



Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology

Stephen Gerry,¹ Timothy Bonnici,² Jacqueline Birks,^{1,3} Shona Kirtley,¹ Pradeep S Virdee,¹ Peter J Watkinson,⁴ Gary S Collins^{1,3}

¹Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK

²Critical Care Division, University College London Hospitals NHS Trust, London, UK

³Oxford University Hospitals NHS Foundation Trust, Oxford, UK

⁴Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK

Correspondence to: S Gerry
stephen.gerry@csm.ox.ac.uk
(ORCID 0000-0003-4654-7311)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;369:m1501
<http://dx.doi.org/10.1136/bmj.m1501>

Accepted: 25 March 2020

ABSTRACT OBJECTIVE

To provide an overview and critical appraisal of early warning scores for adult hospital patients.

DESIGN

Systematic review.

DATA SOURCES

Medline, CINAHL, PsycInfo, and Embase until June 2019.

ELIGIBILITY CRITERIA FOR STUDY SELECTION

Studies describing the development or external validation of an early warning score for adult hospital inpatients.

RESULTS

13 171 references were screened and 95 articles were included in the review. 11 studies were development only, 23 were development and external validation, and 61 were external validation only. Most early warning scores were developed for use in the United States (n=13/34, 38%) and the United Kingdom (n=10/34, 29%). Death was the most frequent prediction outcome for development studies (n=10/23, 44%) and validation studies (n=66/84, 79%), with different time horizons (the most frequent was 24 hours). The most common predictors were respiratory rate (n=30/34, 88%), heart rate (n=28/34, 83%), oxygen saturation, temperature, and systolic blood pressure (all n=24/34, 71%). Age (n=13/34, 38%) and sex (n=3/34, 9%) were less frequently included. Key details of the analysis populations were often not reported in development studies (n=12/29, 41%) or validation studies (n=33/84, 39%). Small sample sizes and insufficient numbers of event patients were common in model development and external validation studies. Missing data were often discarded, with just

one study using multiple imputation. Only nine of the early warning scores that were developed were presented in sufficient detail to allow individualised risk prediction. Internal validation was carried out in 19 studies, but recommended approaches such as bootstrapping or cross validation were rarely used (n=4/19, 22%). Model performance was frequently assessed using discrimination (development n=18/22, 82%; validation n=69/84, 82%), while calibration was seldom assessed (validation n=13/84, 15%). All included studies were rated at high risk of bias.

CONCLUSIONS

Early warning scores are widely used prediction models that are often mandated in daily clinical practice to identify early clinical deterioration in hospital patients. However, many early warning scores in clinical use were found to have methodological weaknesses. Early warning scores might not perform as well as expected and therefore they could have a detrimental effect on patient care. Future work should focus on following recommended approaches for developing and evaluating early warning scores, and investigating the impact and safety of using these scores in clinical practice.

SYSTEMATIC REVIEW REGISTRATION

PROSPERO CRD42017053324.

Introduction

Research towards the end of the 20th century showed the incidence of adverse events and unnecessary deaths in hospital patients.¹⁻⁴ Early warning scores (EWSs) were proposed as a potential solution.⁵ These tools are clinical prediction models that generally use measured vital signs to monitor patients' health during their hospital stay. The models identify the likelihood of patients deteriorating, which is often defined as death or admission to the intensive care unit. When a patient shows signs of deterioration, the EWS triggers a warning so that care can be escalated. Historically EWSs were implemented on paper based observation charts, but now they are increasingly becoming part of electronic health record systems.

EWSs based on vital signs are widely used every day in hospitals to identify patients who are clinically deteriorating. These measures are routinely used in several countries, including the Netherlands, the United States, Australia, and the Republic of Ireland.⁶⁻⁹ In hospitals in the United Kingdom EWS use is mandated as a standard of care by the National Institute for Health and Care Excellence.¹⁰ Because hospital inpatients are usually assessed every few hours by using an EWS, these scores are used hundreds of millions of times each year.¹¹ Requests have also been

WHAT IS ALREADY KNOWN ON THIS TOPIC

Early warning scores are widely used in hospitals to identify clinical deterioration in patients, for example the modified early warning score and the national early warning score

Early warning scores are commonly implemented by using electronic systems
A systematic overview of studies developing and externally validating these systems has been lacking

WHAT THIS STUDY ADDS

An abundance of articles describe the development or validation of early warning scores

Poor methods and inadequate reporting were found in most studies, and all studies were at risk of bias

Methodological problems could result in scoring systems that perform poorly in clinical practice, which might have detrimental effects on patient care

made to increase EWS use across ambulance services, primary care, and community care homes.¹²⁻¹⁶

Articles that describe the development of clinical prediction models abound in many areas of medicine.¹⁷⁻¹⁸ Systematic reviews have shown that the methods used in these papers are often poor.¹⁹⁻²² Although many published prediction models are not put into practice, EWSs are used widely, probably more than any other type of clinical prediction model. Despite extensive development and increasing uptake, comprehensive reviews of EWS articles in the past decade have been lacking. Systematic reviews are needed that assess the methodological and reporting quality of papers describing the development and validation of EWSs. External validation studies, which are vital for assessing the generalisability of EWSs, need to be systematically evaluated. Existing systematic reviews of EWSs have mostly concentrated on predictive performance, but have hinted at methodological flaws.²³⁻²⁴

Hospital patients will probably have their vital signs and other parameters measured several times during their hospital stay, therefore the available datasets might include multiple measurements (or observation sets) for each patient. The most appropriate way to analyse such data is not clear, which increases the complexity of EWS research in comparison to other areas of clinical prediction modelling.²⁵⁻²⁶ Debate also exists about the best choice of outcome measure and time horizon; for example, death or admission to intensive care within a specific time period (eg, 24 hours) or the whole hospital stay.²⁷ Different approaches to these problems might give different results when developing and validating EWSs, and could lead to models being used that do not work.

The great potential for EWSs to assist in clinical decision making might be thwarted by poor methods and inadequate reporting. The widespread use of EWSs means poorly developed and reported EWSs could have a highly detrimental effect on patient care. We carried out a systematic review to assess the methods and reporting of studies that developed or externally validated EWSs for general adult patients.

Methods

Details of the study design and rationale have been previously published.¹¹ In summary, we identified articles that described the development or validation of EWSs. The Medline (Ovid), CINAHL (EBSCOHost), PsycInfo (Ovid), and Embase (Ovid) databases were searched from inception to 30 August 2017. An update search was conducted on 19 June 2019 to identify articles published since the date of the original search. Search strategies were developed by an information specialist (SK) for each database and are reported in the supplementary appendix. Search terms included relevant controlled vocabulary terms (eg, MeSH, EMTREE) and free text variations for early warning or track and trigger scores or systems (including common acronyms), physiological monitoring or health status indicators, combined with development

and validation terms. We did not apply any date or language restrictions to the search. Additional articles were found by searching the references in papers identified by the search strategy, our own personal reference lists, and a Google Scholar search.

Eligibility criteria

We included any primary research articles that described the development or validation of one or more EWSs, defined as a score (with at least two predictors) used to identify general patients admitted to hospital who are at risk of clinical deterioration. External validation studies were only included if an article describing the development of that EWS was also available.

Articles were not eligible if the score was developed for use in a subset of patients with a specific disease or group of diseases, for use in children (<16 years old), or in pregnant women; when the score is intended to be used for outpatients or for patients in the intensive care unit; when no vital signs were included in the final model; or when the article was a review, letter, personal correspondence or abstract, or the article was published in a non-English language.

Study selection and data extraction

One author (SG) screened the titles and abstracts of all articles identified by the search string. Two reviewers (from SG, PSV, and JB) independently extracted data by using a standardised and piloted data extraction form. Conflicts were resolved by discussion between the two relevant reviewers. The form was administered by using the REDCap (research electronic data capture) electronic data capture tool.²⁸ The items for extraction were based on the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist,²⁹ supplemented by subject specific questions and methodological guidance. These items included study design characteristics, patient characteristics, sample size, outcomes, statistical analysis methods, and model performance methods.

Items extracted from studies describing the development of EWSs included the following (for an explanation of some of the technical terms, see box 1): study design (retrospective, prospective), details of population (eg, when and where data were collected, age, sex), method of development (eg, clinical consensus, statistical approach), predicted outcome and time horizon, number and type of predictors, sample size, number of events, missing data approach, modelling approach (eg, type of regression model, method used to select variables, handling of continuous variables, examination of interaction terms), model presentation (eg, reporting of model coefficients, intercept or baseline hazard, simplified model), method of internal validation (eg, split sample, bootstrapping, cross validation), and assessment of model performance (eg, discrimination, calibration). Items extracted from studies describing the external validation of EWSs included study design (retrospective, prospective), details of population (eg, when and where data were

collected, age, sex), predicted outcome and time horizon, sample size, number of events, missing data approach, and assessment of model performance (eg, discrimination, calibration). We define event patients as the number of patients recorded as having the outcome of interest (eg, dying or being admitted to the intensive care unit at any point during their hospital stay). Event observations refer to the number of observation sets that are within a defined period before the outcome occurs.

Assessment of bias

We assessed the risk of bias for each article by using PROBAST (prediction model risk of bias assessment tool), which was developed by the Cochrane Prognosis Methods Group.³⁰ PROBAST consists of 23 signalling questions within four domains (participant selection, predictors, outcome, and analysis). The articles were classified as low, high, or unclear risk of bias for each domain. A study was classified as having an overall low risk of bias only if it was at low risk of bias within each domain.

Evidence synthesis

We summarised the results by using descriptive statistics, graphical plots, and a narrative synthesis. We did not perform a quantitative synthesis of the models because this was not the main focus of the review, and the studies were too heterogeneous to combine.

Patient and public involvement

Patients and members of the public were involved in setting the research question and developing the study, through face-to-face meetings and revisions of the protocol. Patients and members of the public have read and revised the manuscript. There are no plans to disseminate the results of the research to patients or the public.

Results

The search strategy identified 13 171 unique articles, of which 12 794 were excluded based on title and

abstract screening. We screened 377 full texts, 93 of which met the eligibility criteria and were included in the review (fig 1). We identified two more articles by searching the article references, which were also included, giving 95 articles in total. Eleven articles described development of EWSs only, 23 described development and external validation, and 61 articles described external validation only. The articles were published between 2001 and 2019 in 51 journals. One journal, *Resuscitation*, published 21 of the articles. No other journal included more than four of the articles. Ninety three articles used a patient dataset (two used only clinical consensus), with most using data from the UK (n=28) or the US (n=25). The articles represented data from 22 countries across four continents.

Studies describing development of EWSs

Study design

Of the 34 articles describing the development of a new EWS, 29 were based on statistical methods; that is, they used some form of data driven approach to create the model. Three studies developed models based on clinical consensus, where a group of experts chose the variables and associated weights that would form the model. Two studies developed models by modifying an existing score, either through modifying the variable weights or through adding binary variables, to improve predictive performance (table 1), however the rationale for modifying an existing score was not reported.

Most of the 29 studies that were developed by statistical methods used data from retrospective cohorts (n=21, 72%), while 7 (24%) used data from prospectively collected cohort datasets. Data used to develop the models were collected between 2000 and 2017. Twelve of the 29 studies (41%) did not adequately describe their dataset, missing at least one of the following characteristics: average age, distribution of men and women, number of patients with and without the event, and number of observation sets with and without the event (a patient could contribute more than one set of observations).

Twenty three of the 29 studies (79%) used a prediction modelling approach (including regression modelling and machine learning methods). The remaining six studies used a variety of methods. Appendix table C gives further details.

Outcome measures and time horizons

We observed a variety of primary outcome measures in the 23 development studies that used a prediction modelling approach (supplementary table A). Nearly all studies used death, intensive care unit admission, cardiac arrest, or a composite of these. The most common primary outcome measures were death (n=10, 44%) and cardiac arrest (n=4, 17%). A wide variety of prediction time horizons were also used; the most frequent was 24 hours (n=8, 35%). Other common horizons were 12 hours (n=3, 13%), 30 days (n=3, 13%), or in-hospital (n=6, 26%). Figure 2 shows a breakdown of outcomes and their time horizons.

Box 1: Definitions of technical terms

- Apparent performance: evaluation of the model's predictive accuracy with the same data used to develop it
- Internal validation: evaluation of the model's predictive accuracy in the population in which the model is intended for use; the apparent performance is adjusted for the optimism resulting from overfitting
- External validation: evaluation of the model's predictive accuracy with data other than those used to develop it
- Discrimination: ability of the model to distinguish between patients who will and those who will not go on to develop the outcome of interest; typically measured using the C index
- Calibration: agreement between predicted risks and observed event rates
- Prediction horizon: timeframe in which the model is intended to predict the outcome of interest
- Individualised risk prediction: ability of the model to estimate probability of outcome occurring based on patient's characteristics
- Observation set: vital sign measurements of an individual patient at a particular point in time; typically consists of blood pressure, heart rate, respiratory rate, temperature, and oxygen saturation

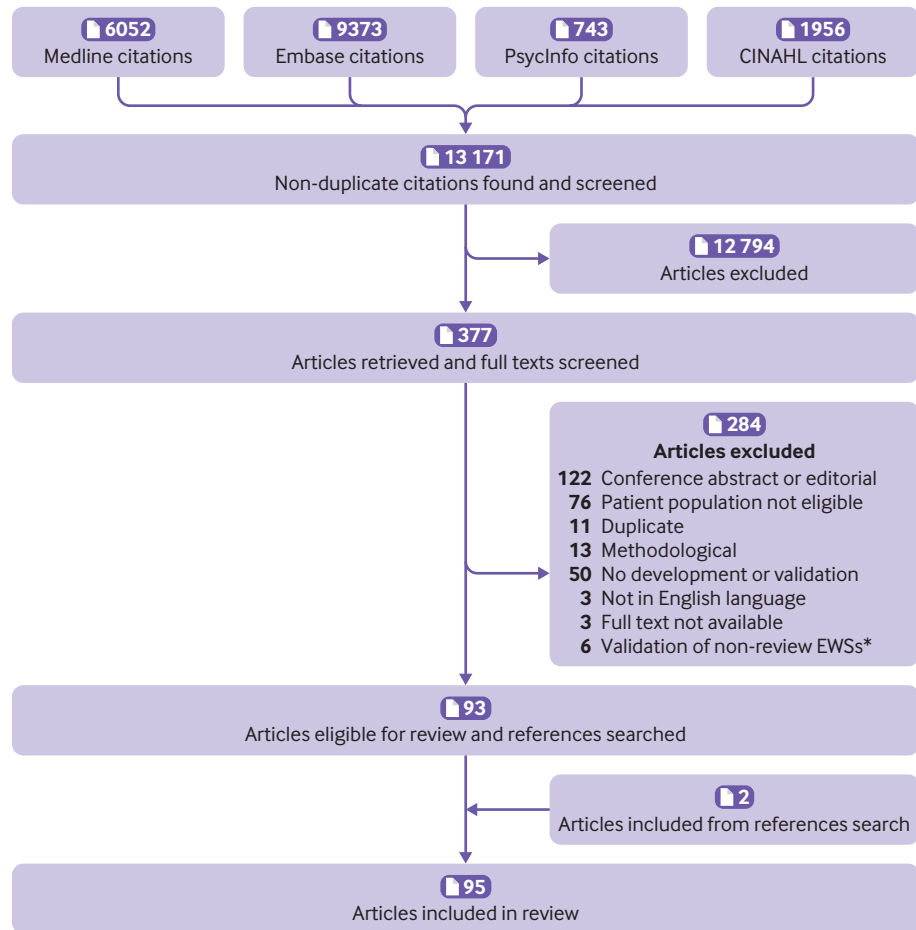


Fig 1 | Flow diagram of article selection. *Validation of non-review EWSs (early warning scores) refers to external studies, which are excluded because the corresponding development paper was ineligible or because no development paper has been published

Predictors

Twenty one of the 23 (91%) development studies that used a prediction modelling approach reported how many candidate predictors were considered for inclusion in the EWS, together reporting a median of 12 (range 4-45) predictors (supplementary table B). The median number of predictors included in the final model was seven (range 3-35). The most common approach for selecting variables for inclusion was backwards elimination (n=9/23, 39%). Six of the 23 models (26%) included all candidate variables, and three studies (13%) carried out univariable screening to reduce the initial number of candidate variables.

The most frequently included predictor in the 34 development studies was respiratory rate (n=30, 88%), followed by heart rate (n=28, 83%), oxygen saturation (n=24, 71%), temperature (n=24, 71%), systolic blood pressure (n=24, 71%), and level of consciousness (n=19, 56%). Thirteen models included age (38%) and three models included sex (9%).

Sample size (for 29 studies developed using statistical method)

The sample size in EWS studies can be complicated because there might be multiple observation sets for

each patient or hospital admission. It was not always clear whether the reported sample size referred to the number of patients, hospital admissions, or observation sets (n=3; supplementary table C). The median patient or hospital admission sample size was 10 712 (range 242-649 418). Eleven of 29 articles (38%) used multiple observation sets for each patient, 15 (52%) used one observation set for each patient, and three (10%) were unclear. Of the 15 studies that used only one observation set for each patient, the first recorded observation was generally used (n=9, 60%).

The median number of events at the patient level was 396 (range 18-19 153) and at the observation set level was 284 (18-15 452). One article did not report the number of events at the patient level, and eight articles did not report the number of events at the observation set level. This difference in denominator explains how the median number of events can be greater at the patient level than at the observation set level.

The events per variable is a key marker of sample size adequacy in prediction modelling studies, and is defined as the number of events divided by the number of candidate predictor variables used. Twenty articles used a prediction modelling approach and provided sufficient information to calculate the patient level

Table 1 | Study design characteristics of 34 articles describing development of early warning score

Reference	EWS	Type of development	Type of data	Country	Years of data	Mean or median age	Male (%)
Albert 2011 ³¹	—	Based on clinical consensus	NA	US	NA	NA	NA
Alvarez 2013 ³²	—	Using statistical methods (based on data)	Retrospective cohort/ database	US	2009-10	51	56
Badriyah 2014 ³³	DTEWS	Using statistical methods (based on data)	Retrospective cohort/ database	UK	2006-08	68	47
Bleyer 2011 ³⁴	Trio of critical vital signs	Using statistical methods (based on data)	Retrospective cohort/ database	US	2009	57	51
Churpek 2012 ⁸	CART	Using statistical methods (based on data)	Retrospective cohort/ database	US	2008-11	54	43
Churpek 2014 ³⁵	—	Using statistical methods (based on data)	Retrospective cohort/ database	US	2008-11	54	43
Churpek 2014 ³⁶	eCART	Using statistical methods (based on data)	Retrospective cohort/ database	US	2008-13	60	40
Churpek 2016 ³⁷	—	Using statistical methods (based on data)	Retrospective cohort/ database	US	2008-13	60	40
Cuthbertson 2010 ³⁸	—	Using statistical methods (based on data)	Prospective cohort	UK	2005	65	51
Douw 2016 ⁹	DENWIS	Modification of existing score	NA	Netherlands	NA	NA	NA
Duckitt 2007 ³⁹	Worthing PSS	Using statistical methods (based on data)	Prospective cohort	UK	2003-05	73	52
Dziadzko 2018 ⁴⁰	APPROVE	Using statistical methods (based on data)	Retrospective cohort/ database	US	2013	58	41
Escobar 2012 ⁴¹	EMR based model	Using statistical methods (based on data)	Retrospective cohort/ database	US	2006-09	65	45
Faisal 2018 ⁴²	CARM	Using statistical methods (based on data)	Prospective cohort	UK	2014-15	67	50
Ghosh 2018 ⁴³	EDI	Using statistical methods (based on data)	Retrospective cohort/ database	US	2012-13	59	Missing
Goldhill 2004 ⁴⁴	—	Using statistical methods (based on data)	Prospective cohort	UK	2002	61	Missing
Harrison 2006 ⁴⁵	GMEWS	Modification of existing score	NA	Australia	NA	NA	NA
Jones 2012 ⁴⁶	NEWS	Based on clinical consensus	NA	UK	NA	NA	NA
Kellett 2006 ⁴⁷	SCS	Using statistical methods (based on data)	Retrospective cohort/ database	Ireland	2000-04	62	52
Kellett 2008 ⁴⁸	HOTEL	Using statistical methods (based on data)	Retrospective cohort/ database	Ireland	2000-04	62	53
Kipnis 2016 ⁴⁹	AAM	Using statistical methods (based on data)	Retrospective cohort/ database	US	2010-13	65	46
Kirkland 2013 ⁵⁰	—	Using statistical methods (based on data)	Other	US	2008-09	72	62
Kwon 2018 ⁵¹	DEWS	Using statistical methods (based on data)	Retrospective cohort/ database	South Korea	2010-17	57	52
Kyriacos 2014 ⁵²	MEWS*	Based on clinical consensus	NA	South Africa	NA	NA	NA
Luis 2018 ⁵³	Short NEWS	Using statistical methods (based on data)	Retrospective cohort/ database	Portugal	2012	Missing	48
Moore 2017 ⁵⁴	UVA	Using statistical methods (based on data)	Retrospective cohort/ database	Gabon, Malawi, Sierra Leone, Tanzania, Uganda, and Zambia	2009-15	36	49
Nickel 2016 ⁵⁵	NEWS and D-dimer	Using statistical methods (based on data)	Retrospective cohort/ database	Denmark	2008-11	62	45
Perera 2011 ⁵⁶	MEWS plus biochemical	Using statistical methods (based on data)	Prospective cohort	Sri Lanka	2009	49	48
Prytherch 2010 ⁵⁷	ViEWS	Using statistical methods (based on data)	Retrospective cohort/ database	UK	2006-08	68	48
Redfern 2018 ⁵⁸	LDTEWS:NEWS	Using statistical methods (based on data)	Retrospective cohort/ database	UK	2011-16	73	49
Silke 2010 ⁵⁹	MARS	Using statistical methods (based on data)	Retrospective cohort/ database	Ireland	2002-07	50	48
Tarassenko 2011 ⁶⁰	CEWS	Using statistical methods (based on data)	Prospective cohort	UK and US	2004-08	60	57
Watkinson 2018 ⁶¹	mCEWS	Using statistical methods (based on data)	Retrospective cohort/ database	UK	2014-15	63	51
Wheeler 2013 ⁶²	TOTAL	Using statistical methods (based on data)	Prospective cohort	Malawi	2012	40	51

AAM=advanced alert monitor; APPROVE=accurate prediction of prolonged ventilation; CARM=computer aided risk of mortality; CART=cardiac arrest risk triage; CEWS=centile early warning score; DENWIS=Dutch early nurse worry indicator score; DEWS=deep learning-based early warning system; DTEWS=decision tree early warning score; eCART=electronic cardiac arrest risk triage; EDI=early deterioration indicator; EMR=electronic medical record; GMEWS=global modified early warning score; HOTEL=hypotension, oxygen saturation, temperature, ECG [electrocardiogram] abnormality, loss of independence; LDTEWS=laboratory decision tree early warning score; MARS=medical admissions risk system; MEWS=modified early warning score; mCEWS>manual centile early warning score; NA=not available; NEWS=national early warning score; PSS=physiological scoring system; SCS=simple clinical score; TOTAL=tachypnoea, oxygen saturation, temperature, alert and loss of independence; UVA=universal vital assessment; ViEWS=VitalPAC early warning score.

*Not the same as original MEWS.

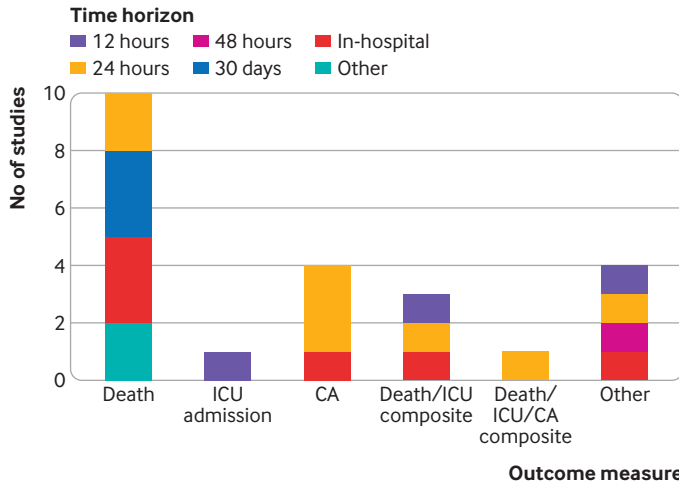


Fig 2 | Summary of development outcomes and time horizons appearing in 23 studies that used regression modelling approach to develop early warning score. CA=cardiac arrest; ICU=intensive care unit

events per variable, with a median of 52 and a range from 1 to 1288. Fifteen studies provided enough information to calculate the observation set level events per variable, with a median of 17 and a range from 1 to 2693.

Statistical methods

Most of the articles that used statistical methods to develop an EWS mentioned missing data (n=25/29, 86%). Supplementary table C lists the methods for dealing with missing data; complete case analysis was the most common approach (n=10/25, 40%). None of the included studies used multiple imputation to handle missing data in the development of an EWS. Four articles mentioned missing data, but did not clearly state which method was used to handle them.

Most of the 23 models developed using a prediction modelling approach used logistic regression (n=15, 65%). Other methods included machine learning (n=4, 17%), Cox proportional hazards regression, multinomial logistic regression, discrete time logistic regression, and naïve Bayes classification combined with logistic regression (all n=1, 4%). The four machine learning studies used decision trees (n=2), artificial neural networks (n=1), or random forests (n=1).

For the handling of continuous predictors and use of interaction terms, all of the 23 prediction models included at least one continuous variable (supplementary table C). The most common approach for handling these variables was to categorise the variable before analysis (n=7, 30%). Other methods included splines (n=6, 26%), linear relations (n=4, 17%), and fractional polynomials (n=2, 9%). Four studies used other methods.

Model presentation

Nine of the 23 (39%) models developed by using a prediction modelling approach reported the complete regression formula, with all coefficients and either the intercept or baseline hazard (supplementary table E). Of

the remaining models, seven (30%) did not report any coefficients, and seven (30%) reported the predictor coefficients but not the intercept or baseline hazard.

Thirteen of the studies (57%) reported enough information for us to calculate individualised risk predictions. Two articles (9%) reported the construction of risk groups. Ten articles (44%) created a simplified model, although only five described how this was done. These simplified models typically reduced the model coefficients to a points based scoring system, with no method of calculating predicted risks.

Apparent predictive performance

Twenty two studies assessed performance by using the same data that were used in the development of the model, thus assessing apparent performance (supplementary table F). Eighteen of these studies (82%) assessed discrimination with the C index, with values ranging from 0.69 to 0.96. Calibration was assessed for eight models (36%); seven used the Hosmer-Lemeshow goodness-of-fit test, and one used a calibration plot. Other performance metrics reported included sensitivity and specificity (n=8, 36%), and positive or negative predictive values (n=4, 18%). Eight studies presented receiver operating characteristic curves.

Internal validation

Supplementary table G shows reporting of internal validation in the 34 development studies. Nineteen models were internally validated. Note that two additional studies included an external validation of their new EWSSs, but not an internal validation. Most studies split their data into development and validation data (n=13/19, 68%). Two articles used bootstrapping, and two used cross validation (both 11%). The remaining two assessed performance by using the derivation data combined with additional data (11%). All studies that assessed discrimination used the C index. Calibration was assessed in four studies, one using a calibration plot and three using the Hosmer-Lemeshow test. Sensitivity and specificity were reported in six studies and eight studies produced receiver operating characteristic curves.

Studies describing external validation of EWS

We included 84 articles that externally validated an EWS (table 2). Twenty three of these also described the development of an EWS. Five developed an EWS and externally validated it in an external dataset, and 18 developed an EWS and externally validated a different EWS by using the development dataset.

Models validated

Twenty two models were validated across the 84 studies (fig 3). The modified early warning score⁶¹ was most frequently validated (n=43), followed by the national early warning score (NEWS)⁴⁴ (n=40). The VitalPAC early warning score,⁶² on which NEWS was based, was validated 10 times, and the original EWS⁵ was validated eight times.

Table 2 | Design characteristics of 84 studies describing external validation of early warning score

Reference	Type of dataset	Country	Year	EWS validated	Mean age	Male (%)
Abbott 2016 ⁶³	Prospective cohort	UK	2013	NEWS	63	48
Abbott 2015 ⁶⁴	Prospective cohort	UK	2013	NEWS	61	46
Alvarez 2013 ³²	Prospective cohort	US	2009-10	MEWS	51	54
Atmaca 2018 ⁶⁵	Prospective cohort	Turkey	2014	NEWS	57	55
Badriyah 2014 ³³	Retrospective cohort/database	UK	2006-08	NEWS	68	47
Bartkowiak 2019 ⁶⁶	Retrospective cohort/database	US	2008-16	eCART, NEWS, MEWS	54	43
Beane 2018 ⁶⁷	Retrospective cohort/database	Sri Lanka	2015	MEWS, NEWS, CART, VIEWS	43	41
Bleyer 2011 ³⁴	Retrospective cohort/database	US	2008-09	NEWS, VIEWS	57	51
Brabrand 2017 ⁶⁸	Retrospective cohort/database	Denmark	2012	NEWS, Worthing, Groarke, Goodacre	67	50
Brabrand 2018 ⁶⁹	Retrospective cohort/database	Denmark	Missing	NEWS	74	49
Cei 2009 ⁷⁰	Prospective cohort	Italy	2005-06	MEWS	79	44
Churpek 2017 ⁷¹	Retrospective cohort/database	US	2008-16	eCART, NEWS, MEWS	57	46
Churpek 2017 ⁷²	Retrospective cohort/database	US	2008-16	NEWS, MEWS	57	48
Churpek 2013 ⁷³	Retrospective cohort/database	US	2008-11	CEWS, MEWS, VIEWS, CART	55	44
Churpek 2014 ³⁶	Retrospective cohort/database	US	2008-13	MEWS	60	40
Churpek 2012 ⁷⁴	Other	US	2008-11	MEWS	59	52
Churpek 2012 ⁸⁵	Retrospective cohort/database	US	2008-11	CART, MEWS	54	43
Churpek 2014 ³⁵	Retrospective cohort/database	US	2008-11	VIEWS	54	43
Cooksley 2012 ⁷⁵	Retrospective cohort/database	UK	2009-11	NEWS, MEWS	63	51
Cuthbertson 2010 ³⁸	Prospective cohort	UK	2005	EWS, MEWS	65	51
De Meester 2013 ⁷⁶	Prospective cohort	Belgium	2009-10	MEWS	59	60
DeVoe 2016 ⁷⁷	Retrospective cohort/database	US	2007-13	MEWS	75	61
Douw 2017 ⁷⁸	Retrospective cohort/database	Netherlands	2013-14	DENWIS	60	47
Duckitt 2007 ³⁹	Prospective cohort	UK	2003-05	EWS	73	52
Dziadzko 2018 ⁴⁰	Retrospective cohort/database	US	2017	APPROVE, MEWS, NEWS	56	33
Eccles 2014 ⁷⁹	Retrospective cohort/database	UK	2012	NEWS	70	50
Escobar 2012 ⁴¹	Retrospective cohort/database	US	2006-09	MEWS	65	45
Fairclough 2009 ⁸⁰	Prospective cohort	UK	2004-06	MEWS	73	43
Faisal 2018 ⁴²	Retrospective cohort/database	UK	2014-15	CARM	68	48
Finlay 2014 ⁸¹	Retrospective cohort/database	US	2009-10	MEWS	65	NR
Forster 2018 ⁸²	Retrospective cohort/database	UK	2015-17	NEWS	63	47
Garcea 2006 ⁸³	Retrospective cohort/database	UK	2002-06	EWS	57	NR
Gardner 2006 ⁸⁴	Prospective cohort	UK	2003	MEWS	59	50
Ghanem 2011 ⁸⁵	Prospective cohort	Israel	2008-09	MEWS	75	52
Ghosh 2018 ⁴³	Retrospective cohort/database	US	2012-13	MEWS, NEWS	59	NR
Green 2018 ⁸⁶	Retrospective cohort/database	US	2008-13	MEWS, NEWS, eCART	62	41
Harrison 2006 ⁴⁵	Retrospective cohort/database	Australia	2000	MEWS	NR	NR
Hodgson 2017 ⁸⁷	Retrospective cohort/database	UK	2012-14	NEWS	74	NR
Hydes 2018 ⁸⁸	Retrospective cohort/database	UK	2010-14	NEWS, EWS, MEWS, MEWS+age, Worthing	57	61
Jo 2016 ⁸⁹	Retrospective cohort/database	South Korea	2013-14	NEWS	70	63
Kellett 2012 ⁹⁰	Retrospective cohort/database	Canada	2005-11	VIEWS	63	49
Kellett 2016 ⁹¹	Prospective cohort	Canada	2005-16	VIEWS	65	49
Kim 2018 ⁹²	Retrospective cohort/database	South Korea	2014-15	NEWS	65	70
Kim 2017 ⁹³	Retrospective cohort/database	South Korea	2008-15	MEWS	61	65
Kipnis 2016 ⁴⁹	Retrospective cohort/database	US	2010-13	eCART, NEWS	65	46
Kovacs 2016 ⁹⁴	Retrospective cohort/database	UK	2011-13	NEWS	57	47
Kruisselbrink 2016 ⁹⁵	Prospective cohort	Uganda	2013	MEWS	43	54
Kwon 2018 ⁵¹	Retrospective cohort/database	South Korea	2017	MEWS	58	50
LeLagadec 2020 ⁹⁶	Retrospective case-control	Australia	2014-17	NEWS	73	53
Lee 2018 ⁹⁷	Retrospective cohort/database	South Korea	2013-14	NEWS	62	58
Liljehult 2016 ⁹⁸	Retrospective cohort/database	Denmark	2012	NEWS	72	50
Luis 2018 ⁵³	Retrospective cohort/database	Portugal	2012	NEWS	NR	48
Moore 2017 ⁵⁴	Retrospective cohort/database	Gabon, Malawi, Sierra Leone, Tanzania, Uganda, and Zambia	2009-15	MEWS	36	49
Mulligan 2010 ⁹⁹	Prospective cohort	UK	2007	EWS	48	85
Öhman 2018 ¹⁰⁰	Retrospective cohort/database	Denmark	2008-10	MARS	65	50
Opio 2013 ¹⁰¹	Retrospective cohort/database	Uganda	2012	VIEWS	45	42
Opio 2013 ¹⁰²	Prospective cohort	Ireland	2011-13	TOTAL	64	53
Pedersen 2018 ¹⁰³	Retrospective cohort/database	Denmark	2014	NEWS	74	42
Perera 2011 ⁵⁶	Prospective cohort	Sri Lanka	2009	MEWS	49	48
Pimentel 2019 ¹⁰⁴	Retrospective cohort/database	UK	2012-16	NEWS	68	48
Plate 2018 ¹⁰⁵	Retrospective cohort/database	Netherlands	2014-16	VIEWS	61	65
Prytherch 2010 ⁵⁷	Retrospective cohort/database	UK	2006-08	EWS, Goldhill, MEWS, MEWS+age, Worthing	68	48
Redfern 2018 ¹⁰⁶	Retrospective cohort/database	UK	2010-16	NEWS	63	47
Redfern 2018 ⁵⁸	Retrospective cohort/database	UK	2016	LDTEWS:NEWS, NEWS	73	50

(Continued)

Table 2 | Continued

Reference	Type of dataset	Country	Year	EWS validated	Mean age	Male (%)
Roberts 2017 ¹⁰⁷	Retrospective cohort/database	Sweden	2014-15	NEWS	NR	60
Romero 2017 ¹⁰⁸	Retrospective cohort/database	US	2011	GMEWS, Kirkland, MEWS, NEWS, ViEWS, Worthing	59	49
Romero 2014 ¹⁰⁹	Retrospective cohort/database	US	2011	MEWS, GMEWS, Worthing, ViEWS, NEWS	59	49
Rylance 2009 ¹¹⁰	Prospective cohort	Tanzania	2005	MEWS	NR	34
Silke 2010 ⁵⁹	Retrospective cohort/database	Ireland	2000-04	MARS	59	48
Smith 2008 ¹¹¹	Retrospective cohort/database	UK	2006	EWS, Goldhill, MEWS, MEWS+age, Worthing	68	48
Smith 2013 ¹¹²	Retrospective cohort/database	UK	2006-08	NEWS, EWS, Goldhill, MEWS, MEWS+age, Worthing	68	47
Smith 2016 ¹¹³	Retrospective cohort/database	UK	2011-13	NEWS	62	48
Smith 2016 ¹¹⁴	Retrospective cohort/database	US	2014-15	NEWS	53	NR
Spagnoli 2017 ¹¹⁵	Prospective cohort	Italy	2013-15	NEWS	72	50
Stark 2015 ¹¹⁶	Retrospective cohort/database	US	2013-14	MEWS	62	65
Stræede 2014 ¹¹⁷	Retrospective cohort/database	Denmark	2008-09	SCS, HOTEL	62	52
Subbe 2001 ¹¹⁸	Retrospective cohort/database	UK	2000	MEWS, MEWS+age	63	45
Suppiah 2014 ¹¹⁹	Prospective cohort	UK	2010	MEWS	56	50
Tirkkonen 2014 ¹²⁰	Prospective cohort	Finland	2010	NEWS	65	53
Tirotta 2017 ¹²¹	Prospective cohort	Italy	2012	MEWS, TOTAL	73	50
Vaughn 2018 ¹²²	Retrospective cohort/database	US	2011-15	MEWS	54	NR
VonLilienfeld-Toal 2007 ¹²³	Retrospective cohort/database	Missing	2002-04	MEWS	40	51
Watkinson 2018 ⁶¹	Retrospective cohort/database	UK	2015-17	CART, CEWS, Goldhill, MEWS, MEWS+age, NEWS	68	49
Wheeler 2013 ⁶²	Prospective cohort	Malawi	2012	MEWS, HOTEL	40	51

APPROVE=accurate prediction of prolonged ventilation; CARM=computer aided risk of mortality; CART=cardiac arrest risk triage; CEWS=centile early warning score; DENWIS=Dutch early nurse worry indicator score; eCART=electronic cardiac arrest risk triage; EWS=early warning score; GMEWS=global modified early warning score; HOTEL=hypotension, oxygen saturation, temperature, ECG [electrocardiogram] abnormality, loss of independence; LDTEWS=laboratory decision tree early warning score; MARS=medical admissions risk system; MEWS=modified early warning score; NEWS=national early warning score; NR=not reported; SCS=simple clinical score; TOTAL=tachypnoea, oxygen saturation, temperature, alert and loss of independence; ViEWS=VitalPAC early warning score.

Study design

Most of the validation articles (n=58/84, 69%) used existing data to externally validate an EWS (table 2). Twenty five (30%) collected prospective data for external validation. The data used to validate the EWSs were all collected between 2000 and 2017. Thirty three of the 84 studies (39%) did not adequately describe their dataset, missing at least one of the following: average age, distribution of men and women, number of patients with or without the event, and number of observation sets with or without the event.

Outcome measures and horizon times

The models were validated against a range of outcomes (fig 4, supplementary table H). The most frequent was death, which was included in 66 of 84 articles (79%), followed by unanticipated admission to intensive care (n=22, 26%), and a composite of death and unanticipated admission to intensive care (n=17, 20%). A variety of prediction horizons were used. In-hospital (that is, the remainder of the hospital stay) was the most frequently used time point (n=58, 69%), followed by 24 hours (n=56, 67%). Figure 4 shows all outcome and time horizon combinations; in-hospital death was the most commonly validated endpoint (n=26, 31%).

Sample size

Supplementary table I shows reported information on the sample size used in each external validation. The number of patients and observation sets could not be identified in eight studies, and the number of

event patients and event observation sets could not be identified in 25 studies. For studies that did report these data, the median number of patients included in the validation articles was 2806 (range 43-649 418) and the median number of observation sets was 3160 (range 43-48 723 248). The median number of event patients was 126, ranging from 6 to 19 153.

Multiple observation sets were used for each patient in 23 of 84 articles (27%), while one observation set for each patient was used in 55 articles (66%). The remaining six studies did not clearly report whether multiple observation sets had been used. Most of the studies that used a single observation set for each patient used the first observation set (n=41/55, 75%).

Statistical methods

Sixty three of the 84 validation articles (75%) mentioned missing data (supplementary table J). The most common approach for dealing with missing data was complete case analysis (n=36, 57%), followed by using the last observation carried forward (n=8, 13%). For seven studies the method was unclear (11%). Two articles reported having no missing data (3%). One article used multiple imputation (2%).

Predictive performance

Sixty nine of the 84 validation studies (82%) assessed model discrimination (supplementary table K). All of these studies used the C index, with values ranging from 0.55 to 0.96. Model calibration was assessed in 15 studies, most commonly by using the Hosmer-Lemeshow test (n=11, 73%). Calibration plots were

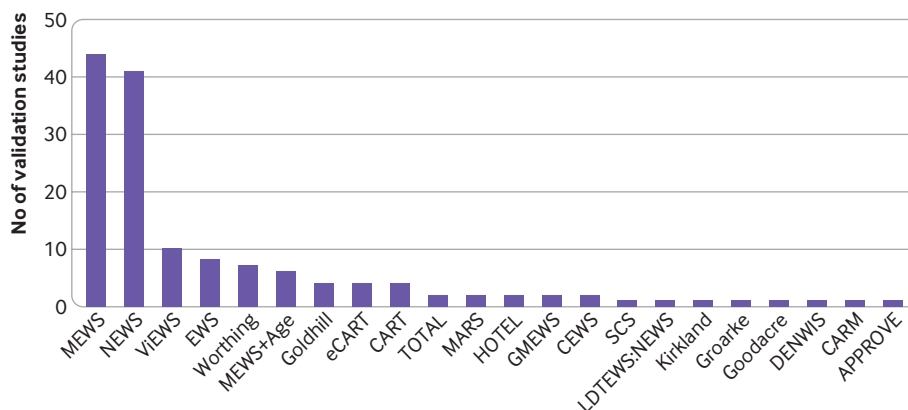


Fig 3 | Frequency of external model validation by early warning score (EWS) in 84 included validation studies. Eight EWSs had never been externally validated. APPROVE=accurate prediction of prolonged ventilation; CARM=computer aided risk of mortality; CART=cardiac arrest risk triage; CEWS=centile early warning score; DENWIS=Dutch early nurse worry indicator score; eCART=electronic cardiac arrest risk triage; GMEWS=global modified early warning score; HOTEL=hypotension, oxygen saturation, temperature, ECG [electrocardiogram] abnormality, loss of independence; LDTEWS=laboratory decision tree early warning score; MARS=medical admissions risk system; MEWS=modified early warning score; NEWS=national early warning score; SCS=simple clinical score; TOTAL=tachypnoea, oxygen saturation, temperature, alert and loss of independence; ViEWS=VitalPAC early warning score.

presented in four studies (27%). Other commonly reported performance metrics included sensitivity and specificity (n=49, 58%), and positive or negative predictive values (n=31, 37%). Overall performance metrics, such as the Brier score and R^2 , were not reported in any of the studies.

Because of the heterogeneity of outcomes and time horizons used in the validation studies, and the relative lack of head-to-head comparisons, we did not quantitatively synthesise performance metrics for specific EWSs.

PROBAST risk of bias assessment

We assessed risk of bias for each study, focusing on participant selection, predictors, outcomes, and analysis (fig 5). Participant selection was at low risk of bias in 42% of studies and at high risk of bias in 55%. For the remaining studies, risk of bias was unclear. Predictors were at low risk of bias in 91% of the

studies, and at high risk in 5%. Outcomes were at low risk of bias in 31% of studies, and at high risk in 66%. The analysis methods were at high risk of bias in all but two studies (98%).

Discussion

Our review of 95 published studies found poor methods and inadequate reporting in most studies that developed or validated EWSs. Problems were observed across all aspects of study design and analysis. We found that handling of statistical issues, such as missing data and regression modelling approaches, was inadequate. Few studies assessed calibration, an essential aspect of model performance, and no study assessed clinical utility using net benefit approaches.¹²⁴ Many studies also failed to report important details, such as sample size, number of events, population characteristics, and details of statistical methods. Several studies failed to report the full model, preventing (independent) external validation or implementation of the model in practice.

EWSs developed using inadequate methods will probably result in poorly performing scoring systems that fail to predict deterioration.¹²⁵ Poor methods in external validation studies could lead to implementation of inferior scoring systems, with false reassurances about their predictive ability and generalisability. These reports could explain why recent systematic reviews have found little evidence of any clinical effectiveness of EWSs.^{24 126} Although formal assessment of the methods and reporting quality in EWSs is needed, some reviews have found that studies describing the development or validation of an EWS were low quality, used poor statistical methods, and were at high risk of bias.^{23 24}

We assessed risk of bias by using PROBAST³⁰ and found that most of the included studies were at risk of bias owing to participant selection, outcome

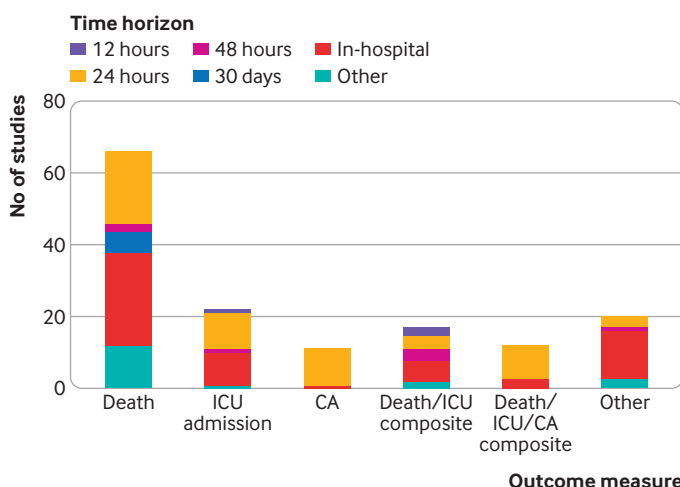


Fig 4 | Summary of outcomes and time horizons used in 84 studies externally validating an early warning score. CA=cardiac arrest; ICU=intensive care unit

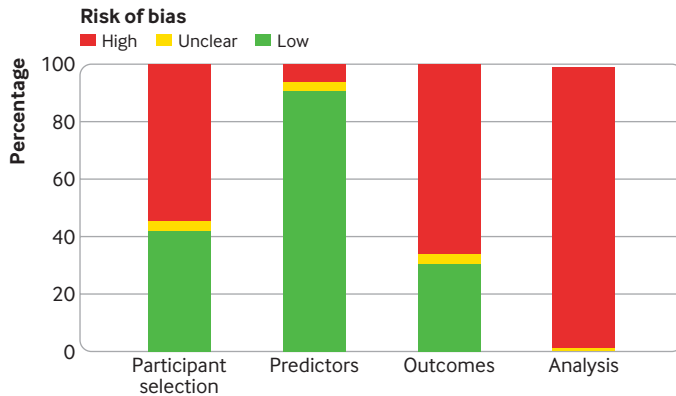


Fig 5 | Summary of risk of bias in four domains of 95 studies developing or validating an early warning score, assessed using PROBAST (prediction model risk of bias assessment tool)

definitions, and statistical analysis. The only domain for which most of the studies were at low risk of bias was predictor selection. Overall, all studies were at high risk of bias.

Our study included more external validation studies than development studies (11 development studies, 61 validation studies, and 23 studies that both developed and validated a model), which differs from reviews conducted in other clinical areas.^{19,21} Our eligibility criteria might partly explain this difference. We stipulated that a development study should not be included if a model is developed for a specific subpopulation (eg, patients with respiratory disease), but an external validation study could be included if an EWS developed for a general population is evaluated in a specific subpopulation. A relatively large number of studies reported the use of prospective data (24% development studies and 30% validation studies). Because the data required for development and validation of EWSs are commonly collected routinely, some authors might be reporting a prospective decision to use future routinely acquired data rather than the implementation of a prospective data acquisition process.

EWSs have historically been implemented as part of bedside paper observation charts. Because the scores were calculated manually, simple scoring systems were necessary. These systems often relied on assigning points to each vital sign, typically three points, and summing the points to get a total score. However, these systems make the unlikely assumption that each vital sign has the same predictive value.⁸ The total score has little meaning, and no obvious correspondence to an absolute risk of an event exists.

Electronic health records are increasingly being used to record vital signs and calculate EWSs.¹²⁷ These records allow more sophisticated EWSs to be implemented that make full use of available data and can be integrated into the clinical workflow of healthcare providers. Because the adoption of digital vital signs charting inevitably leads to further research, it is important that this research is of the highest quality, particularly when interest has surged

in using machine learning or artificial intelligence. The results of our review suggest recommendations for future research (box 2).

Recommendations for future research practice

Many of the recommendations are covered by the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement, which is a reporting guideline for studies developing or validating prognostic (or diagnostic) models.

Describe the data

We found that datasets were often not described in sufficient detail to understand in whom the model was intended for use or in whom the model was evaluated. These details are crucial when interpreting an article that describes an EWS. We recommend that several critical factors should be reported by all studies: number of patients with and without the event of interest; whether multiple observation sets are used per patient—if so, the total number of observation sets with and without an associated event; data source (eg, country, hospital, and wards); patient characteristics (eg, age, sex, and admission method).

Use sufficiently large sample size

Although many of the studies in our review used a large sample acquired from electronic health records, some used a sample that was too small. For example, a quarter of model development studies had fewer than six events for each variable at the patient level, and four at the observation level. A quarter of external validation studies included fewer than 460 patients, and a quarter of studies included fewer than 35 event patients. As the outcomes used in EWS studies are usually rare (~1-2%), and the number of events is a critical factor, large sample sizes are often necessary. Guidance suggests that external validation studies require a minimum of 100 event patients, and preferably more than 200.¹²⁸ Therefore, with their low event rates, EWS studies require data from many thousands of patients. Although defining the necessary sample size for model development studies is more complex, new guidance is available, which should be considered before embarking on new EWS studies.^{129, 130} Data driven variable selection methods increase the chance of overfitting and therefore should be avoided if possible.

Account for missing data

Most of the included studies mentioned missing data (86% of development studies and 75% of validation studies), although most of these studies used a complete case analysis to deal with missing data. Data are usually not missing at random, but are using missing selectively, for example based on patient characteristics or illness severity. Therefore, excluding records with incompletely observed predictor or outcome data can result in serious bias^{131, 132}; for example, by inflating associations between predictors and outcomes.

Box 2: Summary of recommendations for future practice**Provide key details of analysis population**

We suggest that articles report population demographics (eg, age and sex), source of data (country, hospital, and ward), number of patients with and without the event of interest, and number of observation sets with and without the event of interest.

Use large enough sample size

The sample size should be sufficient to robustly answer the question. For model development studies we recommend performing a sample size calculation specific to the context. For external validation studies we suggest including at least 100 event patients.

Describe amount of missing data and use statistical methods to account for missing data

Describe the frequency of missing data for each predictor and outcome. We recommend multiple imputation is the best practice approach for accounting for missing data in the analysis.

Carefully consider outcome measures and time horizons

Use an outcome measure that is clinically meaningful (that is, an outcome measure that can be prevented by appropriate treatment), and a time horizon in which deterioration can reasonably be expected to occur, and thus be predicted, which is probably a few days at most.

Use best practice statistical methods and report full model

If using a regression modelling approach to develop a new EWS, studies should allow for nonlinear predictor-outcome relations (eg, fractional polynomials) and avoid categorising predictors before analysis. Predictor interaction terms and competing risk approaches should be considered if appropriate. Newly developed models should always be fully described to allow independent evaluation and implementation.

Always carry out internal validation of new models

Internal validation is an important way of assessing how optimistic newly developed models might be. Split sample validation should be avoided and bootstrapping should be used.

Test all aspects of model performance

Assess both calibration and discrimination of EWSs. We also recommend using decision curve analysis to evaluate clinical utility.

We recommend that every study should describe how missing data were handled (for example, using complete case analysis, single imputation, or multiple imputation). The studies should also describe the amount of missing data overall, and for each predictor variable and outcome. We recommend that complete case analyses be avoided. Instead imputation approaches should be considered, with missing data imputed based on other known information. These approaches are now easy to implement in all standard software packages. Multiple imputation is widely regarded as the best approach.¹³³⁻¹³⁵ This method allows the uncertainty about missing data to be accounted for by creating multiple imputed datasets, then appropriately combining the results from each dataset. Before implementing multiple imputation the likely missing data mechanism should be thoughtfully considered. If imputation is appropriate, the setup of the imputation model should also be carefully considered (eg, the handling of categorical and skewed variables), and fully reported.¹³⁶

Use appropriate outcome measures and time horizons

The included studies used a variety of outcome measures and time horizons to develop and validate

EWSs. Both development and validation studies frequently used death and unanticipated intensive care unit admission, along with a variety of composite outcomes that included these outcomes. Some debate exists about which outcome measure is most appropriate.²⁷

We found that 39% of development studies and 52% of validation studies included a time horizon that was either in-hospital or 30 days. These long term horizons will not lead to models that give early warning of deterioration.²⁴ Instead the resulting models will identify generally unwell patients who are more likely to die or be admitted to the intensive care unit. We recommend that the time horizon should be limited to a few days at most, as any signs of deterioration linked to an observed outcome will probably not be seen for longer than this period.

Use best practice statistical approaches and report the full model

We observed that several of the articles reported regression modelling approaches that were methodologically weak. Nonlinear relations between predictors and outcomes were only included in 23% of development studies. However, this is an area of research in which such relations might readily exist. For example, both low and high respiratory rates can indicate increased risk. Similarly, interactions between predictors were only considered in 22% of studies. Models to predict the individual outcomes of intensive care unit admission or cardiac arrest were relatively frequent, but few studies accounted for death as a competing risk (intensive care unit admission or cardiac arrest not being possible if death has occurred). Failure to account for death as a competing risk could lead to a biased model, and inaccurate model predictions.¹³⁷ We recommend that future work accounts for competing risks in model development using Fine and Gray, cause specific hazards, or absolute risk regression¹³⁸⁻¹⁴¹ rather than logistic and Cox proportional hazards regression models. External validation of such models also requires that the potential of competing risks is taken into account.^{142 143} We also observed that the full model (all regression coefficients and either an intercept or baseline survival) was poorly reported, with only 39% of studies reporting sufficient information to allow independent validation or implementation.

We recommend that future development studies use best practice statistical methods, including examining plausible interaction terms (which should be chosen a priori and not data driven), examining nonlinear relations, avoiding univariable selection methods, and reporting all regression modelling coefficients. The methods used should be fully described in the publication and follow the recommendations laid out in the TRIPOD statement.¹⁷

Use internal validation for new models

The apparent performance of a newly developed model on its development data is likely to be optimistic, and better than its performance when applied to external

data. This optimism can be driven by a small sample size, many predictors, or categorisation of continuous variables. Internal validation quantifies the optimism and adjusts the apparent performance, and can be used to shrink the regression coefficients.^{144 145} Although many studies randomly split their dataset into two parts, one for model development and one for validation, this approach is weak and inefficient.¹⁴⁴ We found that 16 of our 34 articles included development studies that internally validated their EWS. However, 11 used a split sample approach.

Cross validation and bootstrapping are two preferred approaches for internal validation.¹⁴⁶ These methods use the entire dataset to both develop and validate the model. They also correct for overfitting in the model performance. We prefer bootstrapping because it can account for the optimism associated with the full model building process (eg, variable selection methods), and it can also provide a mechanism to shrink the regression coefficients to compensate for overfitting.^{144 147}

We recommend that new EWSs be internally validated, using bootstrapping if possible. However, we recognise that large datasets are becoming ever more present. In this context bootstrapping can be time consuming and less worthwhile when large datasets are used because overfitting in these instances is less likely. We recommend a form of the split sample approach is carried out with large datasets, where the dataset is not split randomly, but according to time, location, or centre.²⁹

Assess all aspects of model performance

Two key aspects characterise the performance of a prediction model, discrimination and calibration.^{17 148} Discrimination refers to a prediction model's ability to differentiate between those who develop an outcome and those who do not. A model should predict higher risks for those who develop the outcome. Calibration reflects the level of agreement between observed outcomes and the model predictions. In development studies the main emphasis will be on discrimination because the model will, by definition, be well calibrated. However, in external validation studies, both discrimination and calibration are important. Most of our included studies assessed model performance by using the C index, which has been observed in other prediction model reviews.^{18 21 149}

¹⁵⁰ For example, although 82% of external validation articles reported a measure of discrimination, only 18% reported an assessment of calibration. Those that assessed calibration used weak methods that are not recommended. Only four articles (5%) presented the preferred approach to assess calibration, the calibration plot.

We recommend that both discrimination and calibration be assessed in external validation studies, in line with TRIPOD recommendations. Calibration should be assessed with a plot that compares predicted and observed risks, with a smoothed curve plotted using LOESS (locally estimated scatterplot smoothing)

or similar methods, such as fractional polynomials or restricted cubic splines.¹⁵¹ Other metrics of calibration such as the intercept and slope should also be reported.¹⁵² Many EWSs are currently based on an integer scoring system. For example, NEWS ranges from 0 to 20 points. Calibration of an integer scoring system cannot be assessed because it relies on the model producing predicted probabilities. Overall performance measures, which combine discrimination and calibration, should also be considered, such as R^2 and the Brier score.¹⁵² The more clinically meaningful decision curve analysis (or net benefit) approach is also recommended.¹²⁴

Weaknesses of the study

We assessed 34 development studies, which is perhaps fewer than expected compared with previous systematic reviews.¹¹¹ We excluded several existing EWSs that have not been published in peer reviewed academic journals. However, we anticipate that the methods underlying these excluded EWSs will be of a similar standard, and possibly even worse, than those included in our review.

Our eligibility criteria state that external validation studies would only be included if the development study was also included. However, we chose to make an exception for studies that described external validation of Morgan's original 1997 EWS, and Subbe's modified early warning score. Otherwise this eligibility criterion excluded few articles.

Some other details that were not collected could be of interest for future investigation. For example, research could include the rationale for the choice of outcome measure and the prediction time horizon, and whether EWSs have been developed or validated by using and accounting for multicentre or clustered data.¹⁵³

Strengths of the study

This systematic review formally assessed the methods and reporting standards in EWS studies. We performed a thorough assessment of important aspects of development and validation based on the CHARMS checklist,²⁹ and other important subject specific items. We also assessed risk of bias using the PROBAST tool.³⁰

Conclusion

We included 95 articles in our review that developed or externally validated EWSs. We found many methodological and reporting shortcomings. Therefore, EWSs in common use could perform more poorly than reported, with potentially detrimental effects on patient care. Clinical responses to elevated scores have major workload impacts,¹⁵⁴ and the weaknesses of the EWSs affect the resulting workload. Therefore, healthcare professionals and policy makers need to be aware of these weaknesses when recommending particular response strategies.

Our study does not seek to recommend a particular EWS, however NEWS is currently mandated for use throughout the National Health Service in the UK.¹⁵⁵ This system was developed by clinical consensus

rather than by applying statistical methods, which is the usual method for developing prediction models. Claims of extensive validation¹² might be misleading because we found the underlying methodology of EWS validation studies to be generally poor. In reality, clinicians can have little knowledge of how such scores will perform in their clinical setting. Therefore, clinicians should be cautious about relying on these scores to identify clinical deterioration in patients.

The move towards electronic implementation of EWSs presents an opportunity to introduce better scoring systems, particularly with the increasing interest in modern model building approaches, such as machine learning and artificial intelligence. However, if methodological and reporting standards are not improved, this potential might never be achieved.

We acknowledge and thank the patients and members of the public who have provided input into this study.

Contributors: SG, TB, JB, SK, PJW, and GSC designed the study. SG and SK developed the search strings. SG, JB, and PSV extracted the data. SG analysed the data and wrote the first draft of the manuscript. All authors revised the manuscript and approved the final version of the submitted manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. SG acts as the guarantor.

Funding: SG is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (DRF-2016-09-173). JB, PJW, and GSC are funded by the NIHR Oxford Biomedical Research Centre. SK and GSC are funded by Cancer Research UK (grant No C49297/A27294). PV is funded by a NIHR Doctoral Research Fellowship (DRF-2018-11-ST2-057). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care, or Cancer Research UK.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: support from the National Institute for Health Research and Cancer Research UK for the submitted work; PJW is chief medical officer for Sensyne Health and holds shares in the company; TB receives royalties from Sensyne Health; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: No additional data available.

The lead authors affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Dissemination to participants and related patient and public communities: There are no study participants to disseminate the results to. Patients and participants involved in the study design and interpretation will be sent a copy of the article. The authors intend to disseminate the study findings through media organisations so that the results will be available for the wider patient and public community.

This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: <http://creativecommons.org/licenses/by/4.0/>.

- Brennan TA, Leape LL, Laird NM, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N Engl J Med* 1991;324:370-6. doi:10.1056/NEJM199102073240604
- Donaldson M, Corrigan J, Kohn L. To err is human: building a safer health system. 2000. doi.org/10.17226/9728
- Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. *BMJ* 2001;322:517-9. doi:10.1136/bmj.322.7285.517
- Hillman KM, Bristow PJ, Chey T, et al. Duration of life-threatening antecedents prior to intensive care admission. *Intensive Care Med* 2002;28:1629-34. doi:10.1007/s00134-002-1496-y
- Morgan R, Lloyd-Williams F, Wright M, Morgan-Warren RJ. An early warning scoring system for detecting developing critical illness.

1997. <https://www.scienceopen.com/document?vid=28251d22-8476-40a6-916d-1a34796816e4>
- Prince Charles Hospital. Modified Early Warning Score (MEWS), Escalation and ISBAR, The Prince Charles Hospital Procedure TPCHS10085.
- Deteriorating Patient Programme and National Early Warning System (NEWS) - HSE.ie. <https://www.hse.ie/eng/about/who/cspd/ncps/acute-medicine/national-early-warning-score/>.
- Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs. *Crit Care Med* 2012;40:2102-8. doi:10.1097/CCM.0b013e318250aa5a
- Douw G, Huisman-de Waal G, van Zanten AR, van der Hoeven JG, Schoonhoven L. Nurses' 'worry' as predictor of deteriorating surgical ward patients: a prospective cohort study of the Dutch-Early-Nurse-Worry-Indicator-Score. *Int J Nurs Stud* 2016;59:134-40. doi:10.1016/j.ijnurstu.2016.04.006
- National Institute for Health and Care Excellence. Acutely ill adults in hospital: recognising and responding to deterioration. <https://www.nice.org.uk/guidance/cg50>.
- Gerry S, Birks J, Bonnici T, Watkinson PJ, Kirtley S, Collins GS. Early warning scores for detecting deterioration in adult hospital patients: a systematic review protocol. *BMJ Open* 2017;7:e019268. doi:10.1136/bmjopen-2017-019268
- Royal College of Physicians. *Royal College of Physicians National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. Updated report of a working party.* RCP, 2017.
- Silcock DJ, Corfield AR, Gowens PA, Rooney KD. Validation of the National Early Warning Score in the prehospital setting. *Resuscitation* 2015;89:31-5. doi:10.1016/j.resuscitation.2014.12.029
- Corfield AR, Lees F, Zealley I, et al. Scottish Trauma Audit Group Sepsis Steering Group. Utility of a single early warning score in patients with sepsis in the emergency department. *Emerg Med J* 2014;31:482-7. doi:10.1136/emered-2012-202186
- Brangan E, Banks J, Brant H, Pullyblank A, Le Roux H, Redwood S. Using the National Early Warning Score (NEWS) outside acute hospital settings: a qualitative study of staff experiences in the West of England. *BMJ Open* 2018;8:e022528. doi:10.1136/bmjopen-2018-022528
- Inada-Kim M, Nsutebu E. NEWS 2: an opportunity to standardise the management of deterioration and sepsis. *BMJ* 2018;360:k1260. doi:10.1136/bmj.k1260
- Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55-63. doi:10.7326/M14-0697
- Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
- Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: a systematic review. *BMJ* 2016;353:i2416. doi:10.1136/bmj.i2416
- Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9:103. doi:10.1186/1741-7015-9-103
- Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9:1-12. doi:10.1371/journal.pmed.1001221
- Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66:268-77. doi:10.1016/j.jclinepi.2012.06.020
- 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. doi:10.1007/s00134-007-0532-3
- 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. doi:10.1513/AnnalsATS.201403-102OC
- Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24:198-208. doi:10.1093/jamia/ocw042
- Goldstein BA, Pomann GM, Winkelmayr WC, Pencina MJ. A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis. *Stat Med* 2017;36:2750-63. doi:10.1002/sim.7308
- Churpek MM, Yuen TC, Edelson DP. Predicting clinical deterioration in the hospital: the impact of outcome selection. *Resuscitation* 2013;84:564-8. doi:10.1016/j.resuscitation.2012.09.024

- 28 Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377-81. doi:10.1016/j.jbi.2008.08.010
- 29 Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744. doi:10.1371/journal.pmed.1001744
- 30 Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170:51-8. doi:10.7326/M18-1376
- 31 Albert BL, Huesman L. Development of a modified early warning score using the electronic medical record. *Dimens Crit Care Nurs* 2011;30:283-92. doi:10.1097/DCC.0b013e318227761d
- 32 Alvarez CA, Clark CA, Zhang S, et al. Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC Med Inform Decis Mak* 2013;13:28. doi:10.1186/1472-6947-13-28.
- 33 Badriyah T, Briggs JS, Meredith P, et al. Decision-tree early warning score (DTEWS) validates the design of the National Early Warning Score (NEWS). *Resuscitation* 2014;85:418-23. doi:10.1016/j.resuscitation.2013.12.011
- 34 Bleyer AJ, Vidya S, Russell GB, et al. Longitudinal analysis of one million vital signs in patients in an academic medical center. *Resuscitation* 2011;82:1387-92. doi:10.1016/j.resuscitation.2011.06.033
- 35 Churpek MM, Yuen TC, Park SY, Gibbons R, Edelson DP. Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards*. *Crit Care Med* 2014;42:841-8. doi:10.1097/CCM.0000000000000038
- 36 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;190:649-55. doi:10.1164/rccm.201406-1022OC
- 37 Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation* 2016;102:1-5. doi:10.1016/j.resuscitation.2016.02.005
- 38 Cuthbertson BH, Boroujerdi M, Prescott G. The use of combined physiological parameters in the early recognition of the deteriorating acute medical patient. *J R Coll Physicians Edinb* 2010;40:19-25. doi:10.4997/JRCP.2010.105
- 39 Duckitt RW, Buxton-Thomas R, Walker J, et al. Worthing physiological scoring system: derivation and validation of a physiological early-warning system for medical admissions. An observational, population-based single-centre study. *Br J Anaesth* 2007;98:769-74. doi:10.1093/bja/aem097
- 40 Dziadzko MA, Novotny PJ, Sloan J, et al. Multicenter derivation and validation of an early warning score for acute respiratory failure or death in the hospital. *Crit Care* 2018;22:286. doi:10.1186/s13054-018-2194-7
- 41 Escobar GJ, LaGuardia JC, Turk BJ, Ragins A, Kipnis P, Draper D. Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. *J Hosp Med* 2012;7:388-95. doi:10.1002/jhm.1929
- 42 Faisal M, Scally AJ, Jackson N, et al. Development and validation of a novel computer-aided score to predict the risk of in-hospital mortality for acutely ill medical admissions in two acute hospitals using their first electronically recorded blood test results and vital signs: a cross-sectional study. *BMJ Open* 2018;8:e022939. doi:10.1136/bmjopen-2018-022939
- 43 Ghosh E, Eshelman L, Yang L, Carlson E, Lord B. Early Deterioration Indicator: Data-driven approach to detecting deterioration in general ward. *Resuscitation* 2018;122:99-105. doi:10.1016/j.resuscitation.2017.10.026
- 44 Goldhill DR, McNarry AF. Physiological abnormalities in early warning scores are related to mortality in adult inpatients. *Br J Anaesth* 2004;92:882-4. doi:10.1093/bja/ae113
- 45 Harrison GA, Jacques T, McLaws ML, Kilborn G. Combinations of early signs of critical illness predict in-hospital death—the SOCCER study (signs of critical conditions and emergency responses). *Resuscitation* 2006;71:327-34. doi:10.1016/j.resuscitation.2006.05.008
- 46 Jones M. NEWSDIG: The National Early Warning Score Development and Implementation Group. *Clin Med (Lond)* 2012;12:501-3. doi:10.7861/clinmedicine.12-6-501
- 47 Kellett J, Deane B. The Simple Clinical Score predicts mortality for 30 days after admission to an acute medical unit. *QJM* 2006;99:771-81. doi:10.1093/qjmed/hcl112
- 48 Kellett J, Deane B, Gleeson M. Derivation and validation of a score based on Hypotension, Oxygen saturation, low Temperature, ECG changes and Loss of independence (HOTEL) that predicts early mortality between 15 min and 24 h after admission to an acute medical unit. *Resuscitation* 2008;78:52-8. doi:10.1016/j.resuscitation.2008.02.011
- 49 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;64:10-9. doi:10.1016/j.jbi.2016.09.013
- 50 Kirkland LL, Malinchoc M, O'Byrne M, et al. A clinical deterioration prediction tool for internal medicine patients. *Am J Med Qual* 2013;28:135-42. doi:10.1177/1062860612450459
- 51 Kwon JM, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J Am Heart Assoc* 2018;7:26. doi:10.1161/JAHA.118.008678
- 52 Kyriacos O, Uelsma J, James M, Jordan S. Monitoring vital signs: development of a modified early warning scoring (MEWS) system for general wards in a developing country. *PLoS One* 2014;9:e87073. doi:10.1371/journal.pone.0087073
- 53 Luís L, Nunes C. Short National Early Warning Score—developing a modified early warning score. *Aust Crit Care* 2018;31:376-81. doi:10.1016/j.auc.2017.11.004
- 54 Moore CC, Hazard R, Saulters KJ, et al. Derivation and validation of a universal vital assessment (UVA) score: a tool for predicting mortality in adult hospitalised patients in sub-Saharan Africa. *BMJ Glob Health* 2017;2:e000344. doi:10.1136/bmjgh-2017-000344
- 55 Nickel CH, Kellett J, Cooksley T, Bingisser R, Henriksen DP, Brabrand M. Combined use of the National Early Warning Score and D-dimer levels to predict 30-day and 365-day mortality in medical patients. *Resuscitation* 2016;106:49-52. doi:10.1016/j.resuscitation.2016.06.012
- 56 Perera YS, Ranasinghe P, Adikari AM, et al. The value of the Modified Early Warning Score and biochemical parameters as predictors of patient outcome in acute medical admissions a prospective study. *Acute Med* 2011;10:126-32.
- 57 Prytherch DR, Smith GB, Schmidt PE, Featherstone PI. ViEWS—Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation* 2010;81:932-7. doi:10.1016/j.resuscitation.2010.04.014
- 58 Redfern OC, Pimentel MAF, Prytherch D, et al. Predicting in-hospital mortality and anticipated admissions to the intensive care unit using routinely collected blood tests and vital signs: Development and validation of a multivariable model. *Resuscitation* 2018;133:75-81. doi:10.1016/j.resuscitation.2018.09.021
- 59 Silke B, Kellett J, Rooney T, Bennett K, O'Riordan D. An improved medical admissions risk system using multivariable fractional polynomial logistic regression modelling. *QJM* 2010;103:23-32. doi:10.1093/qjmed/hcp149
- 60 Tarassenko L, Clifton DA, Pinsky MR, Hravnak MT, Woods JR, Watkinson PJ. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation* 2011;82:1013-8. doi:10.1016/j.resuscitation.2011.03.006
- 61 Watkinson PJ, Pimentel MAF, Clifton DA, Tarassenko L. Manual centile-based early warning scores derived from statistical distributions of observational vital-sign data. *Resuscitation* 2018;129:55-60. doi:10.1016/j.resuscitation.2018.06.003
- 62 Wheeler I, Price C, Sitch A, et al. Early warning scores generated in developed healthcare settings are not sufficient at predicting early mortality in Blantyre, Malawi: a prospective cohort study [correction: *PLoS One* 2014;9:e91623]. *PLoS One* 2013;8:e59830. doi:10.1371/journal.pone.0059830
- 63 Abbott TEF, Torrance HDT, Cron N, Vaid N, Emmanuel J. A single-centre cohort study of National Early Warning Score (NEWS) and near patient testing in acute medical admissions. *Eur J Intern Med* 2016;35:78-82. doi:10.1016/j.ejim.2016.06.014
- 64 Abbott TE, Vaid N, Ip D, et al. A single-centre observational cohort study of admission National Early Warning Score (NEWS). *Resuscitation* 2015;92:89-93. doi:10.1016/j.resuscitation.2015.04.020
- 65 Atmaca Ö, Turan C, Güven P, Ankan H, Eryüksel SE, Karakurt S. Usage of NEWS for prediction of mortality and in-hospital cardiac arrest rates in a Turkish university hospital. *Turk J Med Sci* 2018;48:1087-91. doi:10.3906/sag-1706-67
- 66 Bartkowiak B, Snyder AM, Benjamin A, et al. Validating the Electronic Cardiac Arrest Risk Triage (eCART) score for risk stratification of surgical inpatients in the postoperative setting: retrospective cohort study. *Ann Surg* 2019;269:1059-63. doi:10.1097/SLA.0000000000002665
- 67 Beane A, De Silva AP, De Silva N, et al. Evaluation of the feasibility and performance of early warning scores to identify patients at risk of adverse outcomes in a low-middle income country setting. *BMJ Open* 2018;8:e019387. doi:10.1136/bmjopen-2017-019387
- 68 Brabrand M, Hallas S, Hansen SN, Jensen KM, Madsen JLB, Posth S. Using scores to identify patients at risk of short term mortality at arrival to the acute medical unit: A validation study of six existing scores. *Eur J Intern Med* 2017;45:32-6. doi:10.1016/j.ejim.2017.09.042
- 69 Brabrand M, Henriksen DP. CURB-65 score is equal to NEWS for identifying mortality risk of pneumonia patients: An observational study. *Lung* 2018;196:359-61. doi:10.1007/s00408-018-0105-y

- 70 Cei M, Bartolomei C, Mumoli N. In-hospital mortality and morbidity of elderly medical patients can be predicted at admission by the Modified Early Warning Score: a prospective study. *Int J Clin Pract* 2009;63:591-5. doi:10.1111/j.1742-1241.2008.01986.x
- 71 Churpek MM, Snyder A, Sokol S, Pettit NN, Edelson DP. Investigating the impact of different suspicion of infection criteria on the accuracy of quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores. *Crit Care Med* 2017;45:1805-12. doi:10.1097/CCM.0000000000002648
- 72 Churpek MM, Snyder A, Han X, et al. Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside the intensive care unit. *Am J Respir Crit Care Med* 2017;195:906-11. doi:10.1164/rccm.201604-0854OC
- 73 Churpek MM, Yuen TC, Edelson DP. Risk stratification of hospitalized patients on the wards. *Chest* 2013;143:1758-65. doi:10.1378/chest.12-1605
- 74 Churpek MM, Yuen TC, Huber MT, Park SY, Hall JB, Edelson DP. Predicting cardiac arrest on the wards: a nested case-control study. *Chest* 2012;141:1170-6. doi:10.1378/chest.11-1301
- 75 Cooksley T, Kitlowski E, Haji-Michael P. Effectiveness of Modified Early Warning Score in predicting outcomes in oncology patients. *QJM* 2012;105:1083-8. doi:10.1093/qjmed/hcs138
- 76 De Meester K, Das T, Hellemans K, et al. Impact of a standardized nurse observation protocol including MEWS after Intensive Care Unit discharge. *Resuscitation* 2013;84:184-8. doi:10.1016/j.resuscitation.2012.06.017
- 77 DeVoe B, Roth A, Maurer G, et al. Correlation of the predictive ability of early warning metrics and mortality for cardiac arrest patients receiving in-hospital Advanced Cardiovascular Life Support. *Heart Lung* 2016;45:497-502. doi:10.1016/j.hrtlng.2016.08.010
- 78 Douw G, Huisman-de Waal G, van Zanten ARH, van der Hoeven JG, Schoonhoven L. Capturing early signs of deterioration: the dutch-early-nurse-worry-indicator-score and its value in the Rapid Response System. *J Clin Nurs* 2017;26:2605-13. doi:10.1111/jocn.13648
- 79 Eccles SR, Subbe C, Hancock D, Thomson N. CREWS: improving specificity whilst maintaining sensitivity of the National Early Warning Score in patients with chronic hypoxaemia. *Resuscitation* 2014;85:109-11. doi:10.1016/j.resuscitation.2013.08.277
- 80 Fairclough E, Cairns E, Hamilton J, Kelly C. Evaluation of a modified early warning system for acute medical admissions and comparison with C-reactive protein/albumin ratio as a predictor of patient outcome. *Clin Med (Lond)* 2009;9:30-3. doi:10.7861/clinmedicine.9-1-30
- 81 Finlay GD, Rothman MJ, Smith RA. Measuring the modified early warning score and the Rothman index: advantages of utilizing the electronic medical record in an early warning system. *J Hosp Med* 2014;9:116-9. doi:10.1002/jhm.2132
- 82 Forster S, Housley G, McKeever TM, Shaw DE. Investigating the discriminative value of Early Warning Scores in patients with respiratory disease using a retrospective cohort analysis of admissions to Nottingham University Hospitals Trust over a 2-year period. *BMJ Open* 2018;8:e020269. doi:10.1136/bmjopen-2017-020269
- 83 Garcea G, Jackson B, Pattenden CJ, et al. Early warning scores predict outcome in acute pancreatitis. *J Gastrointest Surg* 2006;10:1008-15. doi:10.1016/j.gassur.2006.03.008
- 84 Gardner-Thorpe J, Love N, Wrightson J, Walsh S, Keeling N. The value of Modified Early Warning Score (MEWS) in surgical in-patients: a prospective observational study. *Ann R Coll Surg Engl* 2006;88:571-5. doi:10.1308/003588406X130615
- 85 Ghanem-Zoubi NO, Vardi M, Laor A, Weber G, Bitterman H. Assessment of disease-severity scoring systems for patients with sepsis in general internal medicine departments. *Crit Care* 2011;15:R95. doi:10.1186/cc10102
- 86 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
- 87 Hodgson LE, Dimitrov BD, Congleton J, Venn R, Forni LG, Roderick PJ. A validation of the National Early Warning Score to predict outcome in patients with COPD exacerbation. *Thorax* 2017;72:23-30. doi:10.1136/thoraxjnl-2016-208436
- 88 Hydes TJ, Meredith P, Schmidt PE, Smith GB, Prytherch DR, Aspinall RJ. National Early Warning Score accurately discriminates the risk of serious adverse events in patients with liver disease. *Clin Gastroenterol Hepatol* 2018;16:1657-1666.e10. doi:10.1016/j.cgh.2017.12.035
- 89 Jo S, Jeong T, Lee JB, Jin Y, Yoon J, Park B. Validation of modified early warning score using serum lactate level in community-acquired pneumonia patients. The National Early Warning Score-Lactate score. *Am J Emerg Med* 2016;34:536-41. doi:10.1016/j.ajem.2015.12.067
- 90 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. doi:10.1016/j.resuscitation.2011.08.022
- 91 Kellett J, Murray A. Should predictive scores based on vital signs be used in the same way as those based on laboratory data? A hypothesis generating retrospective evaluation of in-hospital mortality by four different scoring systems. *Resuscitation* 2016;102:94-7. doi:10.1016/j.resuscitation.2016.02.020
- 92 Kim D, Jo S, Lee JB, et al. Comparison of the National Early Warning Score+Lactate score with the pre-endoscopic Rockall, Glasgow-Blatchford, and AIMS65 scores in patients with upper gastrointestinal bleeding. *Clin Exp Emerg Med* 2018;5:219-29. doi:10.15441/ceem.17.268
- 93 Kim WY, Lee J, Lee JR, et al. A risk scoring model based on vital signs and laboratory data predicting transfer to the intensive care unit of patients admitted to gastroenterology wards. *J Crit Care* 2017;40:213-7. doi:10.1016/j.jccr.2017.04.024
- 94 Kovacs C, Jarvis SW, Prytherch DR, et al. Comparison of the National Early Warning Score in non-elective medical and surgical patients. *Br J Surg* 2016;103:1385-93. doi:10.1002/bjs.10267
- 95 Kruisselbrink R, Kwizera A, Crowther M, et al. Modified Early Warning Score (MEWS) identifies critical illness among ward patients in a resource restricted setting in Kampala, Uganda: A prospective observational study. *PLoS One* 2016;11:e0151408. doi:10.1371/journal.pone.0151408
- 96 Le Lagadec MD, Dwyer T, Browne M. The efficacy of twelve early warning systems for potential use in regional medical facilities in Queensland, Australia. *Aust Crit Care* 2020;33:47-53. doi:10.1016/j.aucc.2019.03.001
- 97 Lee YS, Choi JW, Park YH, et al. Evaluation of the efficacy of the National Early Warning Score in predicting in-hospital mortality via the risk stratification. *J Crit Care* 2018;47:222-6. doi:10.1016/j.jccr.2018.07.011
- 98 Liljehult J, Christensen T. Early warning score predicts acute mortality in stroke patients. *Acta Neurol Scand* 2016;133:261-7. doi:10.1111/ane.12452
- 99 Mulligan A. Validation of a physiological track and trigger score to identify developing critical illness in haematology patients. *Intensive Crit Care Nurs* 2010;26:196-206. doi:10.1016/j.iccn.2010.03.002
- 100 Öhman MC, Atkins TEH, Cooksley T, Brabrand M. Validation of the MARS: a combined physiological and laboratory risk prediction tool for 5- to 7-day in-hospital mortality. *QJM* 2018;111:385-8. doi:10.1093/qjmed/hcy053
- 101 Opio MO, Nansubuga G, Kellett J. Validation of the VitalPAC™ Early Warning Score (VIEWS) in acutely ill medical patients attending a resource-poor hospital in sub-Saharan Africa. *Resuscitation* 2013;84:743-6. doi:10.1016/j.resuscitation.2013.02.007
- 102 Opio MO, Nansubuga G, Kellett J, Clifford M, Murray A. Performance of TOTAL, in medical patients attending a resource-poor hospital in sub-Saharan Africa and a small Irish rural hospital. *Acute Med* 2013;12:135-40.
- 103 Pedersen NE, Rasmussen LS, Petersen JA, Gerds TA, Østergaard D, Lippert A. Modifications of the National Early Warning Score for patients with chronic respiratory disease. *Acta Anaesthesiol Scand* 2018;62:242-52. doi:10.1111/aas.13020
- 104 Pimentel MAF, Redfern OC, Gerry S, et al. A comparison of the ability of the National Early Warning Score and the National Early Warning Score 2 to identify patients at risk of in-hospital mortality: A multi-centre database study. *Resuscitation* 2019;134:147-56. doi:10.1016/j.resuscitation.2018.09.026
- 105 Plate JD, Peelen LM, Leenen LP, Hietbrink F. Validation of the VitalPAC early warning score at the intermediate care unit. *World J Crit Care Med* 2018;7:39-45. doi:10.5492/wjccm.v7.i3.39
- 106 Redfern OC, Smith GB, Prytherch DR, Meredith P, Inada-Kim M, Schmidt PE. A comparison of the Quick Sequential (sepsis-related) Organ Failure Assessment score and the National Early Warning Score in non-ICU patients with/without infection. *Crit Care Med* 2018;46:1923-33. doi:10.1097/CCM.0000000000003359
- 107 Roberts D, Djärv T. Preceding national early warnings scores among in-hospital cardiac arrests and their impact on survival. *Am J Emerg Med* 2017;35:1601-6. doi:10.1016/j.ajem.2017.04.072
- 108 Romero-Brufau S, Morlan BW, Johnson M, et al. Evaluating automated rules for rapid response system alarm triggers in medical and surgical patients. *J Hosp Med* 2017;12:217-23. doi:10.12788/jhm.2712
- 109 Romero-Brufau S, Huddleston JM, Naessens JM, et al. Widely used track and trigger scores: are they ready for automation in practice? *Resuscitation* 2014;85:549-52. doi:10.1016/j.resuscitation.2013.12.017
- 110 Rylance J, Baker T, Mushi E, Mashaga D. Use of an early warning score and ability to walk predicts mortality in medical patients admitted to hospitals in Tanzania. *Trans R Soc Trop Med Hyg* 2009;103:790-4. doi:10.1016/j.trstmh.2009.05.004
- 111 Smith GB, Prytherch DR, Schmidt PE, Featherstone PI. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation* 2008;77:170-9. doi:10.1016/j.resuscitation.2007.12.004

- 112 Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 2013;84:465-70. doi:10.1016/j.resuscitation.2012.12.016
- 113 Smith GB, Prytherch DR, Jarvis S, et al. A comparison of the ability of the physiologic components of medical emergency team criteria and the U.K. National Early Warning Score to discriminate patients at risk of a range of adverse clinical outcomes. *Crit Care Med* 2016;44:2171-81. doi:10.1097/CCM.0000000000002000
- 114 Smith HJ, Pasko DN, Haygood CLW, Boone JD, Harper LM, Straughn JMjr. Early warning score: An indicator of adverse outcomes in postoperative patients on a gynecologic oncology service. *Gynecol Oncol* 2016;143:105-8. doi:10.1016/j.ygyno.2016.08.153
- 115 Spagnolli W, Rigoni M, Torri E, Cozzio S, Vettorato E, Nollo G. Application of the National Early Warning Score (NEWS) as a stratification tool on admission in an Italian acute medical ward: A perspective study. *Int J Clin Pract* 2017;71. doi:10.1111/ijcp.12934
- 116 Stark AP, Maciel RC, Sheppard W, Sacks G, Hines OJ. An early warning score predicts risk of death after in-hospital cardiopulmonary arrest in surgical patients. *Am Surg* 2015;81:916-21.
- 117 Stræde M, Brabrand M. External validation of the simple clinical score and the HOTEL score, two scores for predicting short-term mortality after admission to an acute medical unit. *PLoS One* 2014;9:e105695. doi:10.1371/journal.pone.0105695
- 118 Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *QJM* 2001;94:521-6. doi:10.1093/qjmed/94.10.521
- 119 Suppiah A, Malde D, Arab T, et al. The Modified Early Warning Score (MEWS): an instant physiological prognostic indicator of poor outcome in acute pancreatitis. *JOP* 2014;15:569-76. doi:10.6092/1590-8577/2829.
- 120 Tirkkonen J, Olkkola KT, Huhtala H, Tenhunen J, Hoppu S. Medical emergency team activation: performance of conventional dichotomised criteria versus national early warning score. *Acta Anaesthesiol Scand* 2014;58:411-9. doi:10.1111/aas.12277
- 121 Tirotta D, Gambacorta M, La Regina M, et al. Evaluation of the threshold value for the modified early warning score (MEWS) in medical septic patients: a secondary analysis of an Italian multicentric prospective cohort (SNOOPII study). *QJM* 2017;110:369-73. doi:10.1093/qjmed/hcw229
- 122 Vaughn JL, Kline D, Denlinger NM, Andritsos LA, Exline MC, Walker AR. Predictive performance of early warning scores in acute leukemia patients receiving induction chemotherapy. *Leuk Lymphoma* 2018;59:1498-500. doi:10.1080/10428194.2017.1376744
- 123 von Lilienfeld-Toal M, Midgley K, Lieberbach S, et al. Observation-based early warning scores to detect impending critical illness predict in-hospital and overall survival in patients undergoing allogeneic stem cell transplantation. *Biol Blood Marrow Transplant* 2007;13:568-76. doi:10.1016/j.bbmt.2006.12.455
- 124 Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: A guide for investigators. *Eur Urol* 2018;74:796-804. doi:10.1016/j.eururo.2018.08.038
- 125 McGaughey J, Alderdice F, Fowler R, Kapila A, Mayhew A, Moutray M. Outreach and Early Warning Systems (EWS) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. *Cochrane Database Syst Rev* 2007;(3):CD005529. doi:10.1002/14651858.CD005529.pub2.
- 126 Alam N, Hobbelenk EL, van Tienhoven AJ, van de Ven PM, Jansma EP, Nanayakkara PWB. The impact of the use of the Early Warning Score (EWS) on patient outcomes: a systematic review. *Resuscitation* 2014;85:587-94. doi:10.1016/j.resuscitation.2014.01.013
- 127 Wong D, Bonnici T, Knight J, Morgan L, Coombes P, Watkinson P. SEND: a system for electronic notification and documentation of vital sign observations. *BMC Med Inform Decis Mak* 2015;15:68. doi:10.1186/s12911-015-0186-y
- 128 Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* 2016;35:214-26. doi:10.1002/sim.6787
- 129 Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes [correction in: *Stat Med* 2019;38:5672]. *Stat Med* 2019;38:1276-96. doi:10.1002/sim.7992
- 130 van Smeden M, Moons KGM, de Groot JAH, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res* 2019;28:2455-74. doi:10.1177/0962280218784726
- 131 Little RJA. Regression with missing X's: a review. *J Am Stat Assoc* 1992;87:1227. doi:10.2307/2290664.
- 132 Janssen KJ, Donders ART, Harrell FEJr, et al. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol* 2010;63:721-7. doi:10.1016/j.jclinepi.2009.12.008
- 133 Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Stat Methods Med Res* 2007;16:277-98. doi:10.1177/0962280206074466
- 134 Little RJA, Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons, Inc, 2002. doi:10.1002/9781119013563
- 135 Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. *J Clin Epidemiol* 2010;63:205-14. doi:10.1016/j.jclinepi.2009.03.017
- 136 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011;30:377-99. doi:10.1002/sim.4067
- 137 Wolbers M, Koller MT, Wittman JCM, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology* 2009;20:555-61. doi:10.1097/EDE.0b013e3181a39056
- 138 Gerds TA, Scheike TH, Andersen PK. Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Stat Med* 2012;31:3921-30. doi:10.1002/sim.5459
- 139 Wolbers M, Koller MT, Stel VS, et al. Competing risks analyses: objectives and approaches. *Eur Heart J* 2014;35:2936-41. doi:10.1093/eurheartj/ehu131
- 140 Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* 2016;133:601-9. doi:10.1161/CIRCULATIONAHA.115.017719
- 141 Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med* 2013;32:5381-97. doi:10.1002/sim.5958
- 142 Austin PC, Lee DS, D'Agostino RB, Fine JP. Developing points-based risk-scoring systems in the presence of competing risks [correction in: *Stat Med* 2018;37:1405]. *Stat Med* 2016;35:4056-72. doi:10.1002/sim.6994
- 143 Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Stat Med* 2014;33:3191-203. doi:10.1002/sim.6152
- 144 Steyerberg EW, Harrell FEJr, Borsboom GJ, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54:774-81. doi:10.1016/S0895-4356(01)00341-9
- 145 Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73. doi:10.7326/M14-0698
- 146 Harrell FEJr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361-87. doi:10.1002/(SIC)1097-0258(19960229)15:4<361::AID-SIM1683>3.0.CO;2-4
- 147 Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567-86. doi:10.1002/sim.1844
- 148 Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016;74:167-76. doi:10.1016/j.jclinepi.2015.12.005
- 149 Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;6:38. doi:10.1186/1472-6947-6-38
- 150 Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Res Treat* 2012;132:365-77. doi:10.1007/s10549-011-1818-2
- 151 Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med* 2014;33:517-35. doi:10.1002/sim.5941
- 152 Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128-38. doi:10.1097/EDE.0b013e3181c30fb2
- 153 Wynants L, Kent DM, Timmerman D, Lundquist CM, Van Calster B. Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting. *Diagn Progn Res* 2019;3:6. doi:10.1186/s41512-019-0046-9
- 154 Linnen DT, Escobar GJ, Hu X, Scruth E, Liu V, Stephens C. Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: a systematic review. *J Hosp Med* 2019;14:161-9. doi:10.12788/jhm.3151
- 155 NHS England. *National Early Warning Score (NEWS)*. <https://www.england.nhs.uk/ourwork/clinical-policy/sepsis/nationalearlywarningscore/>.