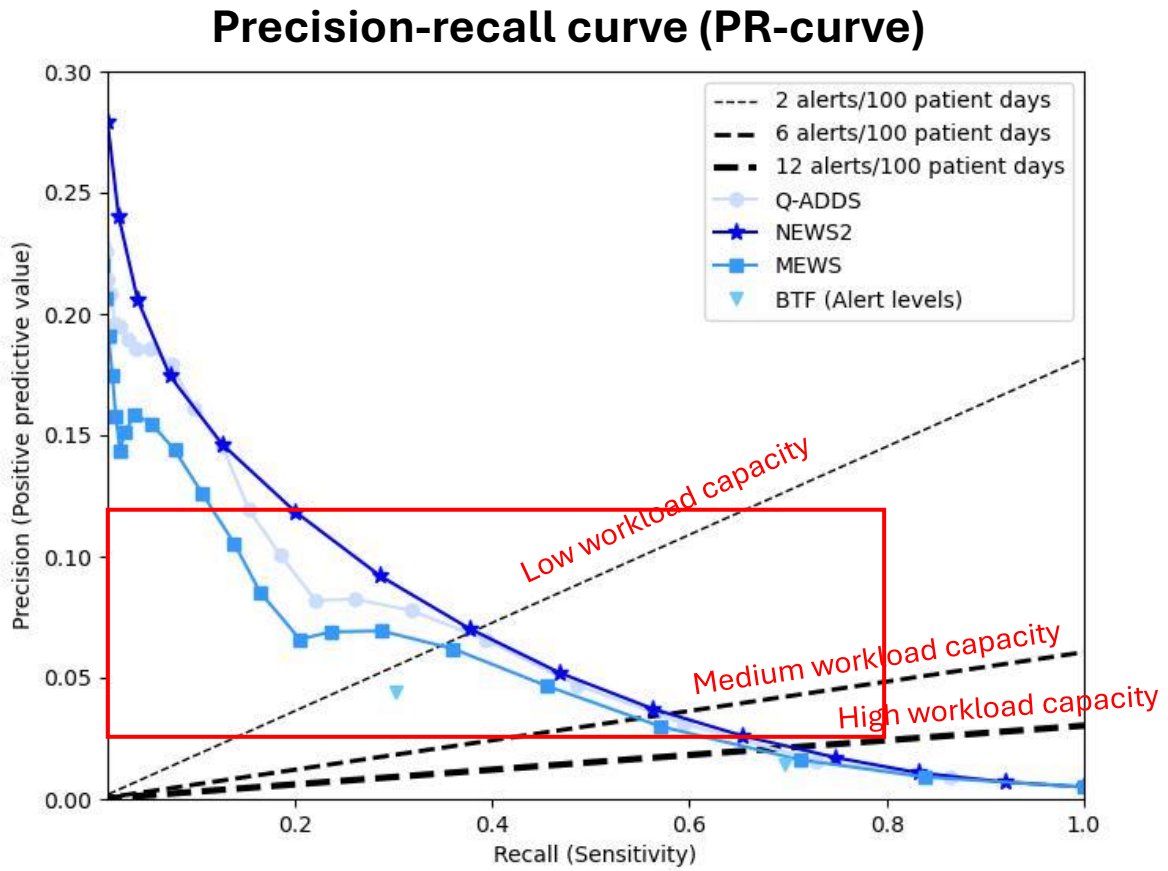
Results integrating workload capacity

- Overall, score-for-score, NEWS2 has higher sensitivity and precision than the other tools
- Note that BTF is represented by 2 points (rather than a curve)
- At this level, it's hard to see what the workload capacity lines show us



Results integrating workload capacity

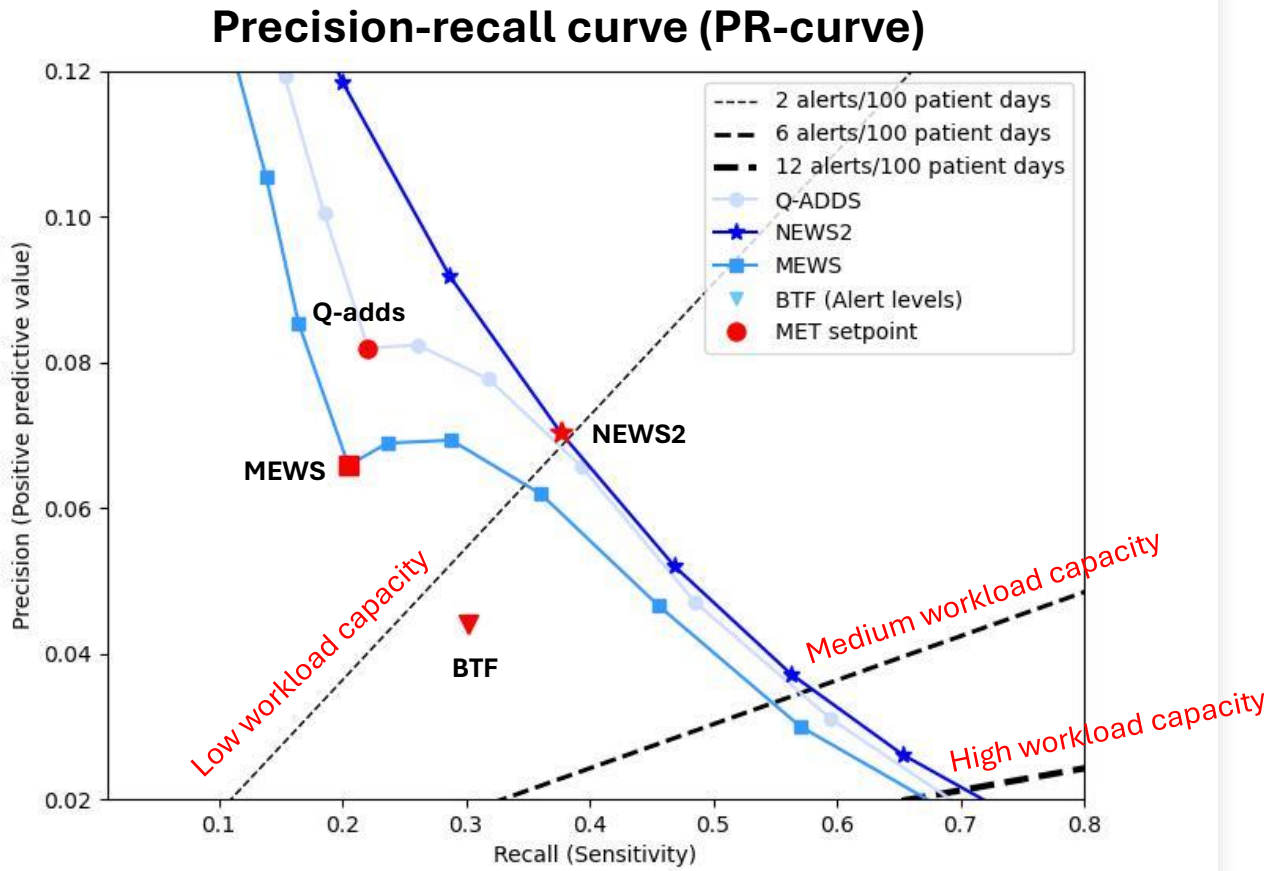
- Overall, score-for-score, NEWS2 has higher sensitivity and precision than the other tools
- Note that BTF is represented by 2 points (rather than a curve)
- At this level, it's hard to see what the workload capacity lines show us



Results integrating workload capacity

At the MET level (red dots highlighted):

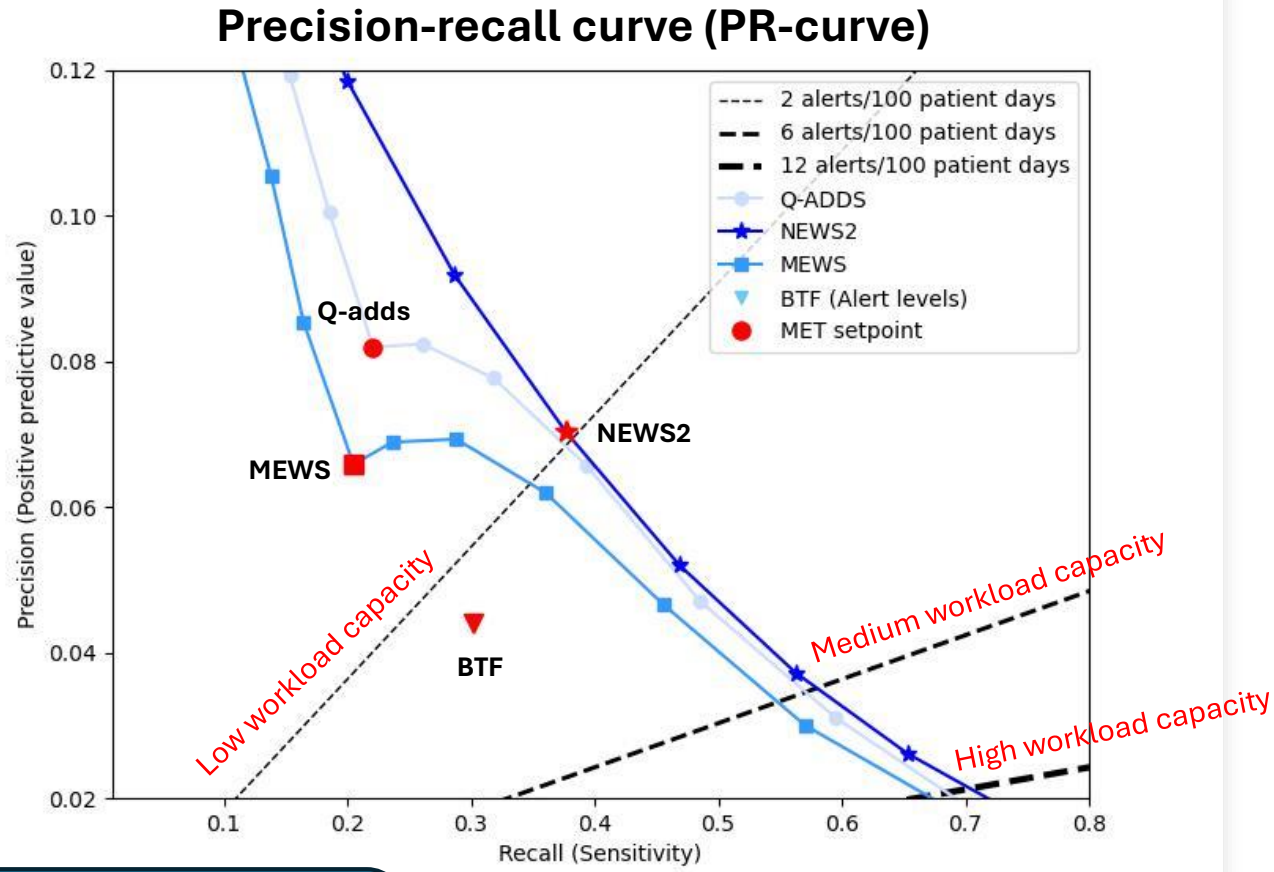
- NEWS2 is sitting on the Low workload capacity line of 2.2 alerts per 100 pats.
- Q-ADDS/MEWS are ~41% less sensitive than NEWS2, but operate at half the workload capacity (0.98 alerts/100 pats)
- BTF has middling sensitivity, but very low precision and the highest workload capacity



Results integrating workload capacity

At the MET level (red dots highlighted):

- NEWS2 is sitting on the Low workload capacity line of 2.2 alerts per 100 pats.
- Q-ADDS/MEWS are ~41% less sensitive than NEWS2, but operate at half the workload capacity (0.98 alerts/100 pats)

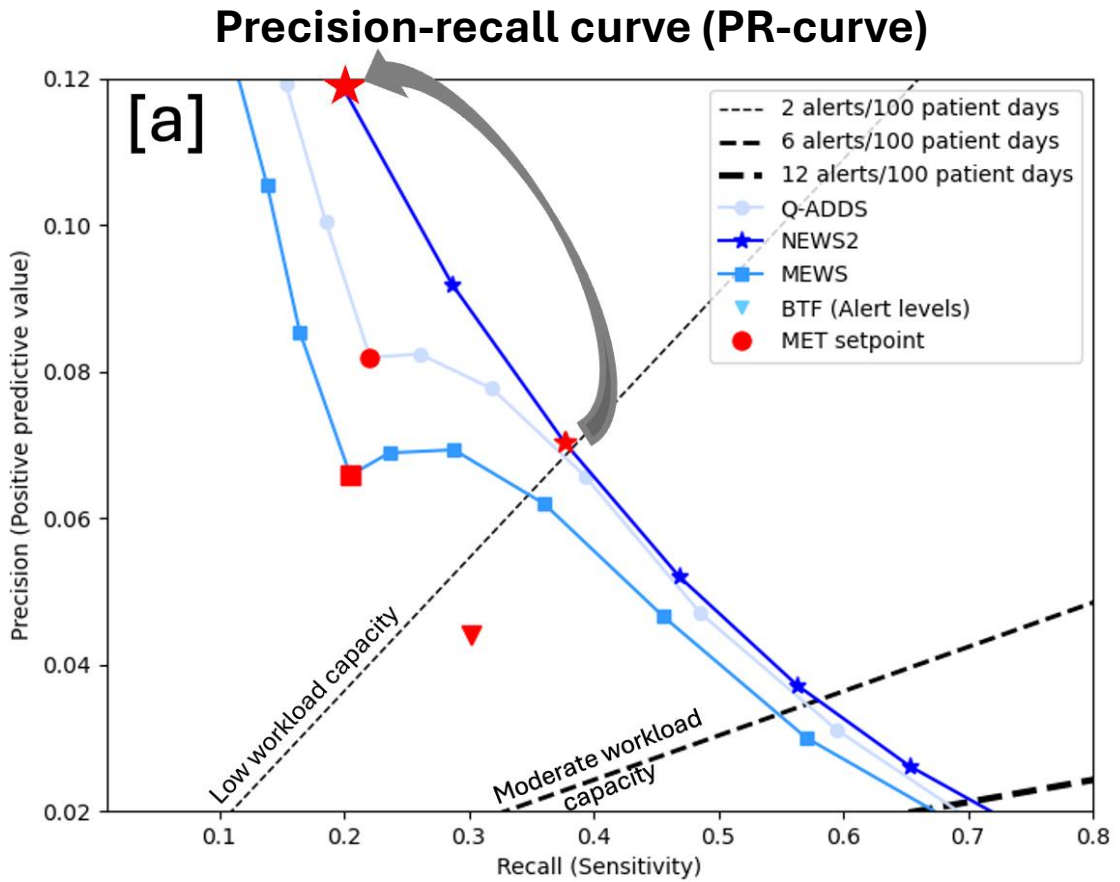


The selection of MET level trigger score appears to be a choice that trades off workload capacity for sensitivity....

Results integrating workload capacity

For example:

- NEWS2 could be operated at a MET level trigger score of 9 instead of 7 and then it would have similar sensitivity to Q-adds and MEWS, far greater PPV and lower workload capacity requirements



Key findings (1)

If you look across all alert thresholds (low, medium & MET), ranking of EWTs by alert rate (workload capacity) was the same at each alert threshold, i.e., at early, medium and MET level

- Q-ADDS and MEWS had the lowest alert rate
- NEWS2 was in the middle
- BTF had the highest alert rate

Key findings (1)

Why do some hospitals use higher sensitivity EWTs and others trade this off for lower workload capacity?

If you look across all alert thresholds (low, medium & MET), ranking of EWTs by alert rate (workload capacity) was the same at each alert threshold, i.e., at early, medium and MET level

- Q-ADDS and MEWS had the lowest alert rate
- NEWS2 was in the middle
- BTF had the highest alert rate

Key findings (1)

Why do some hospitals use higher sensitivity EWTs and others trade this off for lower workload capacity?

If you look across all alert thresholds (low, medium & MET), ranking of EWTs by alert rate (workload capacity) was the same at each alert threshold, i.e., at early, medium and MET level

- Q-ADDS and MEWS had the lowest alert rate
- NEWS2 was in the middle
- BTF had the highest alert rate

Hospitals may:

(i) Implement alert mitigation methods, (e.g., manual patient-specific threshold adjustments, or system alert suppressions), enabling a higher overall alert level system to become manageable;

(ii) Provide generous staffing ratios and/or dedicated MET responders (high workforce capacity), or

(iii) Implement multiple tiers of escalation responsibility, which shares the alert burden while improving efficiency of the EWS as a whole

Key findings (2)

- Although the performance (PR curves and other metrics) were very similar between NEWS2 and Q-ADDS, their applied operating characteristics (PPV, sensitivity and alert burden) were quite different at each threshold of alerting (MET-level to low-risk)

Key findings (2)

- Although the performance (PR curves and other metrics) were very similar between NEWS2 and Q-ADDS, their applied operating characteristics (PPV, sensitivity and alert burden) were quite different at each threshold of alerting (MET-level to low-risk)

This may reflect:

(i) Government mandates
(ii) International expert recommendations
(iii) Resource burden
(iv) Clinical need
(v) Evidence base of EWS effectiveness

Implications for hospitals who have EWTs

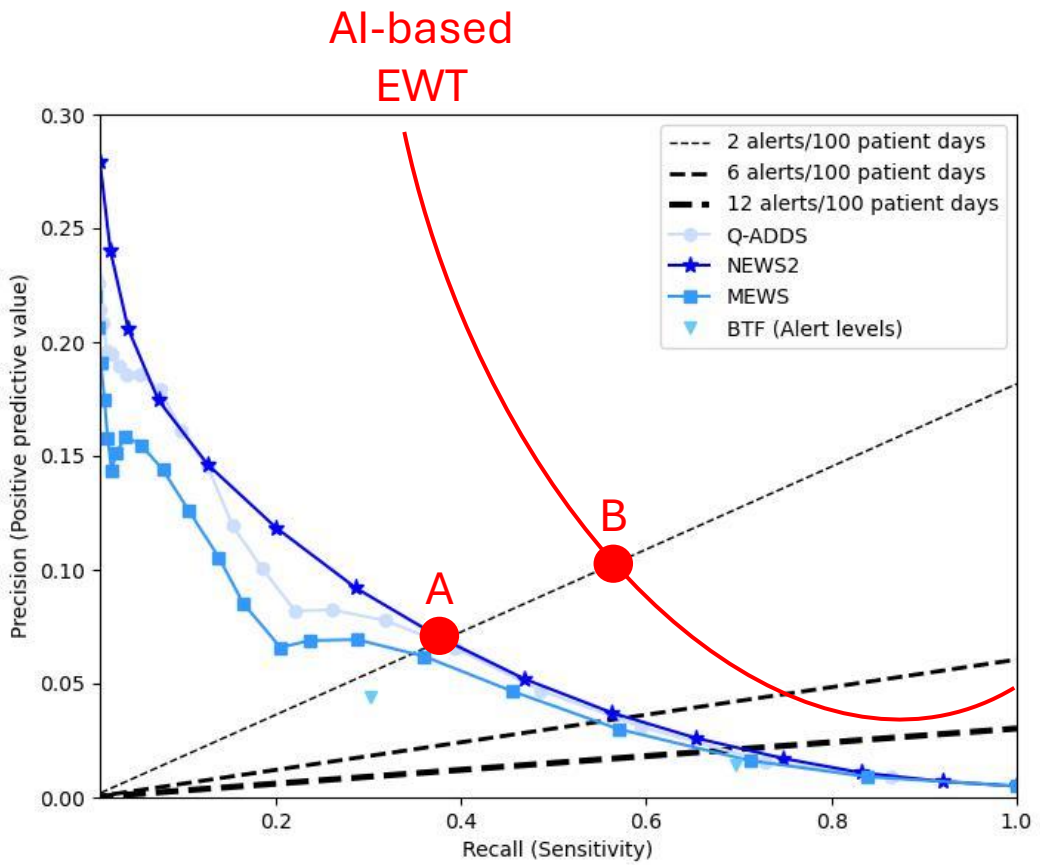
1. Moving to an EWT that is more effective, i.e., that provides greater sensitivity for the same workload capacity.

2. Alter the threshold score of your existing tool to improve sensitivity, if you have the resourcing to support the higher workload capacity

Implications for hospitals considering AI-based EWTs

- Unlike rule-based EWTs, you can select the alert trigger points anywhere along its PR curve. You can use the workload capacity lines and existing EWT thresholds to determine the most suitable alert points for your hospital

- For example, if you introduce a new AI-based EWT, you can set the trigger threshold at B, which will yield exactly the same alert frequency (workload) for a greater PPV and sensitivity



Thank you

Questions