Early warning score validation methodologies and performance metrics: A systematic review

Hao Sen Andrew Fang (andrew.fang.h.s@singhealth.com.sg)
SingHealth https://orcid.org/0000-0003-4761-9356
Wan Tin Lim
Singapore General Hospital
Balakrishnan Tharmmambal
Singapore General Hospital

Research article

Keywords: Early warning score, validation, methodology

Posted Date: April 30th, 2020

DOI: https://doi.org/10.21203/rs.2.16417/v2

Version of Record: A version of this preprint was published on June 18th, 2020. See the published version at https://doi.org/10.1186/s12911-020-01144-8.

Abstract

Background Early warning scores (EWS) have been developed as clinical prognostication tools to identify acutely deteriorating patients. With recent advancements in machine learning, there has been a proliferation of studies that describe the development and validation of novel EWS. Systematic reviews of published studies which focus on evaluating performance of both well-established and novel EWS have shown conflicting conclusions. A possible reason for this is the lack of consistency in the validation methods used. In this review, we aim to examine the methodologies and performance metrics used in studies which describe EWS validation.

Methods A systematic review of all eligible studies in the MEDLINE database from inception to 22-Feb-2019 was performed. Studies were eligible if they performed validation on at least one EWS and reported associations between EWS scores and mortality, intensive care unit (ICU) transfers, or cardiac arrest (CA) of adults within the inpatient setting. Two reviewers independently did a full-text review and performed data abstraction by using standardized data-worksheet based on the TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) checklist. Meta-analysis was not performed due to heterogeneity.

Results The key differences in validation methodologies identified were (1) validation population characteristics, (2) outcomes of interest, (3) case definition, intended time of use and aggregation methods, and (4) handling of missing values in the validation dataset. In terms of case definition, among the 34 eligible studies, 22 used the patient episode case definition while 10 used the observation set case definition, and 2 did the validation using both case definitions. Of those that used the patient episode case definition, 11 studies used a single point of time score to validate the EWS, most of which used the first recorded observation. There were also more than 10 different performance metrics reported among the studies.

Conclusions Methodologies and performance metrics used in studies performing validation on EWS were not consistent hence making it difficult to interpret and compare EWS performance. Standardizing EWS validation methodology and reporting can potentially address this issue.

Background

Early warning scores (EWS) are simple tools to help detect clinical deterioration to improve patient safety in hospitals. EWS are often implemented as part of a wider early warning system, also known as “rapid response system”, whereby detection of a likely deterioration will trigger an alert or pre-planned escalation of care by healthcare providers.[1-2] These EWS use objective parameters such as vital signs and laboratory results, and may include subjective parameters (e.g. “nurses’ worry”)[3] as input; and then output an integer score. A higher score generally indicates a higher likelihood of clinical deterioration, but is not a direct estimate of risk.

The first EWS was published in 1997,[4] and the concept gradually gained traction with the National Institute for Health and Clinical Excellence (NICE) recommending the use of early warning systems to monitor all adult patients in acute hospital setting in a 2007 guideline.[5] Currently there are many different EWS that have become available and are routinely used in hospitals globally, including USA, UK, Netherlands, Denmark and South Korea.[6-8] The recent advancements in machine learning (ML) have also opened up a new paradigm of novel EWS development, using ML techniques such as random forests and deep neural networks, giving rise to arguably better EWS.[9-11]

As EWS have an impact on patient care, it is critical that they are rigorously validated.[12] In this regard, several systematic reviews have already looked at the performance of various EWS.[6-8,13] Notably, these systematic reviews drew conflicting conclusions about EWS performance – Smith ME et al concluded that EWS perform well, while Gao et al and Smith GB et al found conflicting and unacceptable performance. A possible reason for this disagreement is a lack of consistency in the methods used to validate EWS.[8]

An example of difference in validation methods of EWS is between a study by researchers from Google and another study that validated the National Early Warning Score (NEWS).[14,15] Although both study teams validated their respective EWS ability to predict inpatient mortality, the former validated its EWS being used once, at the start of the admission (AUROC 0.93-0.94), while the latter validated their EWS using its score for every observation set of vitals measured for the entire admission (AUROC 0.89-0.90). In terms of case definition, we consider the former to have used the “patient episode” definition, while the latter used the “observation set” definition. Case definition is one of several differences in validation methods.

The aim of this review is to examine the different methodologies and performance metrics used in the validation of EWS so that readers will pay attention to specific aspects of the validation when making comparisons between EWS performance from published studies. As far as we are aware, there has not been any similar work done before. It is not this review’s intention to identify better performing EWS or to perform quality or bias assessment of the studies.

Methods

Search strategy

We used PubMed to search the MEDLINE database from inception to 22 Feb 2019 for studies of EWS in adult populations. We used the keywords “early warning score”, “predict”, “discriminate”, and excluded

"paediatrics", "children" and "systematic review". For completeness, we also sought to include publications that we found but were not returned in the PubMed search. This involved looking at studies from other EWS review papers [6-8, 13], and consulting with experts.

To assess the validation of EWS, we included only articles that performed validation on at least one EWS, in which investigators reported associations between EWS scores and inpatient mortality, intensive care unit (ICU) transfers, or cardiac arrest (CA). Systematic review papers were excluded as they lacked granularity in description of the data handling and statistical analysis. We also excluded studies which did prospective validation whereby the EWS was already in operation to influence care decisions and impact patient outcomes. In these cases, the validation did not purely evaluate the discriminative ability of the EWS, but also included other factors such as staff compliance and availability of rapid response resources.

Investigators then reviewed titles and abstracts of citations identified through literature searches, and eligible articles were selected for full-text review and data abstraction.

Data abstraction

Pre-defined data for abstraction was largely based on the TRIPOD Checklist for Prediction Model Validation,[16] with some added elements which the study team felt were pertinent to EWS.

A full-text review of each eligible study was performed by two investigators independently. Data for abstraction included the specific EWS validated, validation dataset used, number of subjects, population characteristics, outcome of interest (inpatient mortality, ICU transfer, cardiac arrest), method of validation (case definition, time of EWS use, type of aggregation for methods with multiple scores), method of handling missing values, and reported metrics. For discrepancies in the abstracted data, the investigators would perform a repeat review of the paper together to reach a consensus.

Results

The PubMed search yielded a list of 125 study abstracts. From reviewing the study abstracts, 47 studies were selected for full-text review (Figure 1). Of these, we excluded a further 12 studies – 11 (unable to access full study article) and 1 (full study article in Korean, only abstract in English). We included 13 additional relevant studies that we found from review papers and from consulting with experts. In total, 48 studies were included in the final review.[3, 9, 11, 15, 17-60]

A summary table of the selected studies and data abstracted are found in Table 1 [see Additional file 1]. 8 of the 48 studies were published in 2018 or later. Majority of the study populations were from UK (23) and USA (10), with 5 from South Korea and one each from Canada, China, Denmark, Hong Kong, Israel, Italy, Netherlands, Singapore, Sweden and Turkey.

Altogether, there were 54 unique EWS that were reviewed by the different studies, excluding the 33 other EWS assessed in the study by Smith in 2013,[15] and the 44 MET criteria assessed in the study by Smith in 2016.[28] The most reviewed EWS were the Modified Early Warning Score (MEWS) and National Early Warning Score (NEWS), which were included in 22 and 16 studies respectively.

Validation dataset

16 of the studies performed an internal validation, where a proportion of the entire study dataset was used to develop the EWS (training set), with the remaining proportion was used for validation (validation set).[9,11,18,19,21,23,25,29,33,35,36,39,40,42,45,47,58] Varying proportion sizes were used for the validation set ranging from 25.0% to 100%. The other studies did an external validation with a study population different from that used to develop the EWS.

The study size used ranged from 43 to 269,999. Slightly over half (25 of 48) of the studies were performed on general admission cases, with the others focused on populations with specific conditions (e.g. chorioamnionitis[49], community-acquired pneumonia[32,51,52]), patients admitted to a certain specialty (e.g. Obstetrics[40], Haematology[29]), or only a subset of the general admission population (e.g. those reviewed by MET[34,57], those with NEWS ≥ 1[22]).

Outcomes of interest

All the studies included at least one of the outcomes of: inpatient mortality, ICU transfer or cardiac arrest, or a combination of them (Figure 2).

For the 24 studies that evaluated more than one outcome, 17 studies validated EWS using a composite of all the outcomes as the endpoint, while the others validated EWS for each of the individual type of outcomes.

Case definition, time of EWS use and aggregation method

There were two different ways a case was defined – a patient episode or an observation set – and this definition had impact on the subsequent validation steps (Figure 3). The patient episode definition considered an entire admission as a single case and used either a single or multiple recordings of vital signs and other parameters from the admission, while the observation set definition considered each observation set of vital signs and other parameter recordings from the same admission as independent of one another.

In the observation set definition, recordings from each observation set would be used to compute a score to evaluate the EWS association with an outcome within a certain time window from the time of recording of the observation set. In the patient episode definition, because multiple recordings were available, study teams either chose to use a single score, or multiple scores to validate the EWS. This reflected how study teams intended for the EWS to be used in practice, so we termed this component as "time of EWS use". If the time of EWS use was multiple, then the series of scores would be aggregated in evaluating the EWS with the outcome of the patient episode.

Among the 48 studies, 34 used the patient episode method while 12 used the observation set method, and 2 did the validation using both methods.

Of the studies that used the patient episode method, 18 studies used a single point of time score to validate the EWS, most of which used the first recorded observation. For the 18 studies that used multiple recordings, 13 studies used use the maximum score as to aggregate the scores for each patient episode to compare with the outcomes. Most studies used all recordings from the patient episode, but there were 2 studies that excluded readings just prior to outcome to account for the "predictive" ability of the EWS.[29,45]

For the studies that used the observation set method, all of them used EWS values generated within the 0-24 hour time window prior to outcome to validate the EWS, with the exception of one study[11] that used EWS values within the 30 minute to 24 hour time window.

Handling of missing values

As the validation datasets were obtained from real-world data, missing values were inevitable . In general, there were two ways the missing values were handled – either exclude or impute values. 19 studies excluded the missing values, 15 used imputations, 4 used a combination of both methods, and in 10 studies it was not stated. There were a variety of imputation methods used. The most commonly used were imputing a value from the last observation (6 studies), imputing a median value (3 studies), a combination of the last observation and median (5 studies) and imputing with a normal value (4 studies). There were also some sophisticated imputation methods, such as using random forest[18] and multiple imputation[24].

Performance metrics

For performance metrics, we grouped them into two types – discrimination and calibration metrics. Discrimination is the measure of the EWS ability to differentiate significantly between cases with outcome and cases without outcome. Calibration is an evaluation of the extent to which estimated probabilities of the EWS scores agree with observed outcome rates.[61] In the case of EWS, the integer scores can be treated as bins to perform a probabilistic calibration evaluation.

The most commonly reported discrimination metric was the area under the curve of the receiver operating characteristics (AUROC). Only 6 studies did not use this metric. 22 studies reported using any one of sensitivity, specificity, or predictive value (positive and negative). A lesser used alternative to the AUROC was the area under the precision-recall curve (AUPRC) which was used in two more recent studies by Kwon et al[11] and Watkinson et al[21]. The authors reasoned that this is a more suitable metric for verifying false-alarm rates with varying sensitivity.

The EWS efficiency curve was another measure used to visualize the discriminatory ability of EWS in 8 studies. The EWS efficiency curve was first introduced in the study by Smith et al to provide a graphical depiction of the proportion of triggers that would be generated at varying EWS scores.[14]

Six studies performed a statistical test of model calibration. Four used the Hosmer-Lemeshow goodness-of-fit test, one calculated the calibration slope and one used both metrics. Another six studies did not perform a statistical test of calibration, but provided a visualization of the model calibration.

Discussion

Studies that validate EWS used a wide variety of validation methods and performance metrics. [4, 9, 11,15, 17-46] Given that these variations have a bearing on EWS performance measurement, one should be mindful of them when interpreting and comparing bottom-line metrics, like AUROC values.

While the TRIPOD checklist for prediction model validation provides a standardized framework for multivariable predictive model validation reporting,[16] it lacks the finer details for EWS which are more multi-faceted than typical prediction models. Unlike clinical prediction tools, EWS are unique in that they may be intended for use at multiple time-points over a patient episode. Some key differences in validation methodology we found in our review, and propose EWS evaluators take note of, are the validation dataset, outcomes of interest, case definition, time of EWS use, aggregation method, and handling of missing values. These differences could explain the reason for conflicting opinion on whether EWS perform well or otherwise.[6-8,13]

In terms of EWS performance reporting, our review also had similar findings from previous reviews, that studies tended to give more prominence to discrimination and have rarely assessed model calibration.[12,62] We concur with the TRIPOD recommendation that both discrimination and calibration should be considered when judging a model's accuracy.[16] Also, this review found some reporting metrics that can be considered as promising alternatives to AUROC in EWS performance reporting. The AUPRC mentioned earlier is one of them. It was noted to be suitable for verifying false-alarm rates with varying sensitivity.[22,23] Another would be to exclude measurements within a time window just prior to outcome, to account for the "predictive" ability of the EWS.[22,28,45]

We acknowledge that a limitation of our review may be the fairly narrow search strategy to include EWS studies with keywords "predict" and "discriminate", and thus might have unwittingly excluded other studies that performed EWS validation. Future reviews may consider broadening the scope of the initial search.

Conclusions

Current EWS validation methods are heterogenous and this probably contributes to conflicting conclusions regarding their ability to discriminate or predict the patients at risk of clinical deterioration. A standardized method of EWS validation and reporting can potentially address this issue.

List Of Abbreviations

APACHE Acute Physiology and Chronic Health Evaluation,

APPROVE	Accurate Prediction of Prolonged Ventilation Score
ASSIST	Assessment Score for Sick patient identification and Step-up in Treatment
AUPRC	Area under the Precision-Recall curve
AUROC	Area under the curve of the receiver operating characteristics
AWTTS	Aggregate weighted track and trigger systems
CA	Cardiac arrest
CARM	Computer aided risk of mortality
CART	Cardiac Risk Assessment Triage
cCEWS	Continuously-recorded centile-based EWS
eCART	Electronic Cardiac Arrest Risk Triage
ED	Emergency Department
EEC	EWS Efficiency curve
EWS	Early warning score
LDTEWS	Lab decision-tree early warning score
LEWS	Leed’s early warning score
GI-EWS	Gastrointestinal EWS
GMEWS	Global Modified EWS
LOCF	Last observation carried forward
H	Hour
HL	Hosmer-Lemeshow test
ICU	Intensive care unit
LR	Likelihood ratio
MACHP	Mean alarm count per patient per hour
mCEWS	Manually-recorded centile-based EWS

MEDS	Mortality in Emergency Department Sepsis
MET	Medical Emergency Team
MEOWS	Modified Early Obstetric Warning Score
MEWS	Modified Early Warning Score
ML	Machine learning
NEWS	National Early Warning Score
NICE	National Institute for Health and Clinical Excellence
NPV	Negative predictive value
NRI	Net reclassification index
OR	Odds ratio
PARS	Patient-At-Risk Score
PIRO	Predisposition/Infection/Response/Organ Dysfunction
PMEWS	Pandemic Medical Early Warning Score
PPV	Positive predictive value
PSI	Pneumonia Severity Index
qSOFA	Quick Sequential Related Organ Failure Assessment
ROC	Receiver operating characteristics
RAPS	Rapid Acute Physiology Score
REMS	Rapid Emergency Medicine Score
RRT	Rapid Response Team
SAPS	Simplified Acute Physiology Score
SCS	Simple Clinical Score
SCT	Stem cell transplant
SEWS	Standardized EWS

SIRS	Systemic Inflammatory Response Syndrome
SOFA	Sequential Organ Failure Assessment
TRIPOD	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
ViEWS	VitalPac EWS
YI	Youden's index

Declarations

Ethics approval: None sought as this was a systematic review of published studies.

Consent for publication: Not applicable as the paper does not contain any individual person's data.

Availability of data and materials: All data used in the publication of this work were obtained from published studies. The abstracts for these studies are available in the MEDLINE database found on PubMed.

Competing interests: The authors declare that they have no conflict of interest. The authors would like to highlight that they included a paper they wrote in this review.

Funding: The authors did not receive any funding for this work.

Authors' contributions: Dr Fang had the original idea for this work. Dr Fang and Dr Lim performed the review of the studies and data abstraction. Dr Thammambal provided research advice. Dr Fang wrote the first draft of this paper and all the authors subsequently assisted in redrafting and have approved the final version.

Acknowledgements: The authors would like to thank the Medical Informatics Office, SingHealth and Department of Internal Medicine, Singapore General Hospital for their support in this work. Also special

thanks also to Tay Wen Qing and Steve Tam Yew Chong from Education Resource Centre Office, Singapore General Hospital for their assistance in obtaining the full study articles.

References

1. DeVita MA, Hillman K. Why RRS? Where RRS?. Crit Care Clin 2018 Apr;34(2):xi-xii.
2. Alam N, Hobbelink EL, van Tienhoven AJ, van de Ven PM, Jansma EP, Nanayakkara PW. The impact of the use of the Early Warning Score (EWS) on patient outcomes: a systematic review. Resuscitation. 2014 May;85(5):587-94. doi: 10.1016
3. Douw G, Huisman-de Wal G, et al. Nurses' 'worry' as a predictor of deteriorating surgical ward patients: A prospective cohort study of Dutch-Early-Nurse-Worry-Indicator-Score. Intl Journal of Nursing Studies 2016; 134-140.
4. Morgan RJM, Williams F, Wright MM. An early warning scoring system for detecting developing critical illness. Clin Intensive Care 1997;8:100.
5. National Institute for Health and Clinical Excellence: Acute ill patients in hospital: recognition of and response to acute illness in adults in hospital. NICE clinical guideline No. 50. London; 2007.
6. 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.
7. Smith GB, Prytherch DR, Schmidt PE, et al. Review and performance evaluation of aggregate weighted 'track and trigger' systems. Resuscitation 2008;77:170–9.
8. Smith ME, 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.
9. Churpek MM, et al. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Crit Care Med. 2016 Feb; 44(2): 368-374.
10. Xu M, Tam B, et al. A protocol for developing early warning score models from vital signs data in hospitals using ensemble of decision tress. BMJ Open 2015;5:
11. Kwon JM, et al. An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest. J Am Heart Assoc. 2018 Jun 26;7(13). pii: e008678.
12. Gerry S, et al. Early warning scores for detecting deterioration in adult hospital patients: a systematic review protocol. BMJ Open 2017;7:e019268.
13. Smith MEB et al. Early Warning System Scores for Clinical Deterioration in Hospitalized Patients: A Systematic Review. Ann of the American Thoracic Society 2014.
14. Rajkomar A, et al. Scalable and accurate deep learning with electronic health records. Nature Digital Medicine 2018.
15. Smith GB, et al. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation

2013, Apr;84(4):465-70.

16. Collins GS, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. 2015.
17. Lim WT, et al. Use of the National Early Warning Score (NEWS) to Identify Acutely Deteriorating Patients with Sepsis in Acute Medical Ward. 2019. Ann Acad Med Singapore; 48:145-9.
18. Dziadzko MA, et al. Multicenter derivation and validation of an early warning score for acute respiratory failure or death in the hospital. Crit Care. 2018 Oct 30;22(1):286.
19. Faisal M, 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 Dec 6;8(12):e022939
20. Hydes TJ, et al. National Early Warning Score Accurately Discriminates the Risk of Serious Adverse Events in Patients With Liver Disease. Clin Gastroenterol Hepatol. 2018 Oct;16(10):1657-1666.e10.
21. Redfen OC, et al. Predicting in-hospital mortality and unanticipated admissions to the intensive care unit using routinely collected blood tests and vital signs: Development and validation of a multivariable model. 2018 Dec;133:75-81.
22. Spångfors M, et al. The National Early Warning Score predicts mortality in hospital ward patients with deviating vital signs: A retrospective medical record review study. J Clin Nurs. 2018 Dec 5.Watkinson PJ, et al. Manual centile-based early warning scores derived from statistical distributions of observational vital-sign data. 2018 Aug;129:55-60.Goulden R, et al. qSOFA, SIRS and NEWS for predicting inhospital mortality and ICU admission in emergency admissions treated as sepsis. Emerg Med J. 2018 Jun;35(6):345-349.
23. Kim WY, 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 Aug;40:213-217.
24. Tirotta D, 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). 2017 Jun 1;110(6):369-373.
25. Delgado-Hurtado JJ, et al. Emergency department Modified Early Warning Score association with admission, admission disposition, mortality, and length of stay. J Community Hosp Intern Med Perspect. 2016 Apr 25;6(2):31456.
26. Durusu Tanrıöver M, et al. Daily surveillance with early warning scores help predict hospital mortality in medical wards. Turk J Med Sci. 2016 Dec 20;46(6):1786-1791.
27. Hu SB, et al. Prediction of Clinical Deterioration in Hospitalized Adult Patients with Hematologic Malignancies Using a Neural Network Model. PLoS One. 2016 Aug 17;11(8):e0161401.
28. Kovacs C, et al. Comparison of the National Early Warning Score in non-elective medical and surgical patients. Br J Surg. 2016 Sep;103(10):1385-93.
29. Smith GB, 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 Dec;44(12):2171-2181.Jo S, et al. 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 Mar;34(3):536-41.

30. Liu FY, et al. A prospective validation of National Early Warning Score in emergency intensive care unit patients at Beijing. 2015. Hong Kong Journal of Emergency Medicine. Vol. 22(3): 137-144.
31. Yoo JW, et al. A combination of early warning score and lactate to predict intensive care unit transfer of inpatients with severe sepsis/septic shock. Korean J Intern Med. 2015 Jul;30(4):471-7.
32. Churpek MM, et al. Multicenter development and validation of a risk stratification tool for ward patients. Am J Respir Crit Care Med. 2014 Sep 15;190(6):649-55.
33. Churpek MM, et al. Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards. Crit Care Med. 2014 Apr;42(4):841-8.
34. Kim WY, et al. Modified Early Warning Score Changes Prior to Cardiac Arrest in General Wards. PLoS One. 2015 Jun 22;10(6):e0130523. Yu S, et al. Comparison of risk prediction scoring systems for ward patients: a retrospective nested case-control study. Crit Care. 2014 Jun 26;18(3):R132.
35. Badriyah T, et al. Decision-tree early warning score (DTEWS) validates the design of the National Early Warning Score (NEWS). 2014 Mar;85(3):418-23.
36. Carle C, et al. Design and internal validation of an obstetric early warning score: secondary analysis of the Intensive Care National Audit and Research Centre Case Mix Programme database. 2013 Apr;68(4):354-67.
37. Corfield AR, et al. Utility of a single early warning score in patients with sepsis in the emergency department. Emerg Med J 2013;0:1-6.
38. Jarvis SW, et al. Development and validation of a decision tree early warning score based on routine laboratory test results for the discrimination of hospital mortality in emergency medical admissions. 2013 Nov;84(11):1494-9.
39. Romero-Brufau S, et al. Widely used track and trigger scores: are they ready for automation in practice? 2014 Apr;85(4):549-52. Alrawi YA, et al. Predictors of early mortality among hospitalized nursing home residents. QJM. 2013 Jan;106(1):51-7.
40. Churpek MM, et al. Derivation of a cardiac arrest prediction model using ward vital signs. Crit Care Med. 2012 Jul;40(7):2102-8.
41. Cooksley T, et al. Effectiveness of Modified Early Warning Score in predicting outcomes in oncology patients.2012 Nov;105(11):1083-8.
42. Kellet J, et al. Changes and their prognostic implications in the abbreviated Vitalpac™ early warning score (ViEWS) after admission to hospital of 18,853 acutely ill medical patients. 2013 Jan;84(1):13-20.Ghanem-Zoubi NO, et al. Assessment of disease-severity scoring systems for patients with sepsis in general internal medicine departments. Crit Care. 2011;15(2):R95.
43. Lappen JR. Existing models fail to predict sepsis in an obstetric population with intrauterine infection. Am J Obstet Gynecol. 2010 Dec;203(6):573.e1-5.

44. Prytherch DR, et al. ViEWS – Towards a national early warning score for detecting adult inpatient deterioration. Resuscitation 81 (2010) 932-937.
45. Barlow G, et al. The CURB65 pneumonia severity score outperforms generic sepsis and early warning scores in predicting mortality in community-acquired pneumonia. Thorax 2007;62:253-259.
46. Challen K, et al. Physiological-social score (PMEWS) vs. CURB-65 to triage pandemic influenza: a comparative validation study using community-acquired pneumonia as a proxy. BMC Health Services Research 2007,7:33.
47. von Lilienfeld-Toal M, et al. Observation-Based Early Warning Scores to Detect Impending Critical Illness Predict In-Hospital and Overall Survival in Patients Undergoing Allogeneic Stem Cell Transplant. Biology of Blood and Marrow Transplantation 13:568-576(2007).
48. Kellet J, et al. The Simple Clinical Score predicts mortality for 30days after admission to an acute medical unit. Q J Med 2006; 99:771-781.
49. Lam TS, et al. Validation of a Modified Early Warning Score (MEWS) in emergency department observation ward patients. Hong Kong Journal of Emergency Medicine. 2006;13:24-30.
50. Subbe CP, et al. Validation of physiological scoring systems in the accident and emergency department. Emerg Med J 2006;23:841-845.
51. Goldhill DR, et al. A physiologically-based early warning score for ward patients: the association between score and outcome. Anaesthesia, 2005, 60:547-553.
52. Olsson T, et al. Rapid Emergency Medicine score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. Journal of Internal Medicine 2004; 255:579-587.
53. Hodgetts TJ, et al. The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team. Resuscitation 54 (2002) 125-131.
54. Subbe CP, et al. Validation of a modified Early Warning Score in medical admissions. Q J Med 2001; 94:521-526.
55. Steyerberg E, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010 Jan; 21(1): 128-138.
56. Van Calster, B., et al.Calibration: the Achilles heel of predictive analytics. BMC Med 17, 230 (2019).

Figures

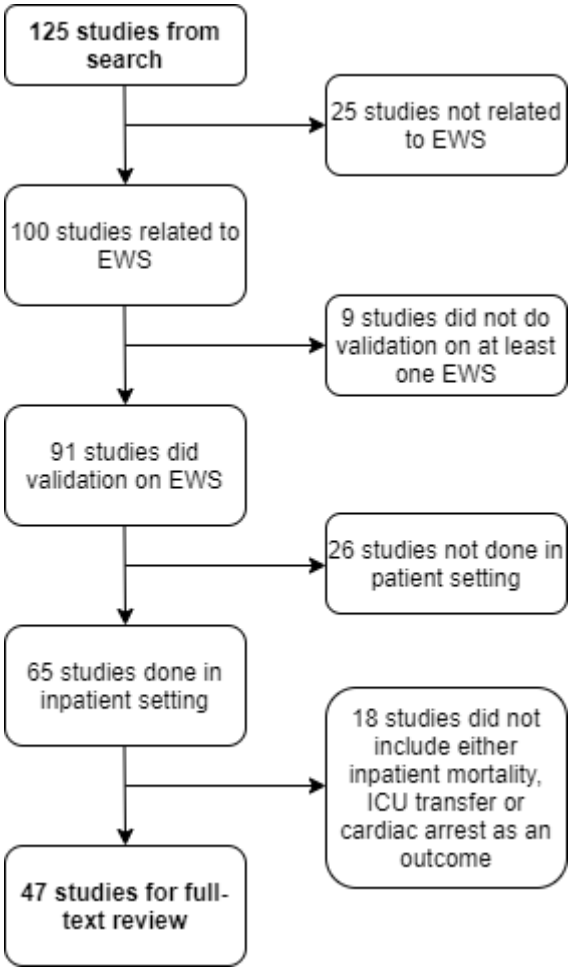


Figure 1

Flow chart describing inclusion of articles for full-text review from search result list

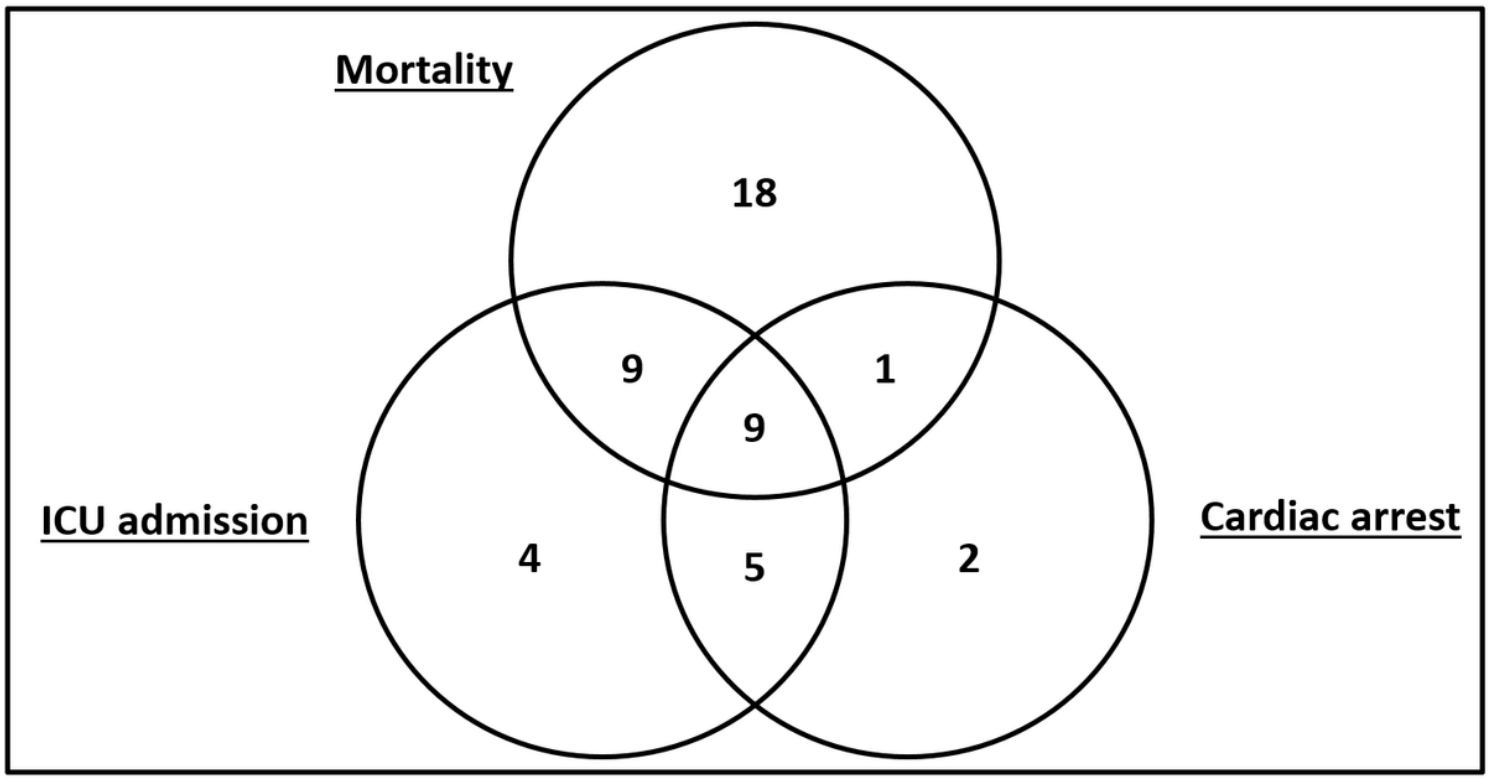
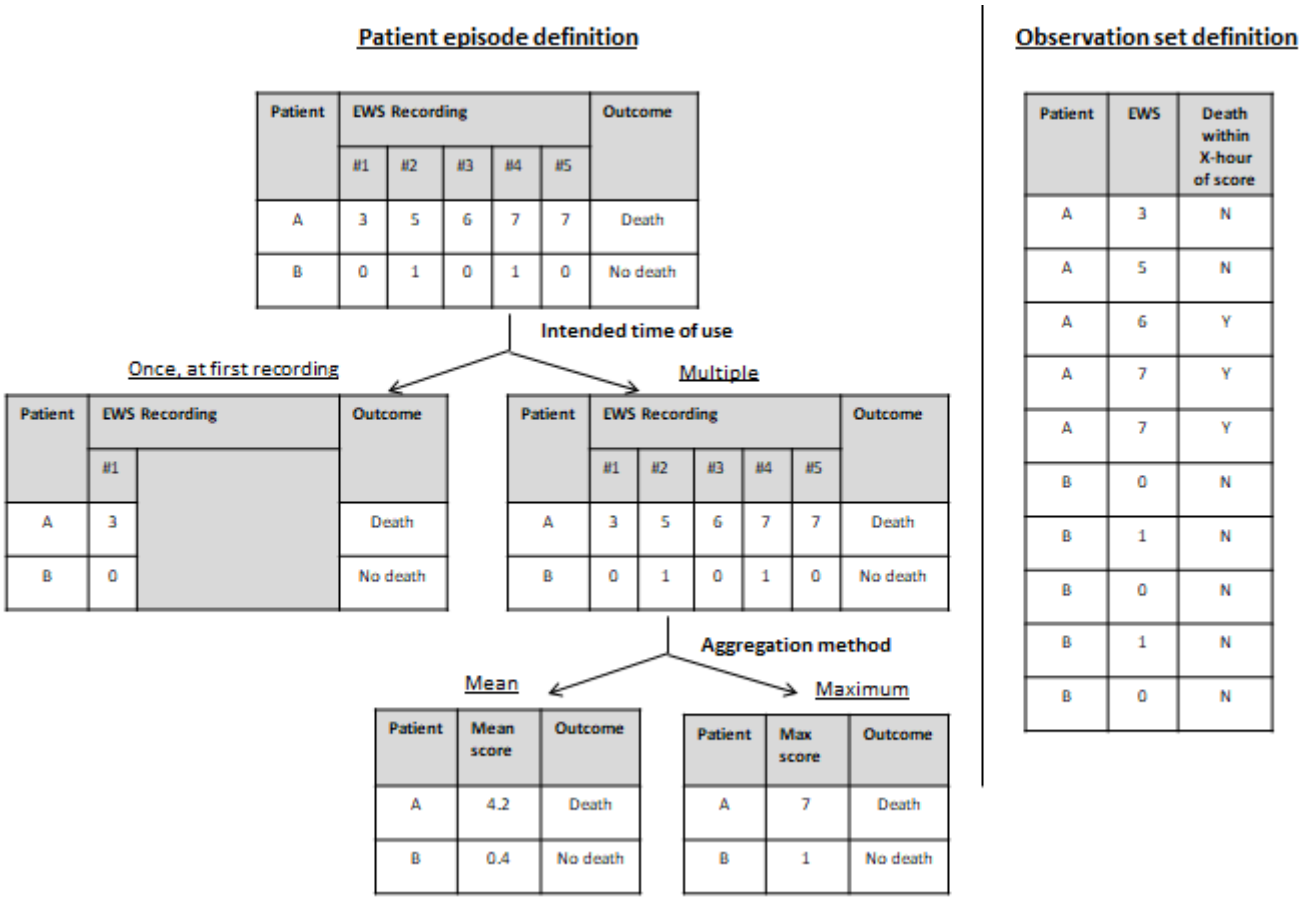


Figure 2

Summary of studies with various combinations of outcomes.



Patient	EWS Recording					Outcome
	#1	#2	#3	#4	#5	
A	3	5	6	7	7	Death
B	0	1	0	1	0	No death

Patient	EWS	Death within X-hour of score
A	3	N
A	5	N
A	6	Y
A	7	Y
A	7	Y
B	0	N
B	1	N
B	0	N
B	1	N
B	0	N

Patient	EWS Recording		Outcome
	#1		
A	3		Death
B	0		No death

Patient	EWS Recording					Outcome
	#1	#2	#3	#4	#5	
A	3	5	6	7	7	Death
B	0	1	0	1	0	No death

Patient	Mean score	Outcome
A	4.2	Death
B	0.4	No death

Patient	Max score	Outcome
A	7	Death
B	1	No death

Figure 3

Illustration of how different case definition affect EWS validation for 2 patients

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1Version220200418.docx