Risk stratification in emergency medicine: towards improved age and sex adjusted risk assessment

Candel, B.G.J.


EMERGENCY

8

DEVELOPMENT AND EXTERNAL VALIDATION OF THE INTERNATIONAL EARLY WARNING SCORE FOR IMPROVED AGE AND SEX ADJUSTED IN-HOSPITAL MORTALITY PREDICTION IN THE EMERGENCY DEPARTMENT.

Bart GJ Candel
Søren Kabell Nissen
Christian H. Nickel
Wouter Raven
Wendy Thijssen
Menno I Gaakeer
Annmarie Touborg Lassen
Mikkel Brabrand
Ewout W Steyerberg
Evert de Jonge
Bas de Groot

Published in Critical Care Medicine – 2023, Volume 51, Issue 7, p881–891

ABSTRACT

Objective Early Warning Scores (EWS) have a great potential to assist clinical decision making in the Emergency Department (ED). However, many EWS contain methodological weaknesses in development and validation and have poor predictive performance in older patients. The aim of this study was to develop and externally validate an International Early Warning Score (IEWS) based on a recalibrated NEWS model including age and sex and evaluate its performance independently at arrival to the ED in three age categories (18-65y; 66-80y; >80years).

Design and setting International multicenter cohort study using data from three Dutch EDs. External validation was performed in two EDs in Denmark.

Patients All consecutive ED patients ≥18years in the Netherlands Emergency department Evaluation Database (NEED) with at least two registered vital signs were included, resulting in 95,553 patients. For external validation, 14,809 patients were included from a Danish Multicenter Cohort (DMC).

Measurements Model performance to predict in-hospital mortality was evaluated by discrimination, calibration curves and summary statistics, reclassification, and clinical usefulness by decision curve analysis.

Main results In-hospital mortality rate was 2.4% (N=2314) in the NEED and 2.5% (N=365) in the DMC. Overall, the IEWS performed significantly better than NEWS with an AUROC of 0.89 (95% confidence Intervals 0.89-0.90) versus 0.82 (0.82-0.83) in the NEED and 0.87 (0.85-0.88) versus 0.82 (0.80-0.84) at external validation. Calibration for NEWS predictions underestimated risk in older patients and overestimated risk in the youngest, while calibration improved for IEWS with a substantial reclassification of patients from low to high risk and a standardized net benefit of 5-15% in the relevant risk range for all age categories.

Conclusions The IEWS substantially improves in-hospital mortality prediction for all ED patients ≥18years.

INTRODUCTION

Early Warning Scores (EWS) are widely used prediction tools to early detect clinical deterioration of patients and trigger intensive care consultation.[9, 10, 52, 169] By aggregating points for the degree of abnormality of each vital sign, EWS provide a likelihood for mortality, which should trigger the nurse or physician to get help or to intensify treatment. These scores are widely used in many settings and they are mandatory as a standard of care in the United Kingdom.[131] The National Early warning Score (NEWS) in particular has been widely implemented and is the most frequently used score to help identify critically ill patients early.[9, 131, 170, 171]

Some limitations of the NEWS and other EWS exist. Calibration of NEWS predictions is poor with relative overestimation of risk in younger ED patients and underestimation of risk in older ED patients.[27, 42, 119] NEWS assigns 0 to 3 points for all vital signs implying that all vital signs have similar predictive value, which has been shown to be unfounded.[11, 82] Furthermore, important risk differences exist between men and women at arrival to the ED.[172, 173] Nonetheless, most studies do not test the performance of early warning scores at older age or include sex differences.[10, 30] As a result, using NEWS may cause serious disadvantages for patient care and wrong treatment or disposition decisions.

The aim of this study was to develop and externally validate an International Early Warning Score (IEWS), by recalibrating NEWS including age and sex, to improve in-hospital mortality assessment at arrival to the ED.

METHODS

Study design and setting

This international multicenter cohort study is based on existing cohorts and reporting adheres to the Transparent Reporting of a multivariable model for Individual Prognosis or Diagnosis (TRIPOD) guidelines for prognostic modelling studies.[136] The Netherlands Emergency department Evaluation Database (NEED) was used as development cohort, consisting of three hospitals in the Netherlands. The NEED is the national quality registry for Emergency Departments (EDs) in the Netherlands and contributes to the improvement of transparency and quality of ED care in the Netherlands by supplying reliable data to the participating centers (see www.stichting-need.nl).[82] Data were prospectively collected and reviewed retrospectively. Data from the three sites spanned slightly different periods: data from one tertiary center (Leiden University Medical Center) included visits between 1 January 2017 – 8 June 2019, and data from the two level II emergency centers (Medical Center Leeuwarden and Catharina Hospital Eindhoven) were from 1 January 2019 – 12 January 2020 and from 1 January 2017 – 31 December 2019, respectively.

For external validation, we used the Danish Multicenter Cohort (DMC) which has been described previously.[118, 119] These data were not only from a different setting, but also from a different period to strengthen our validation. Patients were consecutively sampled in relation to previous prospective studies at two level II emergency centers: University Hospital of Southwest Jutland: (2 October 2008 - 12 February 2009; 23 February 2010 - 26 May 2010; 1 June 2012 - 1 November 2011; 24 April 2013 - 9 December 2013), and Lillebaelt Hospital (1 January 2010 - 30 June 2010).

Ethical considerations

In the Netherlands, the study was approved by the medical ethics committee of the Máxima MC on 2021, February 2 (ref number: Institutional Review Board N21.007). Under Danish law, retrospective registry studies are exempt from the need for approval by an ethics committee.[140] The study has been performed in accordance with the Helsinki declaration of 1975.

Selection of participants

All consecutive ED patients of ≥18 years were included in this study. Patients were excluded in the NEED if none or only one vital sign (systolic blood pressure (SBP), heart rate (HR), peripheral oxygen saturation (SpO_2), respiratory rate (RR) or temperature) were registered, as vital signs were considered missing not at random which prevented the possibility

for imputation (see supplementary file 1). Both studies collected data prospectively, but the DMC did so based on a prospective study design and the NEED was based on a registry. Hence, the missing data mechanisms differed for DMC (see supplementary file 2). Here, patients were excluded if neither systolic blood pressure nor pulse were recorded as these observations were missing not at random, i.e., unrelated to any of the observed variables, including outcomes.

Data collection

Demographic data were extracted from registers for both the NEED and DMC. Implausible physiological values were considered missing. Vital signs were recorded by a nurse in triage before ED treatment as described previously for the NEED,[82] and for the DMC.[118, 119]The first initial set of vital signs was registered before treatment.

NEWS aggregates seven vital signs (see table 1).[9] The NEWS was calculated for each patient (0-23points).[9] The collected Glasgow Coma Scale (GCS) was converted to an AVPU score.[119]

Table 1 The National Early Warning Score (NEWS)

Points	3	2	1	0	1	2	3
RR (/min.)	≤8		9-11	12-20		21-24	≥25
SpO2 (%)	≤91	92-93	94-95	≥96			
supplemental oxygen		Yes		No			
Temperature (°C)	≤35.0		35.1-36.0	36.1-38.0	38.1-39	≥39.1	
SBP (mmHg)	≤90	91-100	101-110	111-219			≥220
Pulse (bpm)	≤40		41-50	51-90	91-110	111-130	≥131
Level of consciousness				A			V, P, or U

RR: Respiratory Rate, $SpO_{2:}$ Peripheral oxygen saturation, SBP: Systolic blood pressure, °C: degrees Celsius, mmHg: millimeter mercury, bpm: beats per minute, A: Alert, V: Verbal, P:Pain, U: Unresponsive.

Outcome

The primary outcome was in-hospital mortality (including death in the ED). This outcome measure allowed us to compare our findings with previous studies.[119, 130] In the NEED, outcome information was registered and collected

from the minimal data set. In the DMC, information regarding mortality was collected retrospectively from the Danish Civil Registration System and the Danish National Patient Register.

Sample size estimation
See supplementary file 3.

Data analyses

Descriptive analyses
Data were presented as mean (standard deviation (SD)) if normally distributed and median (interquartile range (IQR)) if skewed.

Main statistical analyses
Predictive performance of NEWS and a recalibrated NEWS were evaluated in three age categories (18-65; 66-80; >80 years). These age categories were chosen based on previous age stratification.[82, 119] Prior to analyses, we assessed non-linearity of age in univariable logistic regression and explored non-linear terms (quadratic and restricted cubic splines) for best fit. Because patients were included if at least two vital signs were registered, missing data in the NEED were substituted by multiple imputation to reduce information bias described in supplementary file 3.[174]

For each imputation set, we calculated the NEWS. We used the vital sign categories as used in the NEWS as ordinal variables to fit the new model to prevent introducing thresholds different to those professionals are used to in current clinical practice with NEWS. We fitted the model NEWS+age+sex on the imputed data by multivariable logistic regression and, in a backwards selection approach, tested one-way and two-way interactions among predictors and found none of sufficient impact to include in the revised model. After deciding on recalibration, points were assigned and rounded to a recalibrated NEWS score based on a nomogram presentation, i.e., regression coefficients.[175] Points were rounded to nearest integer.

Predictive performance was compared in all three age categories of NEWS, recalibrated NEWS+age and the recalibrated NEWS+age+sex using Area under the Receiving Operating Characteristic (AUROC) with 95% Confidence Intervals (CI) and calibration plots. We averaged regression coefficients and intercepts across imputed sets to incorporate variance introduced by the imputation procedure. The best of the two recalibrated models was named the International Early Warning Score (IEWS).

To compare the net benefit of IEWS with NEWS, decision curves are presented.[176] This plots net-benefit at a range of risk thresholds for in-hospital mortality with the trade-off of benefit (true positive proportion) and harms (false positive proportion) on the same scale, adjusted by an appropriate exchange rate.[146] Because risk thresholds may differ by age group, separate decision curves were produced. To demonstrate how IEWS classifies patients differently than NEWS, a reclassification table was produced in which patients were allocated to low risk, medium risk or high risk subsets, stratified by outcome. In this example, we decided that the threshold from low to medium risk was two times the baseline risk and medium to high risk was three times the baseline risk (mean in-hospital mortality for patients with a NEWS <4points) per age category.

Internal and external validation
See supplementary file 3.

All analyses were performed in R statistical software (packages dplyr (v1.0.7;2021), rms (v6.2;2021), mice(v45;2011). A p-value <0.05 was considered as statistically significant.

RESULTS

Patient characteristics
In total, 95,553 patients could be included for analyses from the NEED with mean age 60.1 years (SD 19.4) and 50.3% male patients. Patient characteristics are described in supplementary file 4. Excluded patients had lower in-hospital mortality and fewer Intensive Care Unit (ICU) admissions than the included patients (see supplementary file 1). For external validation, a total of 14,809 patients were included. They had a mean age of 63 (SD 20) years, 51.9% were male. Patient characteristics in DMC were comparable to the NEED (supplementary file 5).

Main results
Age was used as a linear spline with no age effect assumed below 40 years based on its fit and association with mortality (supplementary file 6). A nomogram for the recalibrated NEWS plus age and sex was presented (supplementary file 7).

Based on the nomogram, points were assigned for a recalibrated NEWS+age score and a recalibrated NEWS+age+sex score resulting in a new risk score (table 2). Based on the calibration plots for NEED data (Figure 1) and for the DMC (Figure 2), the NEWS+age+sex was chosen as the best fit because

calibration improved visually and according to the slope and intercept in the relevant risk range for all age categories while discrimination was not affected by adding sex. The NEWS+age+sex model was therefore proposed as the IEWS. Flexible calibration curves are shown in supplementary file 8 and 9 for NEED data and DMC. Whereas the NEWS showed substantial underestimation of risk in older patients and overestimation of risk in younger patients, calibration for IEWS improved in both the development and validation cohort (Figures 1 and 2).

Table 2 The International Early Warning Score (IEWS)

Points	5	3	2	1	0	1	2	3	5
RR (/min.)					0-20		21-24	≥25	
SpO2 (%)		≤91		92-95	≥96				
supplemental oxygen				Yes	No				
Temperature (°C)	≤35.0	35.1-36.0			≥36.1				
SBP (mmHg)			≤90	91-110	111-219		≥220		
Pulse (bpm)				≤50	51-90	91-110	≥111		
Level of consciousness					A				V,P, or U
Sex				Male	Female				
Age, years					0-40	41-50	51-60	61-65 (66-75= 4points)	76-80=5 (81-90= 6points) (91-100= 7points)

RR: Respiratory Rate, SpO_2: Peripheral oxygen saturation, SBP: Systolic blood pressure, °C: degrees Celsius, mmHg: millimeter mercury, bpm: beats per minute, A: Alert, V: Verbal, P:Pain, U: Unresponsive.
The International Early Warning Score (IEWS) is a recalibrated model of the National Early Warning Score (NEWS) extended with age and sex.

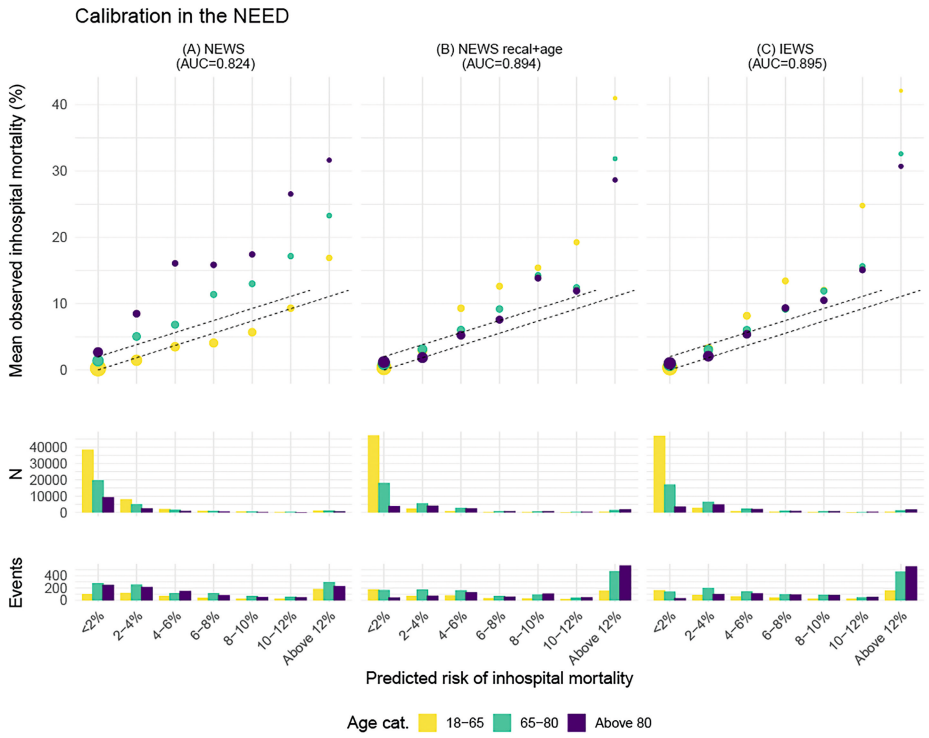


Figure 1 Internal calibration plots for the NEED (Netherlands Emergency department Evaluation Database) for (A) the National Early Warning Score (NEWS), (B) a recalibrated NEWS + age and (C) a recalibrated NEWS+age+sex (the International Early Warning Score (IEWS)). The predicted in-hospital mortality was categorized in steps of 2% in the relevant risk range. Calibration was assessed in three different age categories (18-65y, 66-80y,>80y). The dotted lines represent ideal calibration. The size of the dots indicates the precision of the estimate for observed inhospital mortality in each risk group, the larger, the higher the precision based on the inverse of the standard deviation. Below the calibration figures are the distribution of patients and outcomes presented for all three scores in histograms.

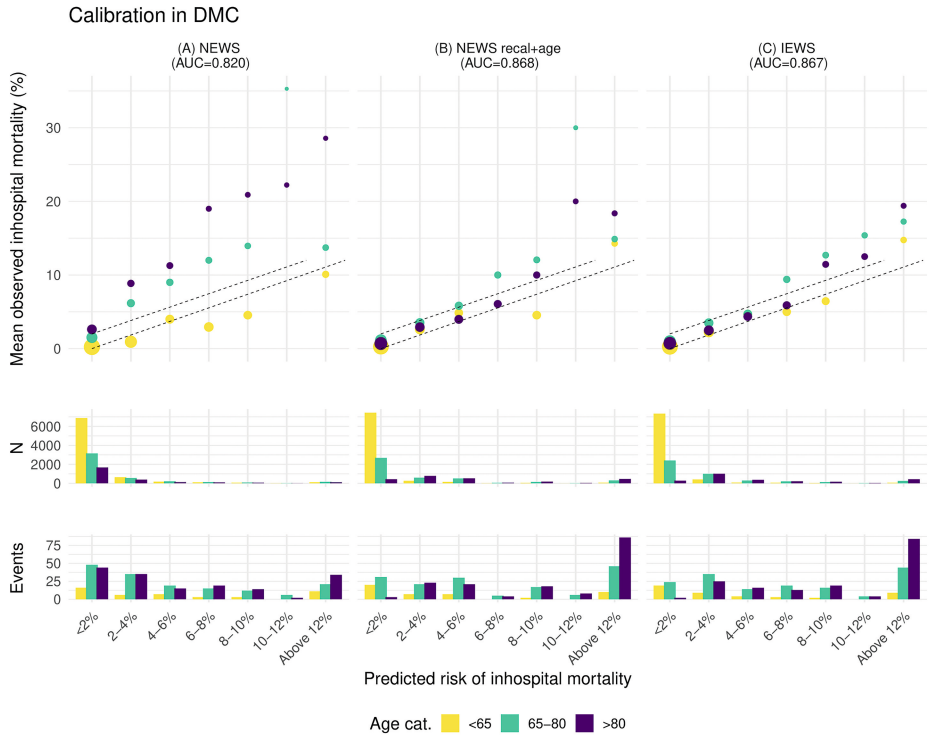


Figure 2 External calibration plots for the Danish Multicenter Cohort (DMC) for (A) the National Early Warning Score (NEWS), (B) a recalibrated NEWS + age and (C) a recalibrated NEWS+age+sex (the International Early Warning Score (IEWS)). The predicted inhospital mortality was categorized in steps of 2% in the relevant risk range. Calibration was assessed in three different age categories (18-65y, 66-80y,>80y). The dotted lines represent ideal calibration. The size of the dots indicates the precision of the estimate for observed inhospital mortality in each risk group, the larger, the higher the precision based on the inverse of the standard deviation. Below the calibration figures are the distribution of patients and outcomes presented for all three scores in histograms.

Overall, AUROC improved substantially for IEWS with 0.89 (95% CI 0.89-0.90) compared to NEWS 0.82 (95% CI 0.82-0.83 in the NEED and in the DMC with AUROC for IEWS 0.87 (95% CI 0.85-0.88) compared to NEWS 0.82 (95% CI 0.75-0.89). For most age categories, discrimination improved substantially (table 3). Internal validation showed good performance of IEWS (Supplementary file 10). Split sample analyses based on hospital location showed similar results (Supplementary file 11).

Table 3 Calibration and discrimination for NEWS and IEWS in the development and validation cohort

	Calibration		Discrimination	
Age groups	**Intercept**	**Slope**	**AUROC**	**95% CI**
Development cohort				
NEWS for in-hospital mortality in the NEED				
18-65y	-0.68	1.32	0.87	0.85-0.88
66-80y	0.38	0.97	0.80	0.79-0.81
>80y	0.97	0.91	0.78	0.77-0.80
Overall	0.18	1.09	0.82	0.82-0.83
IEWS for in-hospital mortality in the NEED				
18-65y	0.15	1.47	0.92	0.90-0.93
66-80y	0.21	1.23	0.85	0.84-0.86
>80y	0.16	1.18	0.83	0.82-0.85
Overall	0.18	1.24	0.89	0.89-0.90
Validation Cohort				
NEWS for in-hospital mortality in the DMC				
18-65y	-1.05	1.09	0.82	0.75-0.89
66-80y	0.43	0.82	0.78	0.74-0.82
>80y	0.94	0.84	0.78	0.74-0.81
Overall	0.20	0.98	0.82	0.80-0.84
IEWS for in-hospital mortality in the DMC				
18-65y	-0.52	1.05	0.86	0.80-0.91
66-80y	-0.04	0.88	0.80	0.76-0.83
>80y	-0.25	0.83	0.77	0.73-0.81
Overall	-0.21	0.94	0.87	0.85-0.88

NEWS: National Early Warning Score, IEWS: International Early Warning score, AUROC: Area under the Receiving Operating Curve, 95% CI: 95percent Confidence Intervals, NEED: Netherlands Emergency Department Evaluation Database, DK: Danish Multicenter Cohort.
The International Early Warning Score(IEWS) is a recalibrated model of the NEWS including the additional variables age and sex.

Decision curve analyses showed for each age category a standardized Net Benefit of 5-15% in the relevant risk range of 1% to 15% (figure 3). As an example, in a population with approximately 24 in-hospital deaths per 1000 patients, for a decision threshold of 5% in-hospital mortality risk, the IEWS would identify 42% additional true deaths (Standardized net benefit at a threshold of 5% for IEWS in figure 3), without increasing the number of false positive predictions compared to not using any model. Compared to using NEWS, the IEWS would identify 15% additional true deaths without increasing the number of false positive predictions.

To give a better insight in the benefit of using IEWS compared to NEWS, a reclassification table is presented in Supplementary file 12.

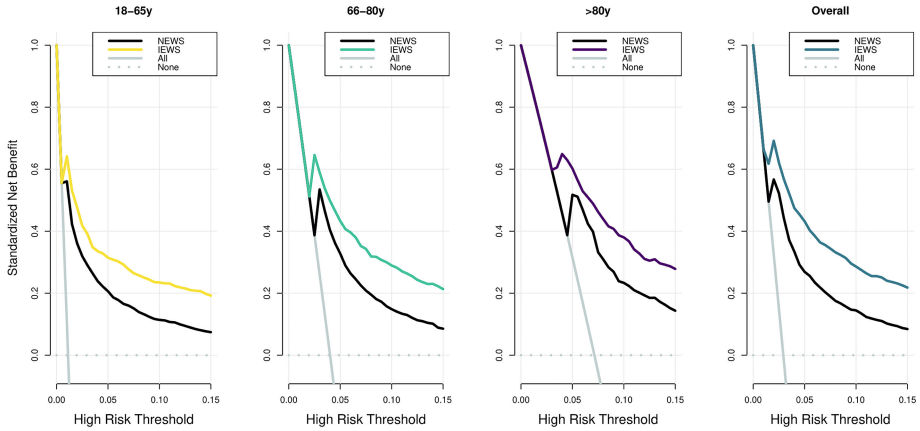


Figure 3 The decision curve analyses show for each age category a standardized Net Benefit of 5-15% in the relevant risk range of 1% to 15%. One physician may be more defensive and use, for example, 5% as a threshold for additional interventions (i.e. Intensive Care consultation or broad spectrum antibiotics) for patients >80years and older whereas another physician may choose to use a threshold of 10% for older patients. With both thresholds, the IEWS performs better than the NEWS, with 10-15% net more true positives (i.e. identified more patients who died corrected for the number of false positive classifications).

DISCUSSION

This large international multicenter cohort study shows that the IEWS, a recalibrated model based on NEWS including age and sex, performs significantly better compared to the widely adopted NEWS for the prediction of in-hospital mortality in ED patients of all age categories in a development and external validation cohort.

Most early warning scores have methodological weaknesses that could have detrimental effects on patient care if used in clinical practice.[130, 177] For example, NEWS, based on the VitalPAC early warning score (ViEWS), did not include age because AUROC only slightly improved after including age.[30, 13]

[3] However, calibration, a key aspect of prediction model performance,[130, 136] has not been assessed and age was used as a dichotomous variable below or above 65 years instead as continuous predictor. Furthermore, points for vital sign disturbances were allocated based on clinical consensus rather than on a statistical approach.[10, 30]

Our results are in line with several studies which have demonstrated that including age to an early warning score improved predictive performance substantially.[42, 133, 135, 151] However, none of these studies followed the recommended steps for development and validation of prediction models neither have they shown a classification in low to high risk.[178, 179] Our decision curve analysis and reclassification table demonstrate that for both younger and older patients the IEWS has considerable incremental value with more young patients correctly classified as low risk, and more importantly, more older patients correctly classified as high risk for in-hospital mortality. Previous studies have shown that predictive performance only improved for younger patients using an age-specific early warning score on a composite outcome of mortality, cardiac arrest and ICU admission compared to NEWS.[166, 180, 181] However, the modeling approach was very different from ours. Points were assigned to vital signs based on their distribution rather than on regression coefficients as recommended for prediction modeling.[130, 174, 178] This may have caused the age-specific model to underperform in older age. Our group has demonstrated previously that the addition of age to NEWS without recalibration of the physiological variables already improved predictive performance for in-hospital mortality.[119]

Early warning scores are designed for prognostication and can be used as early as in the ED and add to the clinical evaluation of a patient's disease severity.[182] While clinical evaluation may vary among physicians depending on years of experience,[183] the IEWS provides a numerical mortality risk (a percentage) that hypothetically may help with clinical decision-making.

For an easily adopted and implemented early warning score it is essential that the variables in the score are easily measured, readily available and strong predictors of the primary outcome.[176, 178] The physiological variables used in NEWS meet all these requirements.[11, 82] In addition, age and sex exhibit the same qualities and are predictors for in-hospital mortality.[119, 172]

Other variables have been proposed to use in early warning scores, such as biomarkers or frailty measures.[73, 156, 184-186] For frailty, only four out of 60 frailty scores could be measured in less than one minute using vignettes.[187] Though, in clinical practice, it may be difficult and not reliable to assess frailty, for example if the level of consciousness is altered and no history is available. Other variables such as biomarkers are not readily available or easily repeated without high costs. For these reasons, we have only evaluated age and sex as additional variables to the seven predictors of NEWS which both met the criteria for reliable predictors and are always known or can at least be estimated precisely.[130, 178] In patients who received prehospital treatments from paramedics or medical emergency services, the physiological variables may already be improved at arrival to the ED and thus the risk may be underestimated by using the NEWS or IEWS, a phenomenon called lead-time bias in literature.[188] Prehospital treatments have not been considered in the model. However, the IEWS still performs better overall than NEWS also in the ED.

The present study has several strengths. We adhered to the TRIPOD guidelines and followed steps recommended for the development of prediction models.[130, 136, 174, 178]. We used a large sample size in relation to the number of predictors for both development and validation and validated our findings externally in a different European country in a different time-period. The IEWS is clinically useful in the relevant risk range for each age category. Further validation is desired to assess generalizability of the proposed IEWS across multiple settings.[189, 190] Other limitations need to be considered. First, a risk of selection bias may be present as we excluded patients in whom less than two vital signs were registered. However, these patients were at very low risk of mortality (for example wounds and fractures) or at very high risk (cardiac arrest) and therefore these patients would have been recognized as low or high risk also without an early warning score. As recommended, we used multiple imputation to prevent information bias so we could include as many patients as possible.[130] Notably, around 90% of AVPU values were missing in the development cohort. However, missingness was clearly related to outcomes and other measured variables (Supplementary file 13), IEWS worked very well in the external validation cohort with a very low missingness of the AVPU variable. Hence, the bias incurred by imputing AVPU is likely negligible, despite a high proportion of missingness.[191] Secondly, in-hospital mortality was chosen as the primary outcome. A time horizon of a few days only is recommended for early warning scores.[130] Nevertheless, the time till patients died in-hospital in our data was short with a median of 4days (IQR 1-9), which allowed us to compare our results with previous studies and assess deterioration of patients.[9, 131, 192] Thirdly, the physiological

variables have been categorized based on the NEWS because physicians are used to work with these thresholds in clinical practice. However, it has been recommended to avoid categorizing predictors in the statistical analysis. For this reason, we repeated our analysis using restricted cubic splines for each physiological variable and presented a nomogram (Supplementary file 14). Using this nomogram would have resulted in similar distribution of points as in the IEWS after rounding. Thus, categorization of variables did not lead to poor modelling. Nonetheless, using different points for each physiological variable may lead to calculation errors as physicians are used to using the NEWS. This could be overcome by calculating the score electronically. Lastly, the NEWS2, a modification of the original NEWS, has not been evaluated in this study for several reasons. First, mortality assessment did not improve using NEWS2 compared to NEWS in a previous large study.[193] Secondly, the two major updates introduced in NEWS2 were separate thresholds for saturation in patients with hypercapnic failure and the addition of confusion in consciousness scale. We did not record confusion, which makes it impossible to use the NEWS2 consciousness scale. Additionally, information about current or previous hypercapnic failure is often not available at arrival or requires arterial blood gas. We therefore bases the IEWS on the foundation laid out in NEWS rather than NEWS2. Comparing the IEWS with other widely adopted EWS, such as the Modified Early Warning Score (MEWS), would have resulted in similar results, as the design of MEWS was neither based on a statistical approach, nor it includes age or sex.[135]

In summary, this large international multicenter cohort study shows that the IEWS performs substantially better than the widely adopted NEWS for predicting mortality in ED patients of all age categories in a development and external validation cohort. Future studies should investigate further evidence for predictive validity and assess whether implementation of IEWS in the ED leads to lower adverse events compared to not using an early warning score or using NEWS.

SUPPLEMENTARY FILE 1

Patient characteristics of excluded patients: Patient characteristics of both included (0-3 missing vitals) and excluded patients (4-5missing vitals) are presented in the table for comparison. Patients with four or five missing vital signs (Systolic blood pressure, temperature, heart rate, peripheral oxygen saturation or respiratory rate) were excluded from the analyses, because vital signs were considered missing not at random in these patients. Excluded patients are described in the table.

Excluded patients had lower in-hospital mortality, less ICU admissions and lower urgency of triage compared to included patients.

		Included patients Number of missing vital signs			**Excluded patients**	
NEED cohort	**0 (N=50378)**	**1 (N=29174)**	**2 (N=10383)**	**3 (N=5618)**	**4 (N=5843)**	**5 (N=47393)**
In-hospital mortality						
died	1406 (2.8%)	614 (2.1%)	215 (2.1%)	79 (1.4%)	57 (1.0%)	478 (1.0%)
Missing	681 (1.4%)	437 (1.5%)	210 (2.0%)	159 (2.8%)	148 (2.5%)	519 (1.1%)
ICU admission						
ICU admission	821 (1.6%)	549 (1.9%)	143 (1.4%)	48 (0.9%)	31 (0.5%)	310 (0.7%)
No ICU admission	49195 (97.7%)	28267 (96.9%)	9965 (96.0%)	4770 (84.9%)	5145 (88.1%)	39976 (84.4%)
Missing	362 (0.7%)	358 (1.2%)	275 (2.6%)	800 (14.2%)	667 (11.4%)	7107 (15.0%)
Triage category*						
immediate	3376 (6.7%)	1370 (4.7%)	396 (3.8%)	110 (2.0%)	86 (1.5%)	632 (1.3%)
very urgent	17015 (33.8%)	7761 (26.6%)	2218 (21.4%)	665 (11.8%)	516 (8.8%)	4913 (10.4%)
urgent	20768 (41.2%)	12855 (44.1%)	4731 (45.6%)	2927 (52.1%)	2515 (43.0%)	16502 (34.8%)
non-urgent	8552 (17.0%)	6600 (22.6%)	2825 (27.2%)	1835 (32.7%)	2595 (44.4%)	20845 (44.0%)
Missing	667 (1.3%)	588 (2.0%)	213 (2.1%)	81 (1.4%)	131 (2.2%)	4501 (9.5%)

*Triage category according to the Manchester Triage System or Dutch Triage Standard.

SUPPLEMENTARY FILE 2

Patient characteristics of excluded patients in DMC: Patients were excluded if neither systolic blood pressure nor pulse were recorded as these observations were missing not at random, i.e., unrelated to any of the observed variables, including outcomes.

DMC Cohort	**Included N=14809**						**Excluded N=2039**				**Pearson χ^2**
	Number of missing vital signs (frequency)										
	0 (9132)	1 (4075)	2 (1357)	3 (217)	4 (21)	5 (7)	3 (4)	4 (7)	5 (201)	6 (1827)	
Inhospital mortality, N (%)	215 (2.4)	105 (2.6)	36 (2.7)	8 (3.7)	1 (4.8)	0 (0)	1 (25.0)	0 (0)	7 (3.5)	50 (2.7)	
Subtotal	365 (2.5)						59 (2.8)				χ^2=1.06, P=0.304
ICU admission, N (%)	220 (2.4)	142 (3.5)	46 (3.4)	8 (3.7)	1 (4.8)	0 (0)	1 (25.0)	1 (14.3)	7 (3.5)	51 (2.8)	
Subtotal	417 (2.8)						60 (2.9)				χ^2=0.1, P=0.746

Missing vital signs include respiratory rate, heart rate, systolic blood pressure, peripheral oxygen saturation, temperature and avpu.

SUPPLEMENTARY FILE 3

Sample size estimation

The NEED contained 148,828 ED visits of patients ≥18 years. We estimated that in ~60% of the ED visits at least two or more vital signs were registered resulting in ~90,000 ED visits which could be used for the analyses with approximately 2300 events (in-hospital mortality). This number is more than sufficient for reliable analyses.[178] Numbers were also large in age-based subgroups. For external validation a minimum of in total 200 events is recommended.[178] The DMC contains approximately 350 events which should be sufficient for external validation.[119]

Multiple imputation procedure

Because patients were included if at least two vital signs were registered, missing data in the NEED were substituted by multiple imputation to reduce information bias.[174] We used the chained equations procedure, after imputation was deemed feasible based on patterns of missingness.[179, 194] For a better multiple imputation procedure, we also used triage category (non-urgent, urgent, very urgent, immediate; according to the Manchester Triage System or Dutch Triage Standard), urea, leukocytes, and fluid administration (0, 0-500ml, >500ml) as a predictor in the imputation procedure. Outcome was imputed if missing. Imputation parameters for DMC have been described in detail previously.[119] We obtained 20 estimates of the missing vital signs for each patient with five iterations each. We checked for collinearity and convergence during the imputation procedure.

Internal and external validation

For internal validation, a bootstrapping was performed with 200 repetitions on the imputed data and the overall AUROC and calibration were presented.[195] Also, a non-randomly split sample analysis was performed based on hospital location in the NEED cohort.

The DMC was used for external validation. Predictive performance in terms of discrimination and calibration was assessed and based on the average estimate of risk. A net benefit curve was produced.

SUPPLEMENTARY FILE 4

Patient characteristics of the NEED, the Development cohort

NEED cohort	18-65years (N=51573)	66-80years (N=29591)	>80years (N=14389)	All (N=95553)
Age, years				
Mean (SD)	45.7 (14.0)	72.9 (4.21)	85.8 (3.87)	60.1 (19.4)
Sex				
male	25424 (49.3%)	16363 (55.3%)	6316 (43.9%)	48103 (50.3%)
Systolic Blood Pressure (mmHg)				
Mean (SD)	128 (28.2)	137 (33.2)	142 (34.3)	133 (31.3)
Missing	4380 (8.5%)	1525 (5.2%)	559 (3.9%)	6464 (6.8%)
Heart rate (bpm)				
Mean (SD)	87.0 (20.7)	85.7 (21.6)	83.4 (20.8)	86.0 (21.0)
Missing	7037 (13.6%)	2647 (8.9%)	1232 (8.6%)	10916 (11.4%)
Respiratory Rate (/min.)				
Median [IQR]	16.0 [14.0-19.0]	17.0 [15.0-21.0]	18.0 [15.0-22.0]	17.0 [14.0-20.0]
Missing	16800 (32.6%)	7103 (24.0%)	3226 (22.4%)	27129 (28.4%)
Peripheral oxygen saturation (%)				
Mean (SD)	97.8 (2.96)	96.4 (3.81)	96.0 (3.84)	97.1 (3.47)
Missing	2443 (4.7%)	1397 (4.7%)	702 (4.9%)	4542 (4.8%)
Level of consciousness				
Alert	4311 (8.4%)	3290 (11.1%)	1840 (12.8%)	9441 (9.9%)
Verbal	178 (0.3%)	118 (0.4%)	114 (0.8%)	410 (0.4%)
Pain	109 (0.2%)	44 (0.1%)	36 (0.3%)	189 (0.2%)
Unresponsive	57 (0.1%)	42 (0.1%)	15 (0.1%)	114 (0.1%)
Missing	46918 (91.0%)	26097 (88.2%)	12384 (86.1%)	85399 (89.4%)
Temperature (degrees Celcius)				
Median [IQR]	37.0 [36.5-37.4]	36.8 [36.5-37.4]	36.8 [36.4-37.3]	36.9 [36.5-37.4]
Missing	9820 (19.0%)	5161 (17.4%)	2762 (19.2%)	17743 (18.6%)

NEED cohort	18-65years (N=51573)	66-80years (N=29591)	>80years (N=14389)	All (N=95553)
Supplemental Oxygen				
yes	1421 (2.8%)	2130 (7.2%)	1449 (10.1%)	5000 (5.2%)
Missing	15931 (30.9%)	6984 (23.6%)	2672 (18.6%)	25587 (26.8%)
Fluid administration (ml)				
0cc	24458 (47.4%)	16065 (54.3%)	7807 (54.3%)	48330 (50.6%)
0-500cc	4593 (8.9%)	2850 (9.6%)	1496 (10.4%)	8939 (9.4%)
>500cc	5248 (10.2%)	3231 (10.9%)	1377 (9.6%)	9856 (10.3%)
Missing	17274 (33.5%)	7445 (25.2%)	3709 (25.8%)	28428 (29.8%)
Triage category, N (%)	*			
non-urgent	11076 (21.5%)	5683 (19.2%)	3053 (21.2%)	19812 (20.7%)
urgent	22703 (44.0%)	12510 (42.3%)	6068 (42.2%)	41281 (43.2%)
very urgent	14413 (27.9%)	9071 (30.7%)	4175 (29.0%)	27659 (28.9%)
immediate	2541 (4.9%)	1871 (6.3%)	840 (5.8%)	5252 (5.5%)
Missing	840 (1.6%)	456 (1.5%)	253 (1.8%)	1549 (1.6%)
Urea (mmol/L)				
Median [IQR]	4.9 [3.8-6.2]	6.7 [5.3-9.0]	8.2 [6.2-11.4]	5.8 [4.4-7.9]
Missing	11273 (21.9%)	4435 (15.0%)	2375 (16.5%)	18083 (18.9%)
Leukocytes (*10^9)				
Median [Min, Max]	9.2 [7.0-12.0]	9.0 [6.9-12.1]	9.2 [7.1-12.2]	9.1 [7.0-12.1]
Missing	10667 (20.7%)	4001 (13.5%)	1921 (13.4%)	16589 (17.4%)
In-hospital mortality, N(%)				
Died	488 (0.9%)	981 (3.3%)	845 (5.9%)	2314 (2.4%)
Missing	608 (1.2%)	554 (1.9%)	325 (2.3%)	1487 (1.6%)
ICU admission, N (%)				

NEED cohort	18-65years (N=51573)	66-80years (N=29591)	>80years (N=14389)	All (N=95553)
ICU admission	828 (1.6%)	582 (2.0%)	151 (1.0%)	1561 (1.6%)
Missing	1385 (2.7%)	319 (1.1%)	91 (0.6%)	1795 (1.9%)

SD: Standard deviation, IQR: Interquartile Range, ICU: Intensive Care Unit
*Triage category according to the Manchester Triage System or Dutch Triage Standard.

SUPPLEMENTARY FILE 5

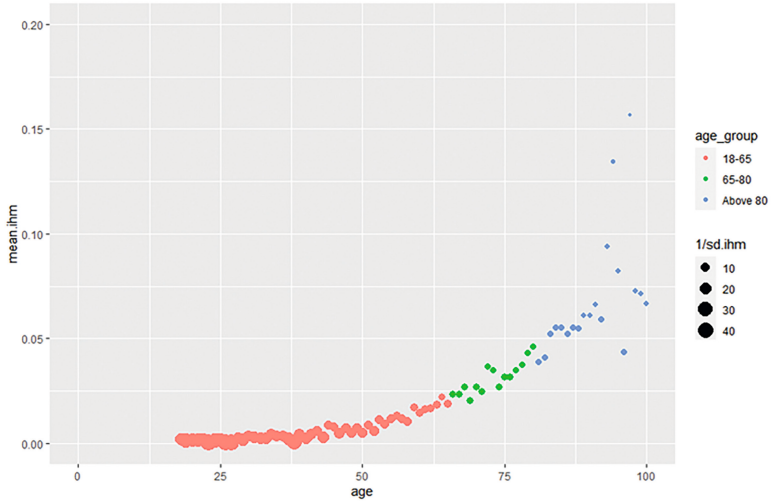
Patient characteristics of the Danish Multicenter Cohort (DMC) used for external validation.

DMC cohort	<65 (N=7990)	65-80 (N=4316)	>80 (N=2503)	All (N=14809)
Age				
Mean (SD)	44.9 (13.9)	73.0 (4.4)	86.1 (4.1)	60.0 (20.0)
Gender				
male	4,186 (52.4%)	2,146 (49.7%)	1,352 (54.0%)	7,684 (51.9%)
Systolic blood pressure (mmHg)				
Mean (SD)	135 (21.6)	140 (26.2)	140 (27.4)	137 (24.1)
Missing	21 (0.3%)	11 (0.3%)	12 (0.5%)	44 (0.3%)
Heart Rate (bpm)				
Mean (SD)	84.8 (19.6)	84.6 (20.3)	81.6 (19.6)	84.2 (19.8)
Missing	54 (0.7%)	40 (0.9%)	29 (1.2%)	123 (0.8%)
Respiratory Rate (/min)				
Median [IQR]	16 [5]	16 [5]	18 [8]	16 [6]
Missing	2,340 (29.3%)	1,073 (24.9%)	624 (24.9%)	4,037 (27.3%)
Peripheral oxygen saturation (%)				
Mean (SD)	97.1 (3.1)	95.4 (4.2)	94.9 (4.7)	96.2 (3.8)
Missing	343 (4.3%)	159 (3.7%)	135 (5.4%)	637 (4.3%)
Level of consciousness				
Alert	6,891 (86.2%)	3,712 (86.0%)	2,039 (81.5%)	12,642 (85.4%)
Vocal	101 (1.3%)	84 (1.9%)	84 (3.4%)	269 (1.8%)
Pain	36 (0.5%)	23 (0.5%)	22 (0.9%)	81 (0.5%)
Unresponsive	23 (0.3%)	10 (0.2%)	8 (0.3%)	41 (0.3%)
Missing	939 (11.8%)	487 (11.3%)	350 (14.0%)	1,776 (12.0%)
Temperature (Degrees Celsius)				
Median [IQR]	37.0 [0.8]	36.9 [0.9]	36.9 [0.9]	37.0 [0.8]
Missing	526 (6.6%)	234 (5.4%)	182 (7.3%)	942 (6.4%)

DMC cohort	<65 (N=7990)	65-80 (N=4316)	>80 (N=2503)	All (N=14809)
Supplementary oxygen				
Yes	353 (4.4%)	427 (9.9%)	309 (12.3%)	1,089 (7.4%)
In-hospital mortality, N (%)				
Died	46 (0.6%)	156 (3.6%)	163 (6.5%)	365 (2.5%)
ICU admission, N (%)				
ICU admission	197 (2.5%)	144 (3.3%)	76 (3.0%)	417 (2.8%)

SUPPLEMENTARY FILE 6

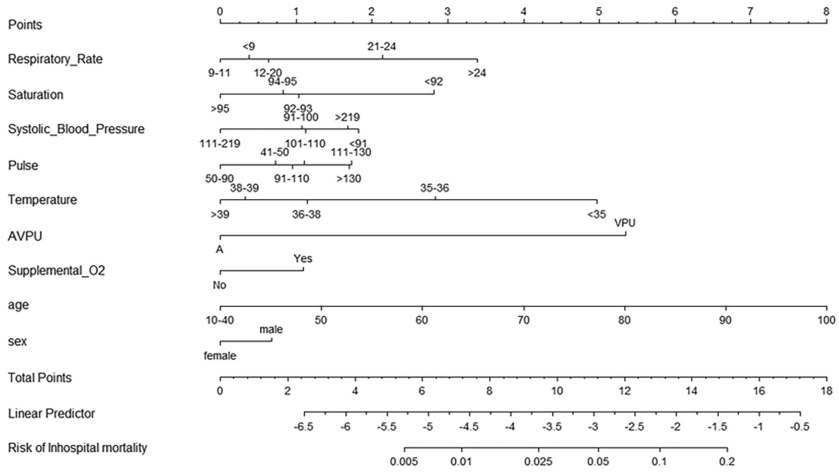
The association between age and in-hospital mortality



Ihm: in-hospital mortality
Included patients from the NEED (N=95,553). Age was used as a linear predictor in the prediction model above 40years old. The size of the dots indicates the precision of the estimate for observed in-hospital mortality and is based on the inverse of the standard deviation. The larger, the higher the precision.

SUPPLEMENTARY FILE 7

Nomogram for a recalibrated National Early Warning Score +age+sex



All predictors were used as categorized variables as used in the original National Early Warning Score (NEWS). A nomogram was developed based on the regression coefficients of all predictors (see below). A nomogram is a graphical presentation of model fit, allowing points to be awarded on a scale that is proportional to the log-odds. Points were rounded to nearest integer.
A nomogram fitted on the primary outcome (in-hospital mortality). As can be seen in the nomogram, age was the strongest predictor for mortality followed by AVPU. Male sex was associated with increased mortality. The nomogram for a recalibrated NEWS score with age and without sex was comparable to the recalibrated NEWS with age and with sex.

Regression coefficients and odds ratio's of multivariable logistic regression

Variables	Regression coefficients	Odds ratio's (95% CI)	Variables	Regression coefficients	Odds Ratio's (95% CI)
Intercept	-1,45	-	Pulse (<41bpm)	-0,22	1.5 (0.9-2.4)
RR (<9/min.)	0,4	0.84 (0.35-2.0)	Pulse (41-50bpm)	0,31	1.4 (1.0-1.9)
RR (9-11/min.)	-0,19	0.82 (0.54-1.6)	Pulse (51-90bpm)	0	Ref.
RR (12-20/min.)	0	Ref.	Pulse (91-110bpm)	-0,84	1.5 (1.3-1.6)
RR (21-24/min.)	1,01	1.87 (1.6-2.2)	Pulse (110-130bpm)	-0,23	2.0 (1.7-2.3)
RR (>24/min.)	1,91	3.1 (2.7-3.6)	Pulse(>130bpm)	-0,83	2.0 (1.6-2.5)
SPO2(>95%)	0	Ref.	Temp(<35dgr)	0,14	4.7 (3.4-6.7)
SpO2(94-95%)	0,35	1.4 (1.2-1.6)	Temp(35-36dgr)	0,71	2.0 (1.7-2.4)
SPO2(92-93%)	-0,29	1.5 (1.3-1.8)	Temp(36-38dgr)	0	Ref.
SPO2(<92%)	0,083	3.1 (2.7-3.6)	Temp(38-39dgr)	-2,44	0.72 (0.60-0.87)
SBP (>219mmHg)	-0,98	2.1 (1.5-3.1)	Temp(>39dgr)	-3,3	0.62 (0.48-0.80)
SBP(111-219mmHg)	0	Ref.	Alert	0	Ref.
SBP (101-110mmHg)	0,43	1.5 (1.3-1.8)	VPU	0,72	8.8 (7.4-10.5)
SBP(91-100mmHg)	-0,43	1.5 (1.3-1.8)	O2 (No)	0	Ref.
SBP(<91mmHg)	-0,58	2.1 (1.8-2.4)	O2 (yes)	0,23	1.6 (1.4-1.8)
Age (counting from 63)	-0,95	4.5 (4.1-5.0) /28y	Sex (female)	-0,29	0.74 (0.68-0.82)

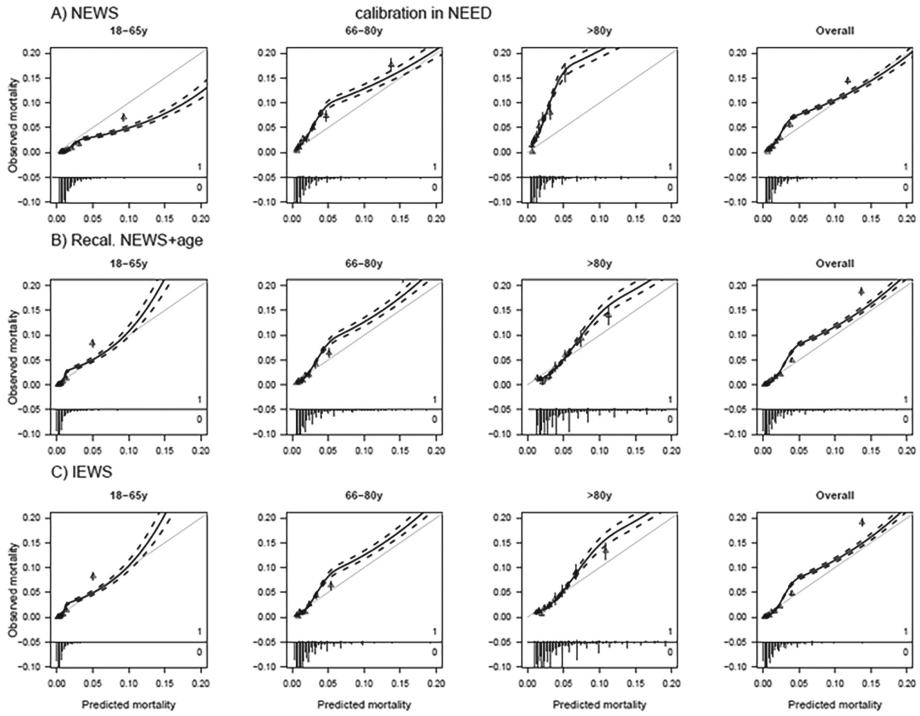
Function

function(Rr2 = 0,Sat = 0,Sbp2 = 0,Pulse2 = 0,Temp2 = 0,Avpu = 0,O2 = 0,age = 63,sex = “male”) {Rr2 <- ordered(Rr2);Sat <- ordered(Sat);Sbp2 <- ordered(Sbp2);Pulse2 <- ordered(Pulse2);Temp2 <- ordered(Temp2);age <- Function(age2)(age);

-1.4534925-0.19432847*(RR== 9-11) +1.0141945*(RR==21-24) +0.40291742*(RR<=9) +1.9103567*(RR>=24)+0.34889429*(SPO2=94-95)-0.28768878*(Sat==92-93) +0.082739612*(Sat<92) +0.43450926*(SBP=101-110)-0.43269485*(Sbp=91-100)-0.57997328*(Sbp<91)-0.97877706*(Sbp>219)+0.30637857*Pulse(41-50)-0.21608917*(Pulse<41)-0.233718-59*(Pulse= 110-130)-0.8381017*(Pulse=91-110)-0.83097791*(Pulse>130)+0.7070425*Temp(35-36)+0.14159026*(Temp<35)-2.4456138*(Temp2=38-39)-3.2986931*(Temp>39)+0.72456848*vpu+0.23021481*O2(yes)-0.94793297*age-0.29486964*(sex==”female”) }

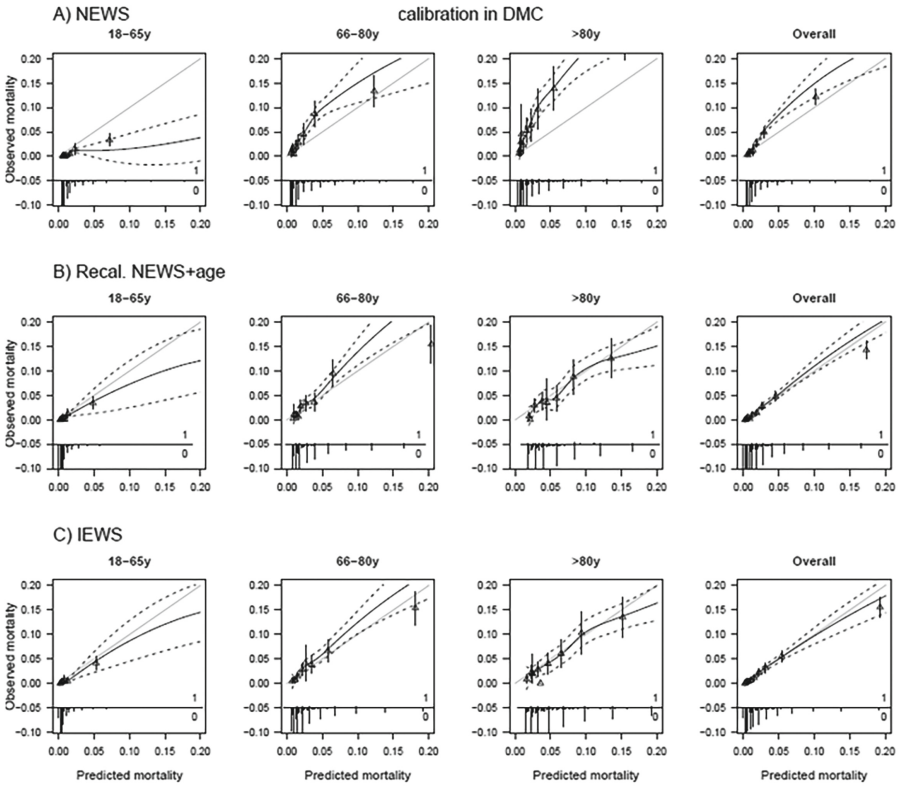
SUPPLEMENTARY FILE 8

Flexible calibration plots in the NEED



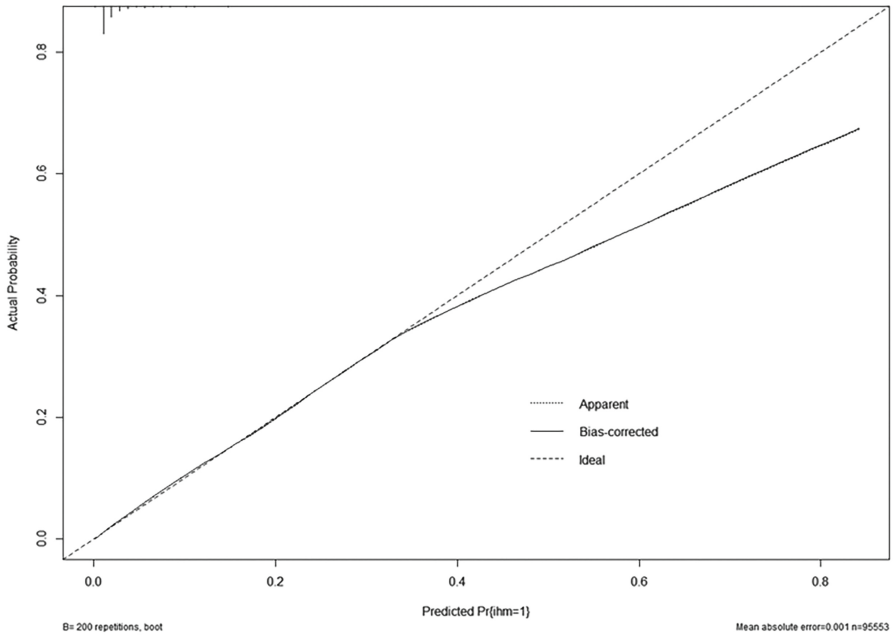
SUPPLEMENTARY FILE 9

Flexible calibration plots in the DMC



SUPPLEMENTARY FILE 10

Internal validation**:** Internal validation using bootstrap with 200 random sample repetitions on the NEED.



AUROC overall = 0.87
Calibration plot. Notes: Apparent refers to apparent performance for calibration; bias-corrected refers to optimism-corrected in internal validation; bootstrap=200.

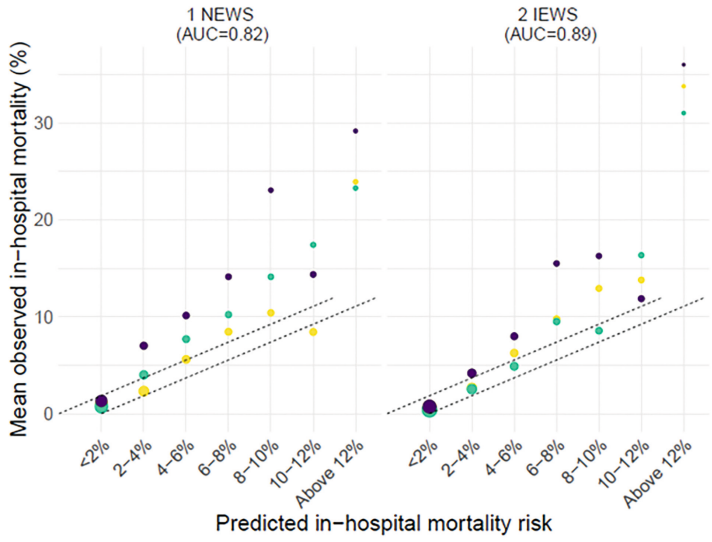
SUPPLEMENTARY FILE 11

Split sample analysis based on hospital location in the NEED.

	Discrimination	
Age groups	**AUROC**	**95% CI**
NEWS for in-hospital mortality		
Hospital 1 (tertiary care center)	0.80	0.78-0.82
Hospital 2	0.86	0.85-0.87
Hospital 3	0.82	0.81-0.84
IEWS for in-hospital mortality		
Hospital 1 (tertiary care center)	0.87	0.86-0.89
Hospital 2	0.90	0.90-0.91
Hospital 3	0.87	0.86-0.89

NEWS: National Early Warning Score, IEWS: International Early Warning score, AUROC: Area under the Receiving Operating Curve, 95% CI: 95percent Confidence Intervals, NEED: Netherlands Emergency Department Evaluation Database.
The International Early Warning Score is a recalibrated model of the NEWS including age and sex as extra variables.

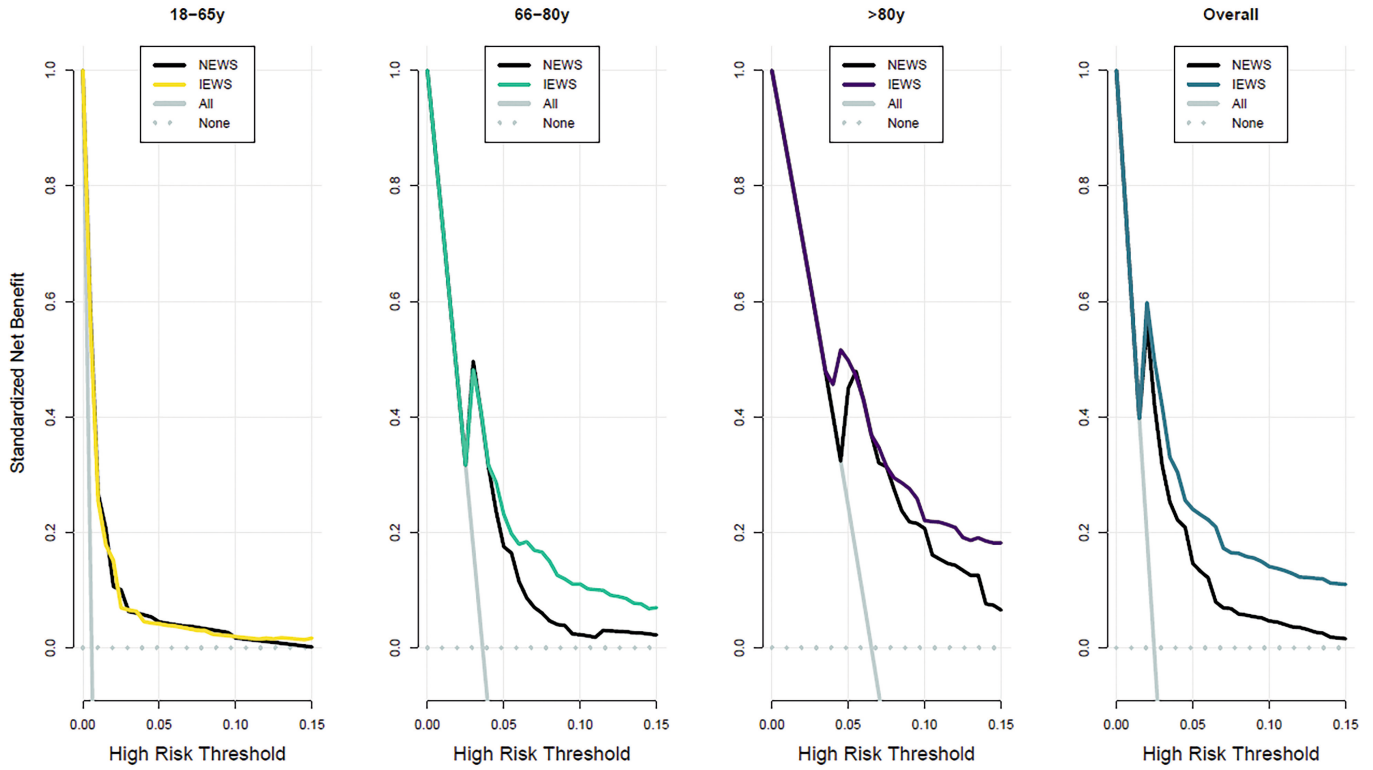
The figure below is a calibration plot for NEWS compared to IEWS stratified by hospital location.



Hospital 1 (yellow), Hospital 2 (green), hospital 3 (purple)

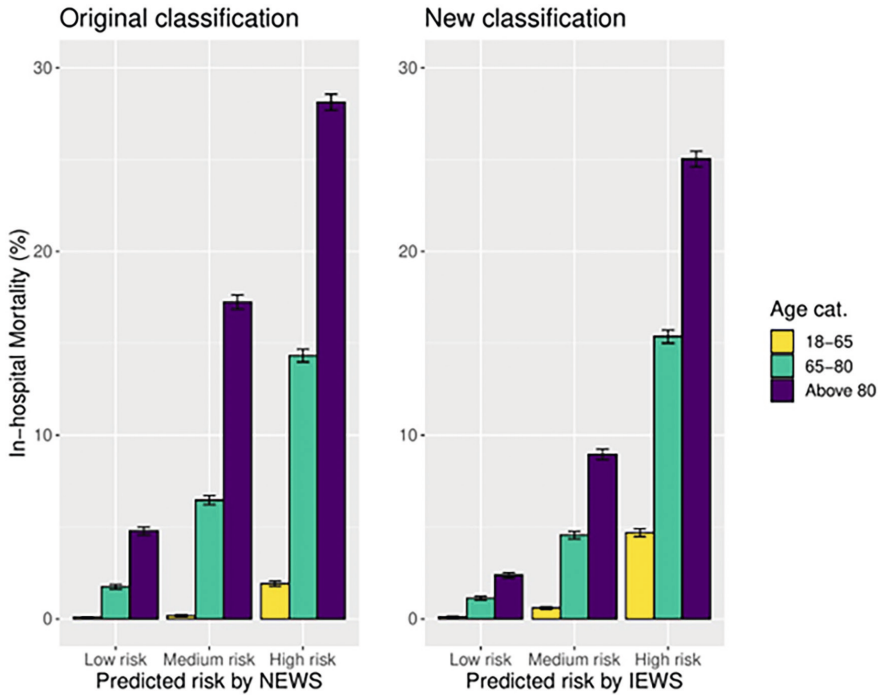
SUPPLEMENTARY FILE 12

Decision Curve Analysis in the DMC



SUPPLEMENTARY FILE 13

Reclassification figure and table for the development cohort.
Because no clear risk thresholds exist, the baseline risk was calculated for each age category by the average mortality risk for patients with a NEWS<4. Patients were considered as lows risk if mortality was lower than 2*baseline risk, medium risk if mortality was between 2*baseline risk and 3*baseline risk, and high risk if mortality was more than 3*baseline risk.
18-65y Low risk <0.68%, medium risk 0.68-1.02% and high risk >1.02%
66-80y Low risk <2.94%, medium risk 2.94-4.41% and high risk >4.41%
>80y Low risk <5.96%, medium risk 5.96-8.94% and high risk >8.94%



	Classification by NEWS			Classification by IEWS		
	Low risk (N45218)	**Medium Risk (N=16657)**	**High risk (N=34178)**	**Low risk (N=66700)**	**Medium risk (N=10588)**	**High risk (N=18765)**
Age category						
18-65y (alive)	9359 (20.7%)	13735 (82.5%)	27991 (81.9%)	34486 (51.7%)	6355 (60.0%)	10244 (54.6%)
18-65y (died)	8 (0.0%)	23 (0.1%)	546 (1.6%)	36 (0.1%)	38 (0.4%)	503 (2.7%)
66-80y (alive)	22631 (50.0%)	2115 (12.7%)	3864 (11.3%)	21379 (32.1%)	2704 (25.5%)	4527 (24.1%)
66-80y (died)	404 (0.9%)	146 (0.9%)	646 (1.9%)	245 (0.4%)	129 (1.2%)	822 (4.4%)
>80y (alive)	12203 (27.0%)	528 (3.2%)	813 (2.4%)	10303 (15.4%)	1240 (11.7%)	2001 (10.7%)
>80y (died)	613 (1.4%)	110 (0.7%)	318 (0.9%)	251 (0.4%	122 (1.2%)	668 (3.6%)

SUPPLEMENTARY FILE 14

Multivariable logistic regression for missing glasgow coma scale**:** A multivariable logistic regression model demonstrated that missing glasgow coma scale is associated with physiological variables and outcome (Chi-square p<0.01), and thus GCS is not missing completely at random or missing not at ramdom and is suitable for multiple imputation. Similar Chi-square p-values were found for the other physiological variables.
Logistic Regression Model
lrm(formula = is.na(gcs) ~ ihm + sbp + saturation + sex + triage +
age, data = NEED2)

Ratio Test
Obs 82306
FALSE 9290
TRUE 73016 Pr(> chi2) <0.0001

Coef	S.E.		Wald Z Pr(>|Z|)
Intercept	2.7829 0.3175	8.77	<0.0001
ihm	-0.2334 0.0630	-3.70	0.0002
sbp	-0.0095 0.0004	-26.49	<0.0001
saturation	0.0032 0.0012	0.37	0.7112
sex=female	0.0375 0.0225	1.67	0.0954
triage=very urgent	0.4236 0.0394	10.76	<0.0001
triage=urgent	0.9677 0.0397	24.37	<0.0001
triage=non-urgent	1.6921 0.0495	34.17	<0.0001
age	-0.0058 0.0007	-8.78	<0.0001

Frequencies of Missing Values Due to Each Variable
is.na(gcs) ihm sbp saturation sex triage age
0 2154 59317 57074 2 6181 0

SUPPLEMENTARY FILE 15

Nomogram for vital signs used as restricted cubic splines. Extremes for each vital sign were excluded.