BMJ Open

Comparing the predictive ability of a commercial artificial intelligence early warning system with physician judgement for clinical deterioration in hospitalised general internal medicine patients: a prospective observational study

Jonathan Arnold [1], Alex Davis,[2] Baruch Fischhoff,[2] Emmanuelle Yecies,[1] Jon Grace,[3] Andrew Klobuka,[4] Deepika Mohan,[5] Janel Hanmer[1]

To cite: Arnold J, Davis A, Fischhoff B, *et al*. Comparing the predictive ability of a commercial artificial intelligence early warning system with physician judgement for clinical deterioration in hospitalised general internal medicine patients: a prospective observational study. *BMJ Open* 2019;**9**:e032187. doi:10.1136/bmjopen-2019-032187

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-032187).

Received 06 June 2019
Revised 07 August 2019
Accepted 30 August 2019

For numbered affiliations see end of article.

Correspondence to
Dr Jonathan Arnold; arnoldjd@pitt.edu

ABSTRACT

Objective Our study compares physician judgement with an automated early warning system (EWS) for predicting clinical deterioration of hospitalised general internal medicine patients.

Design Prospective observational study of clinical predictions made at the end of the daytime work-shift for an academic general internal medicine floor team compared with the risk assessment from an automated EWS collected at the same time.

Setting Internal medicine teaching wards at a single tertiary care academic medical centre in the USA.

Participants Intern physicians working on the internal medicine wards and an automated EWS (Rothman Index by PeraHealth).

Outcome Clinical deterioration within 24 hours including cardiac or pulmonary arrest, rapid response team activation or unscheduled intensive care unit transfer.

Results We collected predictions for 1874 patient days and saw 35 clinical deteriorations (1.9%). The area under the receiver operating curve (AUROC) for the EWS was 0.73 vs 0.70 for physicians (p=0.571). A linear regression model combining physician and EWS predictions had an AUROC of 0.75, outperforming physicians (p=0.016) and the EWS (p=0.05).

Conclusions There is no significant difference in the performance of the EWS and physicians in predicting clinical deterioration at 24 hours on an inpatient general medicine ward. A combined model outperformed either alone. The EWS and physicians identify partially overlapping sets of at-risk patients suggesting they rely on different cues or decision rules for their predictions.

Trial registration number NCT02648828.

Strengths and limitations of this study

- This is a pragmatic prospective observational study of physician decision-making in a real-world clinical setting compared head-to-head with a commercially available automated early warning system.
- Our study is set in the general internal medicine teaching services of a single large tertiary care academic medical centre.
- We include all patients admitted to the general internal medicine teaching services who are eligible for activation of the rapid response team (ie, are not ordered for 'comfort measures only' care).
- Our hospital had previously deployed and the vendor had optimised the automated early warning system (Rothman Index) in our clinical environment as a separate pilot project.
- Our study is limited to the predictive ability of physicians working in this clinical environment, and it does not necessarily generalise to other clinical settings (eg, surgical services, hospitalist services, non-teaching settings).

BACKGROUND AND SIGNIFICANCE

Identifying patients at risk for clinical deterioration is essential for prioritising attention and resources in a hospital setting. Although such identification has historically been the responsibility of medical staff, there is now a drive to automate the identification of at-risk patients with automated early warning systems (EWSs).[1–3] EWSs are designed to detect changes in vital signs or clinical condition that precede clinical deterioration, standardising the assessment of patient stability.[4 5] The discrimination ability of EWSs has not been compared directly to physician judgement on general medicine wards, leaving the relative accuracy and sensitivity of EWSs and medical staff unknown. Nor has the joint performance of physicians and EWSs been studied, examining the extent to which they complement one another.

Prior studies have demonstrated the ability of EWSs to identify patients at high risk for clinical deterioration.[6–11] These studies have assessed the predictive ability of EWSs using the area under the receiving operator curve (AUROC), which measures a test's ability to discriminate positive and negative cases across multiple cut-off thresholds.[12] An AUROC of 0.5 is the same as chance, and 1.0 is perfect discrimination. A study of VitalPac early warning score (ViEWS) found an AUROC of 0.89 (95% CI 0.88 to 0.90) for predicting 24 hours mortality in a non-intensive care unit (ICU) hospitalised medical population.[8] A study of Cardiac Arrest Risk Triage (CART) found an AUROC of 0.84 for predicting cardiac arrest at 48 hours in non-ICU hospitalised patients.[9] The Rothman Index (RI) is a commercial automated EWS, based on a form of artificial intelligence that uses electronic health record (EHR) data to update itself in near real-time; it has a reported AUROC of 0.93 for predicting 24 hours mortality in a population of medical, surgical and ICU patients.[13] Despite these demonstrations of their ability to identify patients at risk, EWSs have not been shown to affect mortality, cardiac arrest or hospital length of stay in adult patients.[14–16] Additionally, a recent large, multicountry randomised control trial demonstrated no decrease in all-cause mortality with implementation of the Bedside Pediatric Early Warning System.[17]

Physicians' ability to predict, on admission to the ICU, which patients will ultimately die has been well studied, with a pooled AUROC of 0.85 across eight studies.[18] Studies in general medical wards have found that physicians have an AUROC of 0.69–0.84 for predicting clinical deterioration at 24 hours.[19 20] Prospectively comparing an EWS to physician prediction, the SUPPORT prognostic model demonstrated equal predictive ability for 180-day mortality compared with ICU physicians and AUROC 0.78 for both.[21] There have, however, been few, if any, studies comparing the predictive performance of EWSs with physician judgement outside of the ICU. The present study directly compares the performance of physicians and an EWS in predicting clinical deterioration for patients admitted to an adult general medicine ward.

METHODS

Study design

We conducted a prospective observational study comparing physicians and an automated EWS in predicting patient clinical deterioration within 24 hours. We conducted the study between July and December 2015 on the academic general medicine floor services of a single 792-bed academic urban tertiary care referral medical centre. Patients on the general medicine floor services are from the local community of the hospital or transferred from other facilities for access to specialty care. Patients are either triaged to the general medicine floor service from the emergency department, directly admitted to the service by a physician with admitting privileges or transferred from an ICU once they no longer required that level of care.

At this medical centre, the rapid response team (RRT) is activated for 'condition A' indicating cardiac or pulmonary arrest or 'condition C' indicating clinical instability. Any staff member can activate the RRT. Nursing criteria for calling a 'condition C' include new onset tachycardia, tachypnoea, increasing oxygen requirement or altered mental status; however, clinical staff has the flexibility to call a condition C whenever they are concerned about patient stability. Medical staff can also transfer patients to the ICU without activating the RRT. In our hospital, patients are transferred to the ICU, with or without a condition C, when they require more than 6 L/min of oxygen, require initiation of mechanical ventilation (invasive or non-invasive), require vasopressors or invasive haemodynamic monitoring, require continuous renal replacement therapy or when they require nursing care more frequently than every 2 hours. We defined our outcome of clinical deterioration for this study as any condition C, condition A or ICU transfer. We excluded ICU transfers occurring immediately postoperatively as these are often preplanned or related to the surgical intervention or anaesthesia. We collected condition calls and ICU transfers from the clinical EHR after study completion.

Study subjects

Physicians

We included interns assigned to work on the general internal medicine teaching services as our physician subjects. Physicians were asked to make clinical predictions only for patients under their direct care. At this training centre at the time of data collection, there were a total of 44 internal medicine, 12 anaesthesia, 7 neurology and 16 transitional/preliminary-year interns who were scheduled for 4-week or 5-week inpatient blocks. Not all eligible interns were scheduled to work on the ward services during the study period. Residents assigned non-overlapping patients to interns. Not all patients were managed by an intern. Interns worked 6 days a week and signed out to an overnight team. We also collected predictions from resident and attending physicians; however, these were not included in this analysis as their predictive ability did not differ significantly from interns and their judgments were assumed to be non-independent of the physicians whom they supervised. Analyses are provided in online supplementary appendix 1.

Automated EWS

Prior to our study, the RI was integrated into the EHR and calibrated by ParaHealth in a separate pilot project. The details of the RI's algorithm are proprietary; however previously published work regarding its development highlight the use of EHR recorded vital signs, nursing assessments, laboratory data and cardiac monitoring data.[13] At the time of data collection, the RI output was accessible in a separate EHR screen displaying a 5-day line graph of scores over time, colour coded for current condition (blue, yellow or red) and accompanied by a risk flag (none, medium, high, very high).

Data collection

The research team collected physician predictions via in-person or phone-based interviews at the end of the day shift on a convenience sample of days, one to two times per week for 6 months. We surveyed physicians independently from each other to minimise contamination from one another's judgments. The research team asked each physician the following question for each patient under his or her care:

> What is the percent chance that this patient will have a condition A, condition C, or be transferred to the ICU within the next 24 hours? Please report your answer on a scale from 0 (definitely won't) to 100 (certainly will).

These questions followed guidelines for expert elicitation.[22 23] Physicians were not given any guidance regarding how to develop their predictions. The research team separately collected the RI colour code and risk assignments of each patient at the same time.

Data analysis

Standardisation of observations

We decided a priori to compare physician and EWS judgments only on patient-days where both were available to minimise the risk of selection bias. The RI was not immediately available for patients following admission or transferred between units and physicians were less likely to be available for data collection when patients were sick. Resident physicians covered intern's patients on their days off. We also excluded patients who were 'comfort measures only' at the time of risk assessment as they would not have a condition called or be transferred to the ICU.

Because the RI assessments were categorical and the physician judgments were continuous, we used the following procedures to compare them:

1. We created an algorithm posthoc for combining the RI colour category and flag to form a single risk judgement, described in online supplementary appendix 2. This approach maximised the discriminatory ability of the RI and had face validity for how the RI might be used in clinical practice. The risk assignment was:
   i. Low risk (RI-Low) if the colour was blue.
   ii. Medium risk (RI-Med) if the colour was yellow or red with no flag.
   iii. High risk (RI-High) if the colour was yellow or red with any flag.
2. Each physician's predictions were represented as standardised scores (z) for each clinical block, using that physician's personal mean (μ) and SD (σ) for that period, $z = (x - \mu)/\sigma$.
3. The pooled physician judgments were divided into high (MD-High), medium (MD-Med) and low (MD-Low) risk categories so that they had the same marginal distribution as the RI judgments from step 1.

Comparison of predictive ability

To compare the ability of the physicians and the RI to predict which patients would experience a clinical deterioration, we used an AUROC approach described by DeLong, DeLong and Clarke-Pearson.[24] A statistical power analysis indicated that we would need 35 clinical deterioration events to detect an AUROC difference of 0.12 with alpha=0.05 assuming 0.8 correlation between physicians and the RI. We additionally calculated the OR for predictions for three discriminations: (1) high vs low risk to evaluate the largest expected difference, (2) high versus not-high to evaluate the relative danger in being flagged as high risk and (3) low vs not-low to evaluate the relative safety in being identified as low risk. ORs were compared using the two-sample z-test.

We additionally performed a net reclassification analysis comparing physicians and the RI for high versus not-high risk assessment. This is presented in online supplementary appendix 3.

Joint prediction

We reported concordance and discordance between the RI and physician judgments and described deterioration rates. We included the RI and physician judgments as independent predictors of clinical deterioration in a logistic regression model. We compared this model's prediction ability, using AUROC, with that of both the RI and physicians using the χ^2 test. We assessed its goodness-of-fit with the outcomes using the likelihood ratio test.

In addition, online supplementary appendix 4 reports a sensitivity analysis comparing continuous physician judgments with the RI rather than stratifying physician judgements into levels matching the RI output.

We performed all statistical calculations using Stata 14. All reported CIs are 95th percentile. All tests are two-tailed.

Patient and public involvement

There was no patient or public involvement in this study. study registration, funding and data sharing.

Study registration, funding and data sharing

The study protocol was approved by the IRBs of Carnegie Mellon University (HS15-410) and the University of Pittsburgh (PRO15040582). The study protocol was registered with clinicaltrials.gov (NCT02648828). Study datasets are available by request to the corresponding author. Informed consent was obtained from all participating physicians prior to the start of the study. JH was supported by the National Institutes of Health through Grant Number KL2 TR001856. BF and AD were partially supported by the Riksbankens Jubileumsfond. JDA was funded by the Ruth L. Kirschstein National Research Service Award/Health Resources and Services Administration T-32, grant number T32HP22240. The funders had no role in design or conduct of the study; collection, management, analysis or interpretation of the data; preparation, review or approval of the manuscript; or decision

Table 1 Characteristics of physicians and patient-days included in the final analysis

Characteristic	Overall
Intern physicians	
Total interns, N (unique)	70 (59)
Female, N (%)	34 (49)
Training programme, N (%)	
Internal medicine	36 (51)
Anaesthesia	12 (17)
Neurology	7 (10)
Preliminary/Transitional year	15 (21)
Blocks included, N (%)	
1	49 (83)
2	9 (15)
3	1 (2)
Predictions per intern per block, median (IQR)	28 (25–34)
Patient-days	
Patient-days, N	1874
Unique patients, N	1106
Age in years, median (IQR)	56 (42–70)
Female, N (%)	939 (50.1%)
Length of stay at data collection, median (IQR)	4.2 (2–9.7)
Clinical deteriorations, N (%)	
Total	35 (1.9)
Condition A*	0 (0)
Condition C†	27 (1.4)
ICU transfer without condition call	8 (0.4)

*Condition A: rapid response team activation indicative of a cardiac or pulmonary arrest.
†Condition C: rapid response team activation for clinical deterioration or instability.
ICU, intensive care unit.

to submit the manuscript for publication. Requests for study data and analysis code should be directed to the corresponding author.

RESULTS

Study subjects and observations

We collected intern predictions and corresponding RI risk-assessments on 1874 patient-days. Patient-days were 50% female (n=939), median patient age was 57 years (IQR 42–70) and median length of stay at the time of data collection was 4.2 days (IQR 2.0–9.7). There were 1106 unique patients over these 1874 patient days. Total 35 patient-days met the study endpoint, a clinical deterioration rate of 1.9%. Table 1 describes the physicians, patients and clinical events included in the analysis. Refer to online supplementary appendix 5 for additional details about the patient-days included.

As described above, we stratified judgements by the physicians and EWS into risk category. We present the resulting judgement distributions, along with rates of clinical deteriorations, in table 2.

Comparison of predictive ability

The AUROC for predicting 24 hours clinical deterioration was 0.70 (CI 0.62 to 0.79) for physicians and 0.73 (CI 0.66 to 0.81) for the RI. Their sensitivity and specificity were similar at both cut-off thresholds. The difference in predictive ability was not statistically significant across the various measures. Table 3 presents these results. The confusion matrices for the physicians and the RI are presented in online supplementary appendix 6.

Joint prediction model

Physicians and the RI agreed on 61 patients being high risk, 7 of these had a clinical deterioration for an event rate of 11.5%. They agreed on 1478 patients being not-high risk; 15 of these patients had a clinical deterioration, for an event rate of 1.0%. They disagreed on 335 patients, with 161 rated high-risk by providers and 174 rated high risk by the RI. There were 13 clinical deteriorations in this group, a rate of 3.9%; these cases involved 6 of 161 ranked high-risk by physicians (3.7%) and 7 of 174 ranked high-risk by the RI (4.0%).

The AUROC for the joint model combining RI and physician judgement 0.78 (CI 0.70 to 0.85) was higher than for either rater independently. Table 4 presents the results of this model. Figure 1 is a Venn diagram of patients receiving high versus not high rating from the

Table 2 Stratified risk prediction and clinical deteriorations (events) by physicians and EWS

	Physicians		EWS	
Risk category*	**Total (%)**	**Events (rate)**	**Total (%)**	**Events (rate)**
Total	1874 (100%)	35 (1.9%)	1874 (100%)	35 (1.9%)
High	222 (11.9%)	13 (5.9%)	235 (12.5%)	14 (6.0%)
Medium	683 (36.5%)	15 (2.2%)	685 (36.6%)	16 (2.3%)
Low	969 (51.7%)	7 (0.7%)	954 (50.9%)	5 (0.5%)

*Risk categories were initially assigned to EWS assessments through a combination of coloured categories and variability flags. Continuous physician predictions were stratified to closely match the same marginal distribution.
EWS, early warning system.

Table 3 Comparison of physician and EWS predictive ability for 24 hours clinical deterioration

	Physicians	EWS	P value
AUROC (CI)	0.70 (0.62 to 0.79)	0.73 (0.66–0.81)	0.571*
Sensitivity/Specificity			
≥High	37.1%/88.6%	40.0%/88.0%	
≥Medium	80.0%/52.3%	85.7%/51.6%	
ORs (CI)			
High versus low	8.5 (3.4 to 21.7)	12.0 (4.3–33.7)	0.322†
High versus not-high‡	4.6 (2.3 to 9.3)	4.9 (2.4–9.7)	0.546†
Low versus not-low§	0.23 (0.10 to 0.52)	0.16 (0.06–0.40)	0.280†

*χ^2.
†Two-sample z-test.
‡Not-high=medium and low risk predictions.
§Not-low=high and medium risk predictions.
AUROC, area under the receiver operating curve; EWS, early warning system.

physicians and the RI, along with the study outcomes. The highest rate of clinical deterioration was for patients identified as high-risk by both physicians and the RI. Each identified at-risk patients that the other did not.

DISCUSSION

This study is the first to compare the predictive ability of physician judgement and an automated EWS for clinical deterioration of patients in a general internal medicine ward. We found that the physicians and the EWS had similar predictive ability and that a joint model combining physician and EWS predictions outperformed either independently. We found that the EWS and physicians identified partially overlapping sets of at-risk patients, suggesting that they were using different cues or different decision rules for their predictions. We found similar patterns when using AUROC and ORs to measure predictive ability, with each maintaining predictive ability in the joint model.

The rate of clinical deterioration (1.9%) observed in our study was similar to that in previous studies in similar clinical environments.[18 19] The AUROC for intern physicians using the probability scale in our study (0.70) was similar to that observed with physicians on general internal medicine teaching services using the single-question, 7-point Likert scale patient acuity rating.[18 19] The AUROC of 0.73 observed here for the RI prediction of clinical deterioration at 24 hours was lower than the AUROC of 0.93 reported for RI predictions of 24 hours mortality in a study that also included ICU patients.[13] As these are clinically different endpoints and populations, the meaning of this difference is unclear. Other studies of EWS predictive ability have also found a higher AUROC, but have used different clinical outcomes or settings and are not directly comparable to our results.[14–16]

Although the EWS and physicians had similar predictive ability, they identified different, but overlapping sets of at-risk patients. When we combined predictions from physicians and the EWS in a joint model, it outperformed either when used alone. This suggests that physicians and the EWS use different cues or decision rules when deriving their predictions and raises the possibility that combining them may improve predictive ability. A direct test of this hypothesis will require considering how best to provide physicians with EWS results or incorporate the additional cues used by physicians into EWSs. Such integration is an area for ongoing and future research, noting that prior efforts to integrate artificial intelligence

Table 4 A joint model combining physician and EWS predictions outperforms either alone

	Joint Model	Physicians	EWS
AUROC (CI)	0.78 (0.70–0.85)	0.70 (0.62 to 0.79)*	0.73 (0.66–0.81)*
P value		P=0.016†	P=0.050†
Goodness-of-fit testing		P<0.001‡	P=0.006‡
Adjusted ORs (CI)			
High versus low		4.8 (1.8 to 12.7)	7.4 (2.52 to 21.9)

*Independent AUROC.
†χ^2.
‡Likelihood ratio test.
AUROC, area under the receiver operating curve; EWS, early warning system.

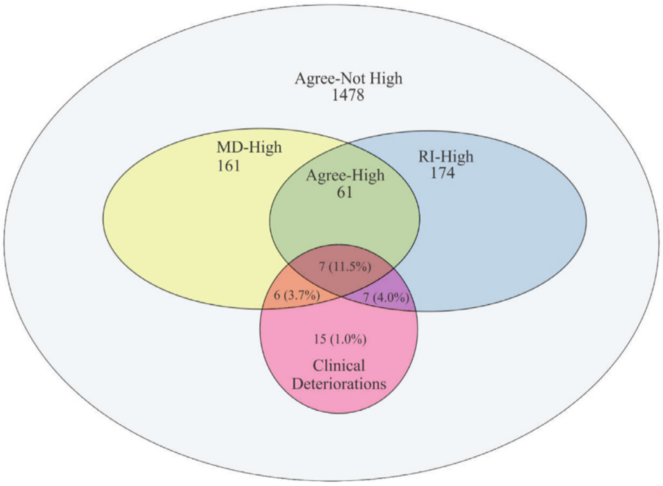

Figure 1 Distribution of high versus not-high risk assessments by physicians and an automated early warning system with rates of 24 hours clinical deterioration.

systems into medical decision making have largely failed to achieve widespread adoption.[25–27]

LIMITATIONS

A primary strength of the present study is its realism. It was conducted under normal clinical conditions for the physicians and with an EWS that had been deployed and optimised for the clinical environment being studied. That realism induced several possible limitations. One was that physicians' judgments might have been influenced by the RI predictions. We believe this not to have been the case. Although, in theory, the physicians had access to the RI through the EHR at the time of this study, the RI was not integrated into clinical workflow and had not been advertised to the clinical staff. Informal polling of physicians on the teaching services before and during the study indicated that the medical teams did not consult the RI. If they had, it would have had an indeterminate impact on physician judgments. As the RI was completely automated, it could not have been affected by physicians' judgments.

A second possible limitation is that the clinical outcome may not be independent of the physician predictions. Those predictions were made by the physicians treating the patients, who would be expected to take measures to reduce the chance of clinical deterioration in patients whom they saw as high risk. To the extent that those measures were successful, it might have reduced the predictive ability of the physician unless already factored into their predictions. Assuming that the RI predictions were unknown to the medical and nursing staff, they would not have influenced clinical deterioration.

This study is, we believe, the largest to evaluate physician predictions of clinical deterioration on general internal medicine floors. Nonetheless, the overall number of clinical deteriorations was still small, leading to large CIs for our estimates.

In addition to objective data, the RI uses EHR-documented nursing assessments in its algorithm and its performance will have depended on the quality of this documentation. While the RI was optimised by the vendor for these clinical settings, it is possible that nursing documentation may have changed over the course of the study which may have degraded the RI's performance.

Finally, though the RI has a raw continuous score, it was not available through the EHR interface. As a result, we used its graphic output, in a way meant to capture how it might be used in clinical care. We chose to interpret the graphic output in a manner that maximised the RI's predictive ability. Using the RI's continuous score or other translations of the graphic output (eg, how risk flags were used) might have improved or degraded the RI's predictive ability. In order to compare the two sets of predictions, we transformed physician judgments to match the distribution of RI predictions, which might have underestimated their relative predictive ability.

CONCLUSION

We offer the first large-scale prospective comparison of predictions made by physicians and an automated EWS in a non-ICU clinical setting for adult medicine patients. We found that physicians and the EWS had similar ability to predict clinical deterioration. A model combining the two sets of predictions outperformed either used alone. Physicians and the EWS identified different sets of at-risk patients, apparently using different predictive cues or decision rules. Further research is needed to understand the differences between these predictions, for combining them into a joint risk prediction model and for using them in clinical practice.

Author affiliations
[1]Division of General Internal Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
[2]Engineering and Public Policy, Carnegie Mellon University College of Engineering, Pittsburgh, Pennsylvania, USA
[3]Division of Pulmonary & Critical Care Medicine, University of Michigan Department of Internal Medicine, Ann Arbor, Michigan, USA
[4]Department of Radiology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
[5]Department of Critical Care Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA

Contributors All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing or revision of the manuscript. Conception and design of study: AD, BF, JH, DM. Acquisition of data: JA, JG, JH, AK, EY. Analysis and/or interpretation of data: JA, AD, JH. Drafting the manuscript: JA. Revising the manuscript critically for important intellectual content and final manuscript approval: JA, AD, BF, JG, JH, AK, DM, EY.

Funding JH was supported by the National Institutes of Health through Grant Number KL2 TR001856, https://www.nih.gov/. BF and AD were partially supported by the Riksbankens Jubileumsfond, https://www.rj.se/en. JA was funded by the Ruth L. Kirschstein National Research Service Award/Health Resources and Services Administration T-32, grant number T32HP22240, https://bhw.hrsa.gov/grants/medicine.

Disclaimer The funders had no role in design or conduct of the study; collection, management, analysis or interpretation of the data; preparation, review or approval of the manuscript or decision to submit the manuscript for publication.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Reasonable requests for study data and analysis code should be directed to the corresponding author, Jonathan Arnold <arnoldjd@pitt.edu>. Study data includes deidentified clinician clinical judgments and extracted Rothman Index color category and risk flags. Analysis code includes Stata routines. Execution of a data sharing agreement may be required.

ORCID iD

Jonathan Arnold http://orcid.org/0000-0003-1185-555X

REFERENCES

1. Kipnis P, Turk BJ, Wulf DA, *et al*. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *J Biomed Inform* 2016;64:10–19.
2. Rusin CG, Acosta SI, Shekerdemian LS, *et al*. Prediction of imminent, severe deterioration of children with parallel circulations using real-time processing of physiologic data. *J Thorac Cardiovasc Surg* 2016;152:171–7.
3. Schwartz SM. Can an automated early warning system derived from continuous physiologic monitoring prevent disaster? *J Thorac Cardiovasc Surg* 2016;152:3–4.
4. Morgan RJ, Williams F, Wright MM. An early warning scoring system for detecting developing critical illness. *Clin Intensive Care* 1997;8.
5. McQuillan P, Pilkington S, Allan A, *et al*. Confidential inquiry into quality of care before admission to intensive care. *BMJ* 1998;316:1853–8.
6. Rothschild JM, Gandara E, Woolf S, *et al*. Single-parameter early warning criteria to predict life-threatening adverse events. *J Patient Saf* 2010;6:97–101.
7. Prytherch DR, Smith GB, Schmidt PE, *et al*. ViEWS—Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation* 2010;81:932–7.
8. Kellett J, Kim A. Validation of an abbreviated Vitalpac early warning score (views) in 75,419 consecutive admissions to a Canadian regional hospital. *Resuscitation* 2012;83:297–302.
9. Churpek MM, Yuen TC, Park SY, *et al*. Derivation of a cardiac arrest prediction model using ward vital signs*. *Crit Care Med* 2012;40:2102–8.
10. Churpek MM, Yuen TC, Huber MT, *et al*. Predicting cardiac arrest on the wards: a nested case-control study. *Chest* 2012;141:1170–6.
11. Smith GB, Prytherch DR, Meredith P, *et al*. 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;84:465–70.
12. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
13. Rothman MJ, Rothman SI, Beals J. Development and validation of a continuous measure of patient condition using the electronic medical record. *J Biomed Inform* 2013;46:837–48.
14. Smith MEB, Chiovaro JC, O'Neil M, *et al*. *Early warning system scores: a systematic review*. Washington (DC): Department of Veterans Affairs, 2014.
15. Alam N, Hobbelink EL, van Tienhoven AJ, *et al*. The impact of the use of the Early Warning Score (EWS) on patient outcomes: a systematic review. *Resuscitation* 2014;85:587–94.
16. Smith MEB, Chiovaro JC, O'Neil M, *et al*. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. *Ann Am Thorac Soc* 2014;11:1454–65.
17. Parshuram CS, Dryden-Palmer K, Farrell C, *et al*. Effect of a pediatric early warning system on all-cause mortality in hospitalized pediatric patients: the epoch randomized clinical trial. *JAMA* 2018;319:1002–12.
18. Sinuff T, Adhikari NKJ, Cook DJ, *et al*. Mortality predictions in the intensive care unit: comparing physicians with scoring systems. *Crit Care Med* 2006;34:878–85.
19. Ratelle JT, Kelm DJ, Halvorsen AJ, *et al*. Predicting and communicating risk of clinical deterioration: an observational cohort study of internal medicine residents. *J Gen Intern Med* 2015;30:448–53.
20. Edelson DP, Retzer E, Weidman EK, *et al*. Patient acuity rating: quantifying clinical judgment regarding inpatient stability. *J Hosp Med* 2011;6:475–9.
21. Knaus WA, Harrell FE, Lynn J, *et al*. The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. study to understand prognoses and preferences for outcomes and risks of treatments. *Ann Intern Med* 1995;122:191–203.
22. O'Hagan A, Buck CE, Daneshkhah A, *et al*. *Uncertain judgements: eliciting expert probabilities*. Chichester: Wiley, 2006.
23. Morgan MG. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proc Natl Acad Sci U S A* 2014;111:7176–84.
24. 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.
25. Alder H, Michel BA, Marx C, *et al*. Computer-based diagnostic expert systems in rheumatology: where do we stand in 2014? *Int J Rheumatol* 2014;2014:672714.
26. Lee J-G, Jun S, Cho Y-W, *et al*. Deep learning in medical imaging: general overview. *Korean J Radiol* 2017;18:570–84.
27. Miller RA. A history of the INTERNIST-1 and Quick Medical Reference (QMR) computer-assisted diagnosis projects, with lessons learned. *Yearb Med Inform* 2010;19:121–36.