The optimal threshold for prompt clinical review : an external validation study of the national early warning score

THE OPTIMAL THRESHOLD FOR PROMPT CLINICAL REVIEW: AN EXTERNAL VALIDATION STUDY OF THE NATIONAL EARLY WARNING SCORE.

RUNNING TITLE: Optimal threshold of NEWS for prompt review.

AUTHORS

Filip HAEGDORENS, Koenraad G. MONSIEURS, Koen DE MEESTER, Peter VAN BOGAERT.

AUTHORS AFFILIATIONS AND DEGREES

Centre for Research and Innovation in Care (CRIC), Department of Nursing and Midwifery Sciences, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium (F HAEGDORENS BN, MScN);

Department of emergency medicine, Antwerp University Hospital, University of Antwerp, Wilrijkstraat 10, 2650 Edegem, Belgium (KG MONSIEURS MD, PhD);

Centre for Research and Innovation in Care (CRIC), Department of Nursing and Midwifery Sciences, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium (K DE MEESTER BN, PhD);

Centre for Research and Innovation in Care (CRIC), Department of Nursing and Midwifery Sciences, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium (P VAN BOGAERT BN, PhD);

CORRESPONDING AUTHOR

Filip HAEGDORENS, Centre for Research and Innovation in Care (CRIC), Department of Nursing and Midwifery Sciences, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium

email - filip.haegdorens@uantwerp.be

phone - 0032 3 265 91 69

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jocn.15493

twitter - @filhgd

ACKNOWLEDGEMENTS

We would like to thank the department of quality of care and patient safety of the Federal Public Service of Health, Food chain safety and Environment of Belgium.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

FUNDING

The Belgian federal government sponsored this study but had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The researchers assume final responsibility.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Approval of the ethics committees of all participating hospitals was obtained beforehand (registration number: B300201317835).

AUTHORS' CONTRIBUTIONS

FH, PVB, KDM, and KM made substantial contributions to the conception and design, or acquisition of data, or analysis and interpretation of data; FH, PVB, and KM were involved in drafting the manuscript or revising it critically for important intellectual content; all authors have given final approval of the version to be published.

DR. FILIP HAEGDORENS (Orcid ID : 0000-0003-1996-8786)

Article type : Original Article

ABSTRACT

AIM

The aim of this study was to determine the optimal threshold for NEWS in clinical practice.

BACKGROUND

The national early warning score (NEWS) is an aggregate early warning score aiming to predict patient mortality. Studies validating NEWS did not use standardised patient outcomes or did not always include clinical workload in their results. Since all patients with a positive NEWS require a clinical workup, it is crucial to determine the optimal threshold to limit false-positive alerts.

DESIGN

External validation study using retrospectively collected data of patient admissions in six Belgian hospitals.

METHODS

We adhered to the STARD guideline for reporting. Two sample groups were selected: the cross-sectional sample (admitted patients, one day every four months) and the serious adverse event (SAE) sample (all patients with unexpected death, cardiac arrest and unplanned admission to the intensive care unit). The maximum registered NEWS value was collected in both groups. Predictive values (PPVs) were used as estimates for clinical workload.

RESULTS

We collected 1523 in the cross-sectional sample and 390 patients in the SAE sample. A NEWS ≥5 had a PPV of 6.8% and a negative predictive value of 99.5% to predict unexpected death, cardiac arrest with cardiopulmonary resuscitation or unplanned admission to intensive care (AUROC 0.841). The performance of NEWS differed between outcome measures. Considering the PPV, the optimal threshold for NEWS is ≥5.

CONCLUSIONS

We validated NEWS to be applied in general hospital wards and confirmed the optimal threshold (≥5).

RELEVANCE TO CLINICAL PRACTICE

When a patient has a NEWS <5, we may assume that in the next 24 hours this patient is less likely to die unexpectedly, receive CPR or be transferred to the ICU. Because of the significant number of false positives when NEWS is ≥5, hospitals should create workable guidelines for clinical practice.

KEYWORDS
validation
early warning score
burden
workload
sensitivity
specificity
positive predictive value
deterioration
mortality

INTRODUCTION

The national early warning score (NEWS) is a widely adopted instrument mostly used by nurses and physicians that estimates the risk of deterioration in hospitalised patients using physiological observations (Smith et al., 2013). It was originally developed and validated by Prytherch and colleagues using a large vital signs database with the aim to detect patient mortality within a 24-hour timeframe (Prytherch et al., 2010). The NEWS comprises respiratory rate, oxygen saturation, administration of supplemental oxygen, temperature, systolic blood pressure, heart rate and the level of consciousness. The main purpose of early warning scores (EWSs) is to screen patients admitted to hospital wards to detect physiological deterioration resulting in a timely and appropriate medical response. Therefore, it is essential to study the predictive performance of EWSs on patient outcomes.

BACKGROUND

Smith et al. showed that the NEWS had the greatest ability to discriminate patients at risk for deterioration compared with 33 other EWSs (Smith et al., 2013). External validation of the NEWS was carried out in multiple studies seemingly confirming its predictive quality in relation to patient outcomes (Spångfors et al., 2019; Mitsunaga et al., 2019; Pimentel et al., 2019; Lee et al., 2018). However, these validation studies have some important limitations.

Firstly, no uniform patient outcomes were used in these studies. Patient deterioration is defined in numerous ways in the EWS literature (Smith et al., 2014). Surrogate measures for deterioration include intensive care unit (ICU) admissions and rapid response team (RRT) calls. The value of these measures is contested as ICU admissions and/or RRT calls are to be expected when patients become unstable and they do not always imply early and efficient detection. The outcomes most frequently studied in previous research are patient mortality, cardiopulmonary resuscitation (CPR) and unplanned ICU admission (Maharaj et al., 2015). As we argued in previous papers, these outcome measures are prone to bias and we therefore proposed unambiguous definitions to be used when studying interventions to prevent deterioration (Haegdorens et al., 2018; Haegdorens et al., 2019)

Secondly, validation studies typically use classic metrics to determine the predictive performance of EWSs such as sensitivity, specificity, likelihood ratios, c-statistics, and area under the receiver operator characteristic curve measures (AUROCs).

Most studies, however, do not take into account the prevalence of the outcomes being measured (Romero-Brufau et al., 2015). We established earlier that the prevalence of clinically relevant serious adverse events in hospital patients admitted to a general ward is much lower than anticipated (Haegdorens et al., 2018). Moreover, outcome measures are not standardised and differ between studies concerning EWSs. Performance metrics are dependent on the outcomes used for validation (e.g. total mortality or cardiac arrest) and could therefore over- or underestimate the usefulness of EWSs. Ideally, the benefit of the EWS (early detection or sensitivity) should be compared with the clinical workload (false positives) it causes. A useful metric to estimate the clinical workload is the positive predictive value (PPV) since it shows the proportion of patients with a positive EWS that eventually will reach the outcome (Trevethan, 2017). Since all patients with a positive EWS require a clinical workup, it is important for hospitals to determine the optimal threshold for clinical review to limit false-positive alerts and subsequently alarm-fatigue.

The aim of this study is to determine the optimal threshold for clinical review of the national early warning score, using a database including patient admissions in six Belgian hospitals, comparing several outcome measures.

METHODS

Study design and participants

In this external validation study, we used retrospectively collected data of 32,722 patient admissions in 24 wards of six Belgian acute hospitals (two surgical and two medical wards per hospital) from February 2014 until May 2015. The original study was submitted to a clinical trial registry (clinicaltrials.gov identifier: NCT01949025) and was approved by ethics committees of all participating hospitals (registration number: B300201317835). The Belgian federal government sponsored this study but had no role in study design, data collection, data analysis, data interpretation, or writing of the report. This study was carried out following STARD 2015 reporting guidelines for diagnostic accuracy studies (supplementary file 1). Non-pregnant patients, older than 16 years, admitted to one of the study wards were consecutively included in the main sample. Two sample groups were selected from the database: the cross-sectional sample and the serious adverse event (SAE) sample. The cross-sectional sample comprised all admitted patients to a study ward in a 24-h timeframe

(between 00:00 and 23:59) on a randomly chosen Tuesday in each four-month period (four times in total during this study). Patients included in the cross-sectional sample did not experience an SAE. Patients who experienced an SAE during their admission were excluded from the cross-sectional sample included in the SAE sample. The SAE sample included all admitted patients to a study ward experiencing an SAE. Patients in the SAE sample were studied in the 24 hours prior to the SAE.

Outcomes

Five different outcome measures and two composite outcomes were studied. Patients were added to the SAE sample in case of unexpected death (a), cardiac arrest with CPR (b) or unplanned ICU admission (c). These outcome measures were defined extensively in a previous publication (Haegdorens et al., 2018). The before mentioned outcomes were added together in the first composite outcome (a+b+c) that represented an undesirable outcome for the patient. It is important to emphasize that this outcome not only contains patient sudden mortality (unexpected death), but also includes the survivors and deceased after cardiac arrest and unplanned ICU admission. Therefore, we added two more outcome measures: death within 72-h after cardiac arrest with CPR (e) and death within 72-h after unplanned ICU admission (f). The second composite outcome measure (a+e+f) served as an estimation for the total undesirable patient mortality and comprised unexpected death, death within 72-h after cardiac arrest with CPR and death within 72-h after unplanned ICU admission. The rationale to follow-up patients' mortality until 72-h after cardiac arrest with CPR or after unplanned ICU admission was to be certain that a patient's death was close enough to the last registered NEWS.

Measures

The index test used in this study was the NEWS as defined by the Royal College of Physicians (RCOP, 2012). The reference standards were the different outcome measures used in this study as defined previously. We used the cut-off measures as pre-specified by the NEWS guideline where a NEWS higher than 4 (medium score) should result in a prompt clinical review of the patient. Furthermore, a NEWS of 7 or more (high score) requires urgent emergency assessment by a critical care team

and could initiate a transfer of the patient to a higher dependency care area. The NEWS was used and registered by nurses at the patient's bedside. Since the NEWS was used prospectively by nurses working at the study wards, they had no prior knowledge concerning the patient outcome when registering the score. To ensure that the NEWS scores used in this study were without errors, all collected scores were checked and corrected by a researcher (F.H.) when necessary using vital sign data. The researchers used a standardised electronic checklist to collect data from patient records. Of all recorded data, only the maximum NEWS value in a 24-h period was retained. Thus, the maximum NEWS up to 24-h before an SAE or in the 24-h cross-sectional sample was used to evaluate the instrument's predictive performance. Data were considered missing if no full set of vital signs was available during the 24-h period.

Statistical analysis

All data were analysed using IBM SPSS Statistics version 25 for MAC OS. Two-sided Fisher's Exact tests were used to compare the proportion of reached outcomes between positive and negative NEWS scores in the SAE sample. Pearson's Chi-Squared tests were used to compare all proportions of characteristics between the cross-sectional and SAE sample. We used the independent t-test to compare patients' age between groups and we chose Mann-Whitney U tests to compare NEWS values and subscores since they had an ordinal level of measurement.

To analyse the screening accuracy of the NEWS we calculated the sensitivity, specificity, positive and negative likelihood ratios and AUROCs. Optimal NEWS cut-off scores were determined using the Youden's J statistic (Fluss et al., 2005). Because this study was based on a retrospectively collected dataset comprising two samples (cross-sectional and SAE), it was not possible to directly calculate screening accuracy statistics integrating the disease prevalence rate (i.e. the positive predictive value (PPV) and negative predictive value (NPV)). The sensitivity and specificity of a specific NEWS threshold are measures of intrinsic accuracy and are the same in prospective or retrospective studies. However, because NEWS aims to screen large groups of hospitalised patients in order to predict rare outcomes (e.g. sudden patient death), it seems meaningful to determine its value in clinical practice using PPVs and NPVs (Lutgendorf & Stoll, 2016). To calculate the PPVs and NPVs in this study, we applied the method proposed by Mercaldo and colleagues and

calculated adjusted estimates of screening accuracy and their corresponding confidence intervals (Mercaldo et al., 2007). We applied Bayes' theorem and used the pre-test probability together with the sensitivity and specificity to calculate the positive predictive values (PPVs) and negative predictive values (NPVs) (Grunau & Linn, 2018). The proportion of positive scores were plotted to estimate the clinical workload for nurses and physicians for each NEWS at or above a given value.

RESULTS

Of all admitted patients (n=32722), 1523 were assigned to the cross-sectional sample and 390 were assigned to the SAE sample after excluding missing data (figure 1). The prevalence of all studied outcomes was compared between the positive NEWS score group (i.e. NEWS ≥5) and the negative NEWS score group (i.e. NEWS <5). The proportion of patients with a positive NEWS score was 11% in the cross-sectional sample and 66% in the SAE sample. In the SAE sample, the proportion of patients who experienced unexpected death or unplanned ICU admission was not significantly higher in the positive NEWS score group compared to the negative NEWS score group (unexpected death 5.4% vs 3.8%, p=0.621; unplanned ICU admission 88.7% vs 82.7%, p=0.116). However, the proportion of patients with cardiac arrest and CPR was significantly lower in the positive NEWS score group compared to the negative NEWS score group (5.8% vs 13.5%, p=0.012). Nonetheless, the proportion of patients who died after CPR was not statistically different between groups (4.3% vs 6.0%, p=0.464). The proportion of death after unplanned ICU admission was significantly higher in the positive NEWS score group compared to the negative NEWS score group (9.4% vs 2.3%, p=0.010). And finally, the undesirable patient mortality (composite outcome a+e+f) was not significantly higher in the positive NEWS score group compared to the negative NEWS score group (19.1% vs 12.0%, p=0.086).

Figure 1. Flow chart

Patients in the SAE sample were significantly older than in the cross-sectional sample (table 1). The mean maximum NEWS was significantly higher in the SAE sample (mean difference: +3.69) and above the assigned cut-off score by the Royal College of Physicians (RCOP, 2012). In the SAE sample, significantly more patients had NEWS scores of 5-6 and at or above 7 compared with the other sample (5-6: 24% vs 8% and ≥7: 42% vs 3%, p<0.001). We compared the NEWS vital signs subscores between the two samples. The available range in which these subscores can vary is between zero and three. All NEWS subscores were significantly higher in the SAE sample with the exception of consciousness (AVPU) and temperature. The highest and lowest divergent vital signs in the SAE sample were oxygen saturation

and temperature, respectively. The mean difference between the cross-sectional sample and the SAE sample subscores, was the highest in oxygen saturation (mean difference +1.06) and the lowest in systolic blood pressure (mean difference +0.41).

Table 1. Comparison of characteristics between the cross-sectional sample and the SAE sample.

The screening accuracy of a NEWS ≥5 was evaluated in table 2 comparing two different outcomes. Using NEWS to predict composite outcome (a+b+c) was less sensitive but more specific when compared with composite outcome (a+e+f). The precision of the NEWS was higher to detect composite outcome (a+b+c) compared with composite outcome (a+e+f) (PPV 6.76% vs 0.77%). We may assume that of all patients with a NEWS ≥5, 6.76% would die unexpectedly, experience a cardiac arrest or would be transferred to the ICU. More importantly, the negative predictive values were very high for both outcomes. This implies that the probability of reaching one of the outcomes for patients with a negative screening test (NEWS <5) is very low (i.e. negative post-test probability composite outcome (a+b+c): 0.48% and (a+e+f): 0.07%).

Table 2. Confusion matrices and measures for diagnostic accuracy for the two composite outcomes in this study.

Receiver operator curves were plotted for both composite outcomes including Youden's J statistics (figure 2). Both AUROCs were higher than 0.800 but the AUROC of composite outcome (a+b+c) was higher than the AUROC of composite outcome (a+e+f) (0.841 vs 0.815). According to Youden's J statistics, the statistically optimal threshold was lower for composite outcome (a+b+c) than for composite outcome (a+e+f). The highest and lowest AUROCs were found for death after unplanned ICU admission (AUROC: 0.885) and CPR (AUROC: 0.716), respectively (table 3). The optimal NEWS threshold in the three outcomes with the highest discriminatory value was ≥4.

Figure 2. Receiver operator curves.

AUROC: area under the receiver operator curve; J: Youden's J statistic; NEWS: national early warning score

Table 3. Comparing area under the curve measures between different outcomes of the National Early Warning Score.

Lastly, in figure 3, we plotted the percentage of positive scores for each NEWS (estimating the total clinical workload) and the corresponding PPV for the two outcomes studied at or above a given value. The aim of this figure is to compare instrument accuracy and workload between NEWS thresholds. A significant proportion of patients (77.4%) had a NEWS greater than or equal to one. The proportion of patients with a NEWS ≥4, ≥5 and ≥7 were 20.1%, 11.3% and 3.2%, respectively. Additionally, PPVs for both composite outcome measures were plotted. The PPV for composite outcome (a+b+c) with a NEWS ≥4 equalled 4.44%. The PPV and NPV differences between NEWS ≥4 and NEWS ≥5 were +2.32% and -0.16%, respectively. PPVs for the composite outcome (a+e+f) were very low across all NEWS values (range 0.0-1.9%).

Figure 3. Percentage of positive scores and PPV for each NEWS at or above a given value

PPV: Positive predictive value; NPV: Negative predictive value;
J: Youden's J statistic; NEWS: national early warning score

DISCUSSION

Before adopting a new diagnostic instrument into clinical practice, it is essential to evaluate its technical characteristics first. This is classically done by building a confusion matrix and by calculating various diagnostic statistics (e.g. sensitivity and specificity) which allow clinicians to evaluate its accuracy and precision. However, different approaches to calculate these statistics exist, yielding very different results (Grunau & Linn, 2018). Firstly, sensitivity and specificity do not provide enough information to determine the practical usefulness of a particular test. PPVs and NPVs are more appropriate measures to determine if a test is effective in categorising patients as having or not having a condition (Trevethan, 2017). We calculated adjusted PPVs and NPVs using the pre-test probability of outcomes in the total sample (n = 32722).

We found that the PPV of the NEWS to predict the most mentioned outcome (a+b+c) in the EWS literature was rather low. Because the PPV is a function of the disease prevalence, the PPV of the second composite outcome (a+e+f) was even lower. Low or moderate PPVs could be acceptable if the negative consequences of false positives are limited. The NEWS is an easily performed screening test designed to be used in a heterogeneous population to detect very serious and possibly avoidable conditions. Furthermore, all positive scores require a prompt clinical assessment and follow-up to determine if an intervention is necessary. The clinical burden after screening a patient with a false positive NEWS score may be considered a negative consequence of the application of the NEWS score. This burden comprises the time spent by clinicians to evaluate the patient's condition, the cost and risk of additional investigations, and the possible discomfort experienced by the patient. However, we hypothesise that clinicians function as a barrier between a false positive score and potential harmful or costly investigations.

The NPVs for both outcomes were very high. This is desirable in case of the NEWS because the outcome being predicted has massive consequences for the patient and should be avoided. If a hospitalised patient has a NEWS <5, we may safely assume that in the next 24 hours this patient is not going to die unexpectedly, receive CPR or be transferred to the ICU.

The AUROCs calculated for the different outcomes in this study varied significantly (range 0.716-0.885). Studies comparing the performance of different EWSs yielded diverse results (Linnen et al., 2019). Typically, studies reporting a comparison between an externally developed EWS and an in-house derived score report higher c-statistics in the latter. However, this does not always imply that the in-house derived score is superior. We showed that AUROCs depend on the predicted outcome and since no standardised outcome set is used in the literature, AUROCs cannot be compared between studies. Furthermore, in-house developed aggregate weight or statistical modelling scores will always perform technically better than externally developed scores (e.g. the NEWS) because of their derivation from the score development dataset which depends on a specific population and time period. Additionally, scores developed using machine learning could introduce other practical problems (e.g. recalibration, difficult interpretability of the output, possible deskilling of clinicians, and loss of context) (Stead, 2018; Cabitza et al., 2017)

The strength of the NEWS is that it is an easy to use decision support system that can be applied in clinical practice without the need for complex datasets or computational power. We externally validated the NEWS in this study and can confirm that its technical characteristics are sufficient to determine patients' risk for serious adverse events. Furthermore, in our population we showed that the NEWS is best at predicting death after unplanned ICU admission and worst in predicting CPR. More patients received CPR and less patients died after ICU admission in the negative score group (NEWS <5). We hypothesise that in patients with a negative score, deterioration was not clinically visible or sudden and therefore not detectable timely, resulting in CPR. Moreover, if patients with a NEWS <5 slowly deteriorated, their illness could be less severe resulting in less deaths after ICU admission compared with positive scores.

Depending on the outcome studied, different optimal thresholds were calculated. The statistically optimal threshold for the composite outcome (a+b+c) was ≥4 which is lower than mentioned in guidelines (Smith et al., 2013; RCOP, 2012). We noticed that if we would use a NEWS ≥5, this would increase the PPV significantly and the loss of NPV would be acceptable. To decide which threshold is optimal in clinical practice, it is important to consider the clinical workload that it implies. The clinical

workload when NEWS is above a certain threshold can be subdivided in the clinical follow-up (evaluation and treatment) and the increase in observation frequency. Nurses are concerned about the workload associated with an increased frequency of taking vital signs, particularly in patients with chronic diseases with deviating vital signs (Jensen et al., 2019). Hospitals should carefully consider the increase in the observation frequency (NEWS ≥1: once every six hours, NEWS ≥5: once every hour) to create a workable and achievable guideline. The added value of NEWS in practice is that it aims to improve patient safety. However, it is well known that if nurses' workload is increased, rationing of nursing care will occur (Griffiths et al., 2018). The way NEWS is implemented in practice should be balanced between detection accuracy, the associated clinical burden and the benefit for the patient (Romero-Brufau et al., 2015). We confirm using a NEWS cut-off ≥5 to initiate a prompt clinical response to evaluate the patient's condition. The associated increase of observation frequency should be revised at that time to prevent unworkable situations for nurses on the ward. Ideally, this becomes standardised practice in the evaluation of the patient.

We compared NEWS subscores of the vital signs between the two groups and found that the most deviating vital signs preceding an SAE were the blood oxygen saturation, systolic blood pressure, and heart rate. This contradicts previous research stating that the respiratory rate is the most important predictor for clinical deterioration (Rolfe, 2019; Cretikos et al., 2008). Nonetheless, it is the combination of different vital signs that makes the NEWS a useful predictor.

The most important limitations in this study were: the use of an existing database and the possibly biased selection of the cross-sectional sample from the main sample. Moreover, our population was confined to surgical and medical wards in Belgian hospitals which limits the transferability of the results. Furthermore, we did not use the latest NEWS updated by the Royal College of Physicians in 2017 (NEWS-2) that includes a new scoring system for patients with type II respiratory failure (RCOP, 2017). However, NEWS-2 showed no improvements in technical capabilities and is more complicated to use (Pimentel et al., 2019)

CONCLUSION

We validated the NEWS in surgical and medical wards in acute hospitals confirming the optimal news threshold of ≥5. When a patient has a NEWS <5, we may safely assume that in the next 24 hours this patient is less likely to die unexpectedly, receive CPR or be transferred to the ICU.

RELEVANCE TO CLINICAL PRACTICE

The follow-up of patients with a NEWS ≥5 is justifiable when considering clinical workload. However, hospitals should carefully consider how to achieve workable guidelines for clinical practice since NEWS ≥5 generates a significant number of false positives.

The NEWS is a widely adopted track-and-trigger system aiming to support nurses and physicians in detecting and responding to deterioration in hospitalised patients and has been validated in previous research (Churpek et al., 2017; Spångfors et al., 2019; Lane et al., 2019). NEWS is a valuable instrument assisting nurses in the clinical evaluation of their patients in order to improve practice. However, we discovered in this study that NEWS has a significant number of false positives. International guidelines do not always take into account the associated workload when implementing NEWS in practice (RCOP, 2017). Previous research reported poor compliance rates when implementing EWS protocols (Considine et al., 2016). We hypothesise that the combination of false positives and the associated workload could possibly affect the face validity of NEWS resulting in poor adoption by nurses and physicians. We advise hospitals to provide sufficient and correct information to clinicians concerning the accuracy and precision of NEWS in detecting patient mortality. NEWS works best for ruling out the possibility of a patient's death and should be applied in practice bearing this in mind. An adequate, but also efficient and workable response strategy should be thought out for each setting. Future research should not only focus on the effects of EWSs on patient outcomes but should also investigate which guidelines are workable in clinical practice.

REFERENCES

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended Consequences of Machine Learning in Medicine. *JAMA*, *318*(6), 517-518. https://doi.org/10.1001/jama.2017.7797

Churpek, M. M., Snyder, A., Han, X., Sokol, S., Pettit, N., Howell, M. D., & Edelson, D. P. (2017). 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. *American Journal of Respiratory and Critical Care Medicine*, *195*(7), 906-911. https://doi.org/10.1164/rccm.201604-0854oc

Considine, J., Jones, D., Pilcher, D., & Currey, J. (2016). Patient physiological status at the emergency department-ward interface and emergency calls for clinical deterioration during early hospital admission. *J Adv Nurs*, *72*(6), 1287-1300. https://doi.org/10.1111/jan.12922

Cretikos, M. A., Bellomo, R., Hillman, K., Chen, J., Finfer, S., & Flabouris, A. (2008). Respiratory rate: the neglected vital sign. *Medical Journal of Australia*, *188*(11), 657-659.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biom J*, *47*(4), 458-472. https://doi.org/10.1002/bimj.200410135

Griffiths, P., Recio-Saucedo, A., Dall'Ora, C., Briggs, J., Maruotti, A., Meredith, P., Smith, G. B., Ball, J., & Missed, C. S. G. (2018). The association between nurse staffing and omissions in nursing care: A systematic review. *J Adv Nurs*, *74*(7), 1474-1487. https://doi.org/10.1111/jan.13564

Grunau, G., & Linn, S. (2018). Commentary: Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health*, *6*. https://doi.org/10.3389/fpubh.2018.00256

Haegdorens, F., Monsieurs, K. G., De Meester, K., & Van Bogaert, P. (2019). An intervention including the national early warning score improves patient monitoring practice and reduces mortality: A cluster randomized controlled trial. *J Adv Nurs*, *75*(9), 1996-2005. https://doi.org/10.1111/jan.14034

Haegdorens, F., Van Bogaert, P., Roelant, E., De Meester, K., Misselyn, M., Wouters, K., & Monsieurs, K. G. (2018). The introduction of a rapid response system in acute hospitals: A pragmatic stepped wedge cluster randomised controlled trial. *Resuscitation*, *129*, 127-134. https://doi.org/10.1016/j.resuscitation.2018.04.018

Jensen, J. K., Skår, R., & Tveit, B. (2019). Introducing the National Early Warning Score – A qualitative study of hospital nurses' perceptions and reactions. *Nursing Open*. https://doi.org/10.1002/nop2.291

Lane, D. J., Wunsch, H., Saskin, R., Cheskes, S., Lin, S., Morrison, L. J., & Scales, D. C. (2019). Assessing Severity of Illness in Patients Transported to Hospital by Paramedics: External Validation of 3 Prognostic Scores. *Prehosp Emerg Care*, 1-9. https://doi.org/10.1080/10903127.2019.1632998

Lee, Y. S., Choi, J. W., Park, Y. H., Chung, C., Park, D. I., Lee, J. E., Lee, H. S., & Moon, J. Y. (2018). Evaluation of the efficacy of the National Early Warning Score in predicting in-hospital mortality via the risk stratification. *J Crit Care*, *47*, 222-226. https://doi.org/10.1016/j.jcrc.2018.07.011

Linnen, D. T., Escobar, G. J., Hu, X., Scruth, E., Liu, V., & Stephens, C. (2019). Statistical Modeling and Aggregate-Weighted Scoring Systems in Prediction of Mortality and ICU Transfer: A Systematic Review. *J Hosp Med*, *14*(3), 161-169. https://doi.org/10.12788/jhm.3151

Lutgendorf, M. A., & Stoll, K. A. (2016). Why 99% may not be as good as you think it is: limitations of screening for rare diseases. *J Matern Fetal Neonatal Med*, *29*(7), 1187-1189. https://doi.org/10.3109/14767058.2015.1039977

Maharaj, R., Raffaele, I., & Wendon, J. (2015). Rapid response systems: a systematic review and meta-analysis. *Crit Care*, *19*, 254. https://doi.org/10.1186/s13054-015-0973-y

Mercaldo, N. D., Lau, K. F., & Zhou, X. H. (2007). Confidence intervals for predictive values with an emphasis to case-control studies. *Stat Med*, *26*(10), 2170-2183. https://doi.org/10.1002/sim.2677

Mitsunaga, T., Hasegawa, I., Uzura, M., Okuno, K., Otani, K., Ohtaki, Y., Sekine, A., & Takeda, S. (2019). Comparison of the National Early Warning Score (NEWS) and the Modified Early Warning Score (MEWS) for predicting admission and in-hospital mortality in elderly patients in the pre-hospital setting and in the emergency department. *PeerJ*, *7*, e6947. https://doi.org/10.7717/peerj.6947

Pimentel, M. A. F., Redfern, O. C., Gerry, S., Collins, G. S., Malycha, J., Prytherch, D., Schmidt, P. E., Smith, G. B., & Watkinson, P. J. (2019). 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*, *134*, 147-156. https://doi.org/10.1016/j.resuscitation.2018.09.026

Prytherch, D. R., Smith, G. B., Schmidt, P. E., & Featherstone, P. I. (2010). ViEWS--Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, *81*(8), 932-937. https://doi.org/10.1016/j.resuscitation.2010.04.014

RCOP. (2012). National Early Warning Score (NEWS): Standardising the assessment of acute- illness severity in the NHS. Report of a working party.

RCOP. (2017). National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. Updated report of a working party.

Rolfe, S. (2019). The importance of respiratory rate monitoring. *British Journal of Nursing*, *28*(8).

Romero-Brufau, S., Huddleston, J. M., Escobar, G. J., & Liebow, M. (2015). Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Crit Care*, *19*, 285. https://doi.org/10.1186/s13054-015-0999-1

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., & Featherstone, P. I. (2013). 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*, *84*(4), 465-470. https://doi.org/10.1016/j.resuscitation.2012.12.016

Smith, M. E., Chiovaro, J. C., O'Neil, M., Kansagara, D., Quiñones, A. R., Freeman, M., Motu'apuaka, M. L., & Slatore, C. G. (2014). Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. *Ann Am Thorac Soc*, *11*(9), 1454-1465. https://doi.org/10.1513/AnnalsATS.201403-102OC

Spångfors, M., Bunkenborg, G., Molt, M., & Samuelson, K. (2019). The National Early Warning Score predicts mortality in hospital ward patients with deviating vital signs: A retrospective medical record review study. *J Clin Nurs*, *28*(7-8), 1216-1222. https://doi.org/10.1111/jocn.14728

Stead, W. W. (2018). Clinical Implications and Challenges of Artificial Intelligence and Deep Learning. *JAMA*, *320*(11), 1107-1108. https://doi.org/10.1001/jama.2018.11029

Trevethan, R. (2017). Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Front Public Health*, *5*, 307. https://doi.org/10.3389/fpubh.2017.00307

IMPACT STATEMENT

- The National Early Warning Score (NEWS) is a valid screening instrument that can be used for adult patients admitted to general hospital wards.

- When a patient has a NEWS <5, we may assume that in the next 24 hours this patient is less likely to die unexpectedly, receive CPR or be transferred to the ICU.

- Because of the significant number of false positives when NEWS is ≥5, hospitals should limit workload by creating workable guidelines for clinical practice.

Table 1. Comparison of characteristics between the cross-sectional sample and the SAE sample.

	Cross-sectional sample (n = 1523)	**SAE sample (n = 390)**	**sig.**
Males, %	51.2	53.4	0.588 [a]
Age, mean (SD)	63.61 (14.46)	69.68 (14.46)	< 0.001 [b]
Max. NEWS, mean (SD)	2.09 (1.94)	5.78 (3.08)	< 0.001 [c]
Max. NEWS 0, %	22.6	4.4	< 0.001 [a]
Max. NEWS 1-4, %	66.1	29.7	
Max. NEWS 5-6, %	8.1	24.1	
Max. NEWS ≥ 7, %	3.2	41.8	
NEWS subscores, mean (SD)			
SATO2	0.76 (1.00)	1.82 (1.23)	< 0.001 [c]
SYSBP	0.86 (0.65)	1.27 (1.28)	< 0.001 [c]
HR	0.57 (0.50)	1.17 (1.05)	< 0.001 [c]
RR	0.27 (0.74)	1.01 (1.25)	< 0.001 [c]
AVPU	0.35 (0.09)	0.89 (0.04)	0.525 [c]
TEMP	0.64 (0.45)	0.84 (0.39)	0.374 [c]

a: Pearson's Chi-Square, b: independent samples t-test, c: Mann-Whitney U

AVPU: consciousness score, HR: heartrate, Max.: maximum, NEWS: national early warning score, RR: respiratory rate, SatO2: oxygen saturation, SD: standard deviation, sig: significance, SYSBP: systolic blood pressure, TEMP: temperature

Table 2. Confusion matrices and measures for diagnostic accuracy for the two composite outcomes in this study.

A. composite outcome (a+b+c)

	Outcome reached	**Outcome not reached**	
NEWS ≥ 5	257	172	429
NEWS < 5	133	1351	1484
	390	1523	

		95% confidence interval	
		lower	**upper**
Sensitivity	0.659	0.612	0.706
Specificity	0.887	0.871	0.903
Positive likelihood ratio	5.835	4.983	6.833
Negative likelihood ratio	0.384	0.335	0.442
Pre-test probability	1.23 %		
Positive predictive value	6.76 %		
Negative predictive value	99.52 %		

B. composite outcome (a+e+f)

	Outcome reached	**Outcome not reached**	
NEWS ≥ 5	49	380	429
NEWS < 5	16	1468	1484
	65	1848	

	95% confidence interval	
	lower	**upper**

Sensitivity	0.754	0.649	0.859
Specificity	0.794	0.776	0.813
Positive likelihood ratio	3.666	3.107	4.325
Negative likelihood ratio	0.310	0.202	0.474
Pre-test probability	0.21 %		
Positive predictive value	0.77 %		
Negative predictive value	99.93 %		

A. Composite outcome (a+b+c): unexpected death + CPR + unplanned ICU admission

B. Composite outcome (a+e+f): unexpected death + death after CPR + death after unplanned ICU admission

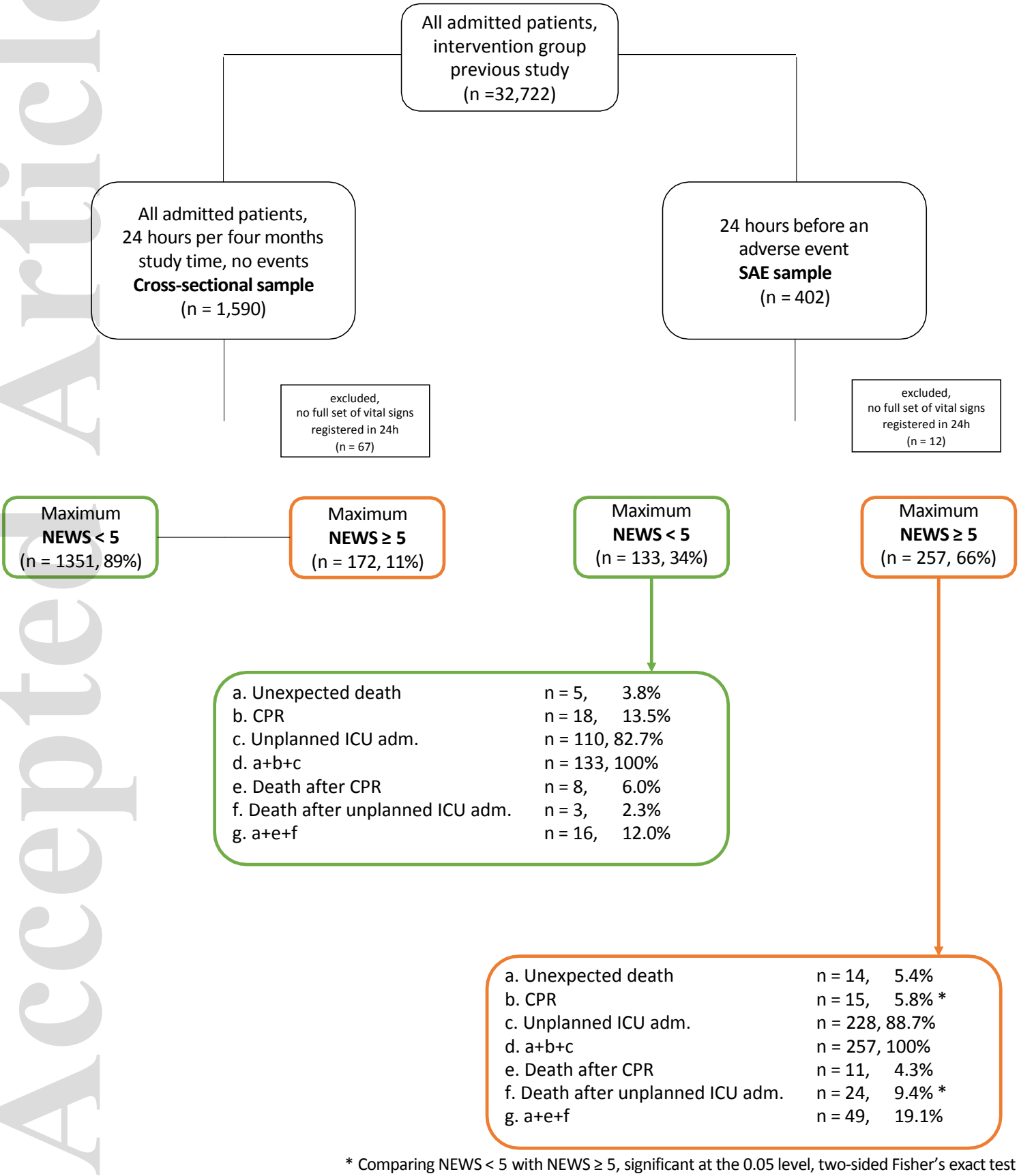
NEWS: National Early Warning Score

Table 3. Comparing area under the curve measures between different outcomes of the National Early Warning Score.

		95% confidence interval		optimal NEWS
	AUROC	**lower**	**upper**	**threshold**
Death after unplanned ICU admission (f)	0.885	0.830	0.940	≥ 4
Composite outcome (a+b+c)	0.841	0.817	0.865	≥ 4
Unplanned ICU admission (c)	0.836	0.811	0.862	≥ 4
Composite outcome (a+e+f)	0.815	0.760	0.870	≥ 5
Total mortality without DNAR	0.814	0.763	0.866	≥ 5
Total mortality	0.801	0.766	0.837	≥ 5
Unexpected death (a)	0.781	0.659	0.903	≥ 5
Death after CPR (e)	0.726	0.622	0.831	≥ 3
CPR (b)	0.716	0.643	0.789	≥ 3

AUROC: area under the receiver operating characteristic, CPR: cardiopulmonary resuscitation, DNAR: do not attempt resuscitation, ICU: intensive care unit, NEWS: national early warning score.

Optimal NEWS thresholds were calculated using Youden's J statistic.



* Comparing NEWS < 5 with NEWS ≥ 5, significant at the 0.05 level, two-sided Fisher's exact test