Original Article

The accuracy of the National Early Warning Score 2 in predicting early death in prehospital and emergency department settings: a systematic review and meta-analysis

Shengfeng Wei[#], Dan Xiong[#], Jia Wang[#], Xinmeng Liang, Jingxian Wang, Yuee Chen

Department of Emergency Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Contributions: (I) Conception and design: All authors; (II) Administrative support: S Wei, D Xiong, Jia Wang; (III) Provision of study materials or patients: X Liang, Jingxian Wang, Y Chen; (IV) Collection and assembly of data: S Wei, D Xiong, Jia Wang; (V) Data analysis and interpretation: X Liang, Jingxian Wang, Y Chen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Xinmeng Liang; Jingxian Wang; Yuee Chen. Department of Emergency Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. Email: Liangxm29@mail.sysu.edu.cn; 540946583@qq.com; chenye25@mail.sysu.edu.cn.

Background: Many studies have explored the accuracy of the National Early Warning Score 2 (NEWS2) in predicting mortality in prehospital and emergency settings, but their findings are inconsistent. Whether NEWS2 is reliable for the pre-examination and triage of patients in prehospital settings and emergency departments remains debatable. Hence, this study aimed to evaluate the accuracy of NEWS2 in predicting mortality in prehospital settings and emergency departments.

Methods: We searched PubMed, Embase, Cochrane Library, Web of Science, CNKI, Wan Fang Data, Vip Database and SinoMed from the inception of each database to January 2023. The inclusion criteria: (I) patients in the prehospital settings or emergency departments; (II) the NEWS2 for predicting 2-day mortality, 30-day mortality, and in-hospital mortality; (III) sufficient data, such as sensitivity, specificity, overall survival, and deaths, were provided for the study; (IV) the type of study was accuracy prediction study. Two authors independently extracted data, including authors, year of publication, country of origin, study design, sample size, threshold cutoff values of NEWS2, and mortality. The PROBAST was used to assess the risk of bias in the included studies.

Results: Thirty studies with 185,835 participants were included. Among the 30 included studies, 13 have a high risk of bias, and 17 have a low risk of bias. The pooled sensitivity, specificity and AUC of 2-day mortality (early mortality), 30-day mortality and in-hospital mortality were 0.81 *vs.* 0.76 *vs.* 0.72 (95% CI: 0.61, 0.80), 0.81 *vs.* 0.69 *vs.* 0.78 (95% CI: 0.49, 0.93) and 0.88 *vs.* 0.80 *vs.* 0.78 (95% CI: 0.74, 0.82), respectively.

Conclusions: NEWS2 has excellent sensitivity and specificity in predicting early mortality in patients in the prehospitals setting and emergency departments. Nonetheless, it has poor performance in predicting in-hospital mortality and 30-day mortality. Our findings underpin the use of NEWS2 as a pre-examination and triage tool to predict early death in the prehospital settings and emergency departments. To improve the predictive accuracy, it should be used to monitor patients continuously rather than at a single point-in-time.

Keywords: National Early Warning Score 2 (NEWS2); prehospital; systematic review; meta-analysis

Submitted Dec 16, 2022. Accepted for publication Jan 11, 2023. Published online Jan 31, 2023.
doi: 10.21037/atm-22-6587
View this article at: https://dx.doi.org/10.21037/atm-22-6587

Introduction

The prehospital setting and emergency rooms are areas with high demand for emergency care, where patients are characterized by critical conditions, multiple disease comorbidities, sudden onset, and rapid change of condition (1-6). Recognizing and responding to clinical deterioration is a priority for patient safety, as well as for emergency care research (7-9). Numerous studies have shown that during emergency care for deteriorating patients, failure to recognize early symptoms and provide intervention is associated with an increase in high mortality adverse events (10-14). Many scoring systems, such as Early Warning Score (EWS), National Early Warning Score (NEWS), and Modified Early Warning Score (MEWS), are used to perform a simple, rapid, effective, and accurate on-site assessment of patients to judge their severity of illness. However, these scoring systems have shown many shortcomings over time (15). Thus, it is necessary to introduce newer and improved triage systems. To optimize the initial treatment management of patients, ensure the reasonable allocation of resources, and reduce the incidence of mortality.

In December 2017, the National Early Warning Score 2 (NEWS2) was published by the Royal College of Physicians (RCP) as an improved update to NEWS 2012. In January 2019, NEWS2 was rolled out across the National Health Service (NHS) in England (16). NEWS2 contains six physiological parameters, and each parameter is scored from 0 (the least severe) to 3 (the most severe). Compared to NEWS, NEWS2 provides a better prediction of exacerbation in patients with hypercapnia respiratory failure. In chronic obstructive pulmonary disease (COPD) patients with hypercapnia, the use of an oxygen saturation metric score scale remains controversial (17).

The EWS score has been used in multiple health care settings, including hospital wards, emergency departments, and the prehospital community (15). Many studies have explored the accuracy of the National Early Warning Score 2 (NEWS2) in predicting mortality in prehospital and emergency settings, but their findings are inconsistent. Medina-Lozano *et al.* (18), for instance, found that the sensitivity, specificity, and area under curve (AUC) of NEWS2 in predicting 2-day mortality were 1.0, 0.89, and 0.962, respectively. On the contrary, Martín-Rodríguez *et al.* (19) reported that the sensitivity, specificity, and AUC of NEWS2 in predicting 2-day mortality were 0.67, 0.75, and 0.756, respectively. The inconsistent findings may be attributable to their different cut-off values of NEWS2. The former adopted a cut-off value of 8 points, whereas the latter used a 11-point cut-off. Whether NEWS2 is reliable for the pre-examination and triage of patients in the prehospital settings and emergency departments remains debatable. This systematic review and meta-analysis aimed to confirm and describe the sensitivity, specificity, positive likelihood ratio (PLR), negative likelihood ratio (NLR), diagnostic odds ratio (DOR), and AUC of NEWS2 for 2-day mortality, 30-day mortality, and in-hospital mortality in the prehospital setting and emergency department at different 'cutoff' values. We present the following article in accordance with the PRISMA-DTA reporting checklist (20) (available at https://atm.amegroups.com/article/view/10.21037/atm-22-6587/rc).

Highlight box

Key findings

- NEWS2 has excellent sensitivity and specificity in predicting 2-day mortality, but a poor sensitivity and specificity for predicting in-hospital, 30-day mortality in the prehospital setting and emergency room.

What is known and what is new?

- EWS, NEWS, and MEWS *et al.* are used to perform a simple, rapid, effective, and accurate on-site assessment of patients to judge their severity of illness.
- Based on the updated version of NEWS2, we aimed to confirm and describe the sensitivity, specificity, PLR, NLR, DOR, and AUC of NEWS2 for 2-day mortality, 30-day mortality, and in-hospital mortality in the prehospital setting and emergency department at different 'cutoff' values.

What is the implication, and what should change now?

- Our results support the use of NEWS2 as a tracking and triggering aid in the assessment of conditions and the allocation of emergency resources when prescreening and triaging patients in prehospital and emergency settings.

Methods

Study design

A predefined protocol has been registered in PROSPERO (CRD42022377935).

Study selection and data extraction

We systematically searched PubMed, Embase, Cochrane Library, Web of Science, CNKI, Wan Fang Data, Vip

Database and SinoMed from the inception of each database to January 2023. The specific search strategies are shown in Appendix 1. All studies were screened through EndNote X9. Two authors independently removed duplicates, screened titles, abstracts, and full texts, and agreed on final study eligibility.

The basic inclusion criteria of the literature search included the following: (I) patients in the prehospital or emergency department area were recruited by the study; (II) the NEWS2 for predicting 2-day mortality, 30-day mortality, and in-hospital mortality was applied by the study; (III) sufficient data, such as sensitivity, specificity, overall survival, and deaths, were provided for the study; and (IV) the type of study was accuracy prediction study. The exclusion criteria were as follows: (I) the article was written in a language other than English; (II) insufficient data in the study to calculate the true positive (TP), false-positive (FP), false-negative (FN) and true negative (TN) results; (III) letters, case reports, conferences, meta-analysis, and reviews; and (IV) NEWS2 limited to a composite outcome [e.g., combination of in-hospital mortality with intensive care unit (ICU) admission, adverse outcomes], and (V) the study subjects were animal studies.

Two authors independently extracted data, including authors, year of publication, country of origin, study design, sample size, threshold cutoff values of NEWS2, and mortality (e.g., in-hospital mortality, 2-day mortality, 30-day mortality). If multiple threshold values for NEWS2 were reported in one study, we preferred the maximum value for analyses. If any two researchers had discrepancies during literature screening or data extraction, we resolved them through discussion with the third author.

Risk of bias in the included studies

Two authors independently assessed the risk of bias using the prediction model risk of bias assessment tool (PROBAST). This assessment tool comprises 20 signaling questions in four domains: participants, predictors, outcomes, and analysis. When all 20 questions were answered as yes, the overall risk of bias was rated as low; otherwise, the overall risk of bias was graded at high. Thirteen of the included studies were considered to have a high risk of bias due to inappropriate data sources, such as retrospective cohort studies. When the included studies were of low quality, our pooled data were compromised to some extent. Any disagreement was resolved through discussion. All details of the quality assessment criteria are reported in Appendix 2.

Statistical analysis

When the I^2 was equal to or higher than 50%, a random-effects model was used for data analysis; otherwise, a fixed-effects model was adopted. A two-tailed P value <0.05 indicated a statistical difference. The summary area under the curve (SAUC) was pooled as point estimates with 95% confidence intervals (CIs). The summary point estimates of sensitivity and specificity were illustrated through the summary receiver operating characteristic (SROC) curve. In general, when the AUC is 0.5, the diagnostic test has no diagnostic value, 0.7 to 0.8 is considered as acceptable, 0.8 to 0.9 is considered as excellent, and greater than 0.9 has outstanding accuracy (21). We considered sensitivity and specificity greater than 0.8 as an excellent prediction threshold. A significant heterogeneity may affect our results. Hence, we conducted a subgroup analysis according to different thresholds (≥4 *vs.* ≥9) and studied continents (Europe *vs.* other continents) to explore the source of heterogeneity. We used Deek's test for funnel plot asymmetry to assess publication bias (22). All the statistical analyses were conducted using Stata SE 15.1 (Stata Corp. LD, College Station, Texas, USA).

Results

Included studies and their characteristics

A total of 1,458 articles were identified initially, of which 464 articles were duplicated, and 994 were screened out through titles and abstracts. The remaining 82 were considered for a full-text review. After excluding 52 studies, the remaining 30 original studies were included in the final synthesis (the reasons for exclusion are given in *Figure 1*).

All characteristics of the 30 included studies involving 185,835 participants are shown in *Table 1*. Among all the included studies, the lowest cutoff value of NEWS2 was greater than 1, and the highest cutoff value of NEWS2 was greater than 11. Three studies were from Asia, and the countries were China (24), India (26), and Japan (33). Two studies were from North America, including Canada (31) and United States (32). Twenty-five studies were from Europe, and the countries were the United Kingdom, Italy, Spain, and Sweden, of which five studies (23,28-30,34)

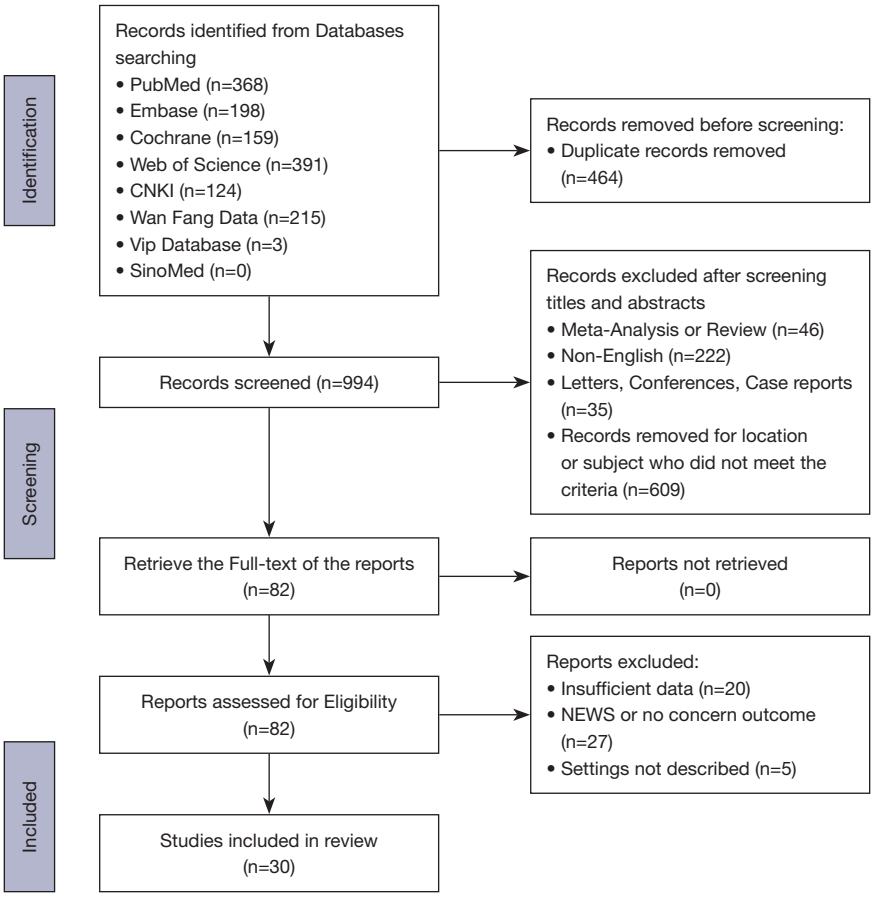


Figure 1 Flow chart of study inclusion.

belonged to the United Kingdom, fifteen (18-19,27,35-40, 42-43,47-50) to Spain, two (25,,45) to Italy, and three (41,44,46) to Sweden. In addition, seventeen studies (18-19, 26-27,35,37-44,46-50) were prospective, and the other 13 studies were retrospective. All included studies took place from 2019 to 2022.

Quality assessment

The summary of quality assessment using PROBAST is shown in *Table 2*. Overall, 13 retrospective cohort studies (23-25,28-34,36,41,45) had a high risk of bias, principally because we assumed that subjects had systematic differences in the accuracy of reporting past information, resulting in recall bias (51,52). The details of the quality assessment are recorded in Appendix 2.

The results of synthesis

The forest plots of sensitivity, specificity, PLR, NLR, and diagnostic odds ratio (DOR) for NEWS2 are illustrated in *Figures 2-5* and show the summary ROC (SROC) curves for NEWS2. Overall, the pooled sensitivity, specificity, DOR, and AUC of 2-day mortality were 0.81 (95% CI: 0.76, 0.84), 0.81 (95% CI: 0.78, 0.84), 18 (95% CI: 12, 26), and 0.88 (95% CI: 0.85, 0.90), respectively (*Table 3*). The pooled sensitivity, specificity, DOR, and AUC of 30-day mortality were 0.76 (95% CI: 0.68, 0.83), 0.69 (95% CI: 0.59, 0.78), 7 (95% CI: 6, 9), and 0.80 (95% CI: 0.76, 0.83), respectively. For in-hospital mortality, the pooled sensitivity, specificity, DOR, and AUC were 0.72 (95% CI: 0.61, 0.80), 0.78 (95% CI: 0.49, 0.93), 9 (95% CI: 3, 28), and 0.78 (95% CI: 0.74, 0.82), respectively.

Table 1 The characteristics of the included studies

References	Year	Country	Design	Sample size	Cutoff	Outcome
Medina-Lozano *et al.* (18)	2020	Spain	Prospective	346	8	2-day mortality
Martín-Rodríguez *et al.* (19)	2020	Spain	Prospective	3,081	11	2-day mortality
Marincowitz *et al.* (23)	2022	England	Retrospective	7,549	1	30-day mortality
Hu *et al.* (24)	2022	China	Retrospective	319	10	In-hospital mortality
Guarino *et al.* (25)	2022	Italy	Retrospective	437	7	In-hospital mortality, 30-day mortality
Chikhalkar *et al.* (26)	2022	Indian	Prospective	814	9	In-hospital mortality
Villanueva Rábano *et al.* (27)	2021	Spain	Prospective	638	10	2-day mortality
					9	30-day mortality
Thomas *et al.* (28)	2021	UK	Retrospective	20,891	4	30-day mortality
Sivayoham *et al.* (29)	2021	UK	Retrospective	2,594	8	In-hospital mortality
Richardson *et al.* (30)	2021	England	Retrospective	6,444	5	In-hospital mortality, 2-day mortality
Reardon *et al.* (31)	2021	Canada	Retrospective	4,022	5	In-hospital mortality
Prasad *et al.* (32)	2021	America	Retrospective	23,837	5	In-hospital mortality
Osawa *et al.* (33)	2021	Japan	Retrospective	2,900	6	In-hospital mortality
Masson *et al.* (34)	2021	England	Retrospective	91,871	5	2-day mortality, 30-day mortality
Martín-Rodríguez *et al.* (35)	2021	Spain	Prospective	3,273	7	2-day mortality
Martín-Rodríguez *et al.* (36)	2021	Spain	Retrospective	663	7	2-day mortality
López-Izquierdo *et al.* (37)	2021	Spain	Prospective	941	8	30-day mortality
Durantez-Fernández *et al.* (38)	2022	Spain	Prospective	1,716	6.5	2-day mortality
					5.5	30-day mortality
Durantez-Fernández *et al.* (39)	2021	Spain	Prospective	445	6	2-day mortality
					5	30-day mortality
Clar *et al.* (40)	2021	Spain	Prospective	201	5	In-hospital mortality
Mellhammar *et al.* (41)	2020	Sweden	Retrospective	941	5	30-day mortality
Martín-Rodríguez *et al.* (42)	2020	Spain	Prospective	209	10	2-day mortality
Martín-Rodríguez *et al.* (43)	2020	Spain	Prospective	2,335	9	2-day mortality
					7	30-day mortality
Magnusson *et al.* (44)	2020	Sweden	Prospective	4,465	5	2-day mortality, 30-day mortality
Covino *et al.* (45)	2020	Italy	Retrospective	334	4	2-day mortality
Mellhammar *et al.* (46)	2019	Sweden	Prospective	1,171	5	30-day mortality
Martín-Rodríguez *et al.* (47)	2019	Spain	Prospective	1,054	9	2-day mortality
Martín-Rodríguez *et al.* (48)	2019	Spain	Prospective	707	9	2-day mortality
					8	30-day mortality
Martín-Rodríguez *et al.* (49)	2019	Spain	Prospective	349	10	2-day mortality
Martín-Rodríguez *et al.* (50)	2019	Spain	Prospective	1,288	9	2-day mortality

Table 2 The PROBAST results

References	ROB				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Medina-Lozano *et al.* 2020 (18)	+	+	+	+	+	+	+	+	+
Martín-Rodríguez *et al.* 2020 (19)	+	+	+	+	+	+	+	+	+
Marincowitz *et al.* 2022 (23)	–	+	+	+	+	+	+	–	+
Hu *et al.* 2022 (24)	–	+	+	+	+	+	+	–	+
Guarino *et al.* 2022 (25)	–	+	+	+	+	+	+	–	+
Chikhalkar *et al.* 2022 (26)	+	+	+	+	+	+	+	+	+
Villanueva Rábano *et al.* 2021 (27)	+	+	+	+	+	+	+	+	+
Thomas *et al.* 2021 (28)	–	+	+	+	+	+	+	–	+
Sivayoham *et al.* 2021 (29)	–	+	+	+	+	+	+	–	+
Richardson *et al.* 2021 (30)	–	+	+	+	+	+	+	–	+
Reardon *et al.* 2021 (31)	–	+	+	+	+	+	+	–	+
Prasad *et al.* 2021 (32)	–	+	+	+	+	+	+	–	+
Osawa *et al.* 2021 (33)	–	+	+	+	+	+	+	–	+
Masson *et al.* 2021 (34)	–	+	+	+	+	+	+	–	+
Martín-Rodríguez *et al.* 2021 (35)	+	+	+	+	+	+	+	+	+
Martín-Rodríguez *et al.* 2021 (36)	–	+	+	+	+	+	+	–	+
López-Izquierdo *et al.* 2021 (37)	+	+	+	+	+	+	+	+	+
Durantez-Fernández *et al.* 2022 (38)	+	+	+	+	+	+	+	+	+
Durantez-Fernández *et al.* 2021 (39)	+	+	+	+	+	+	+	+	+
Clar *et al.* 2021 (40)	+	+	+	+	+	+	+	+	+
Mellhammar *et al.* 2020 (41)	–	+	+	+	+	+	+	–	+

Table 2 (*continued*)

Table 2 (*continued*)

References	ROB				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Martín-Rodríguez *et al.* 2020 (42)	+	+	+	+	+	+	+	+	+
Martín-Rodríguez *et al.* 2020 (43)	+	+	+	+	+	+	+	+	+
Magnusson *et al.* 2020 (44)	+	+	+	+	+	+	+	+	+
Covino *et al.* 2020 (45)	–	+	+	+	+	+	+	–	+
Mellhammar *et al.* 2019 (46)	+	+	+	+	+	+	+	+	+
Martín-Rodríguez *et al.* 2019 (47)	+	+	+	+	+	+	+	+	+
Martín-Rodríguez *et al.* 2019 (48)	+	+	+	+	+	+	+	+	+
Martín-Rodríguez *et al.* 2019 (49)	+	+	+	+	+	+	+	+	+
Martín-Rodríguez *et al.* 2019 (50)	+	+	+	+	+	+	+	+	+

'+' represents low ROB/low concern regarding applicability; '–' represents high ROB/high concern regarding applicability. PROBAST, Prediction model Risk of Bias Assessment; ROB, risk of bias.

Subgroup analysis

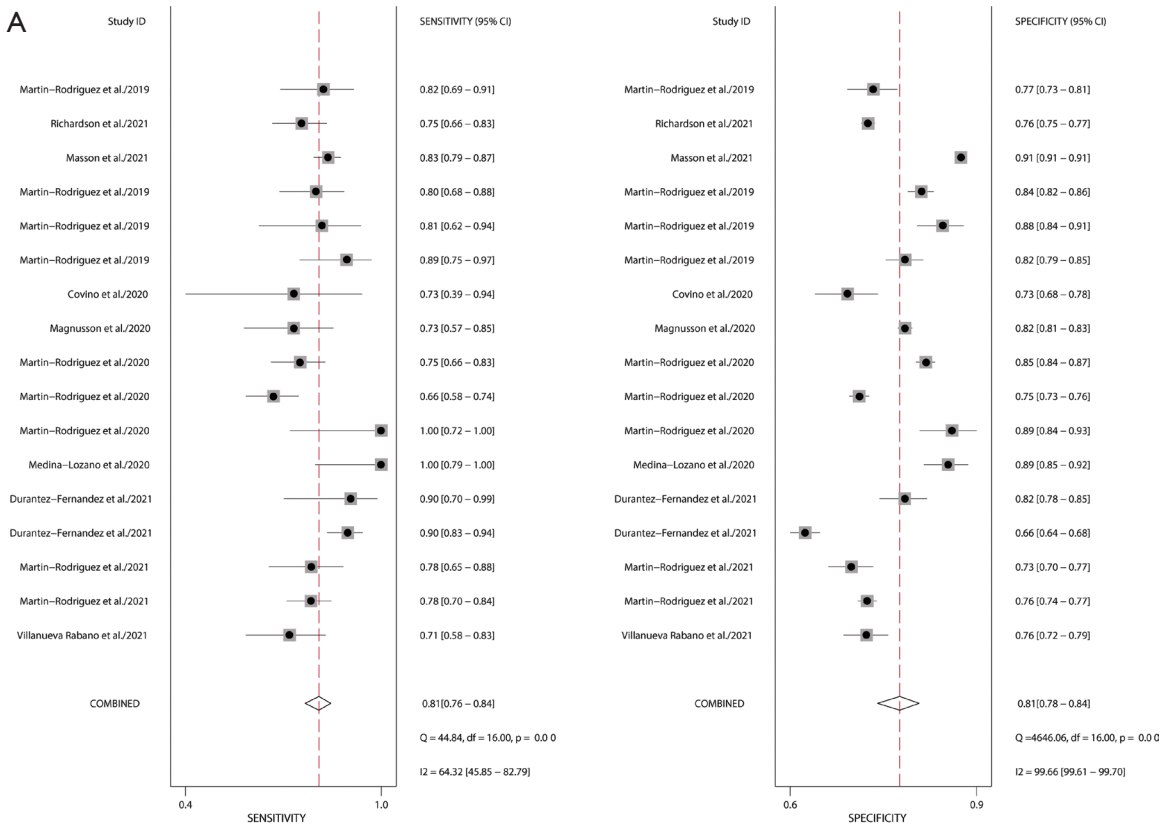
There is relevant evidence that the prognostic performance of NEWS2 is not significantly different in different subgroups (*Table 3*). The sensitivity and specificity of the NEWS2 in predicting early mortality (2-day mortality) in prehospital and emergency settings are high, with excellent accuracy. For example, in 9 studies using a threshold ≥4, the pooled sensitivity, specificity and AUC were 0.82 (95% CI: 0.77, 0.86), 0.80 (95% CI: 0.74, 0.85), and 0.88 (95% CI: 0.84, 0.90); in the 8 studies using a threshold ≥9, the combined sensitivity, specificity, and AUC were 0.78 (95% CI: 0.71, 0.84), 0.83 (95% CI: 0.79, 0.86), and 0.87 (95% CI: 0.84, 0.90). In addition, from the pooled data of NEWS2 in different continents for predicting in-hospital mortality in prehospital and emergency settings, the accuracy rate in Europe and other continents is acceptable, with a similar AUC (0.74 *vs.* 0.76). Among them, the European study had a low sensitivity of 0.77 (95% CI: 0.55, 0.90) and a poor specificity of 0.61 (95% CI: 0.40, 0.78); studies from other continents (Asia, North America) had a low sensitivity of 0.69 (95% CI: 0.59, 0.78) and an outstanding specificity of 0.90 (95% CI: 0.39, 0.99).

The results of publication bias

Figure 6 shows the results of publication bias by using Deeks' funnel plot asymmetry test. The P values of NEWS2 for patients in 2-day, 30-day, and in-hospital mortality were 0.98, 0.99, and 0.07, respectively. This indicates that there was no significant publication bias.

Discussion

Throughout this systematic review and meta-analysis, we found that the AUC curve of 2-day mortality in the emergency department and the prehospital settings ranged from 0.85 to 0.90. The AUC curves of in-hospital mortality and 30-day mortality ranged from 0.74 to 0.82 and 0.76 to 0.83, respectively. NEWS2 is relatively reliable



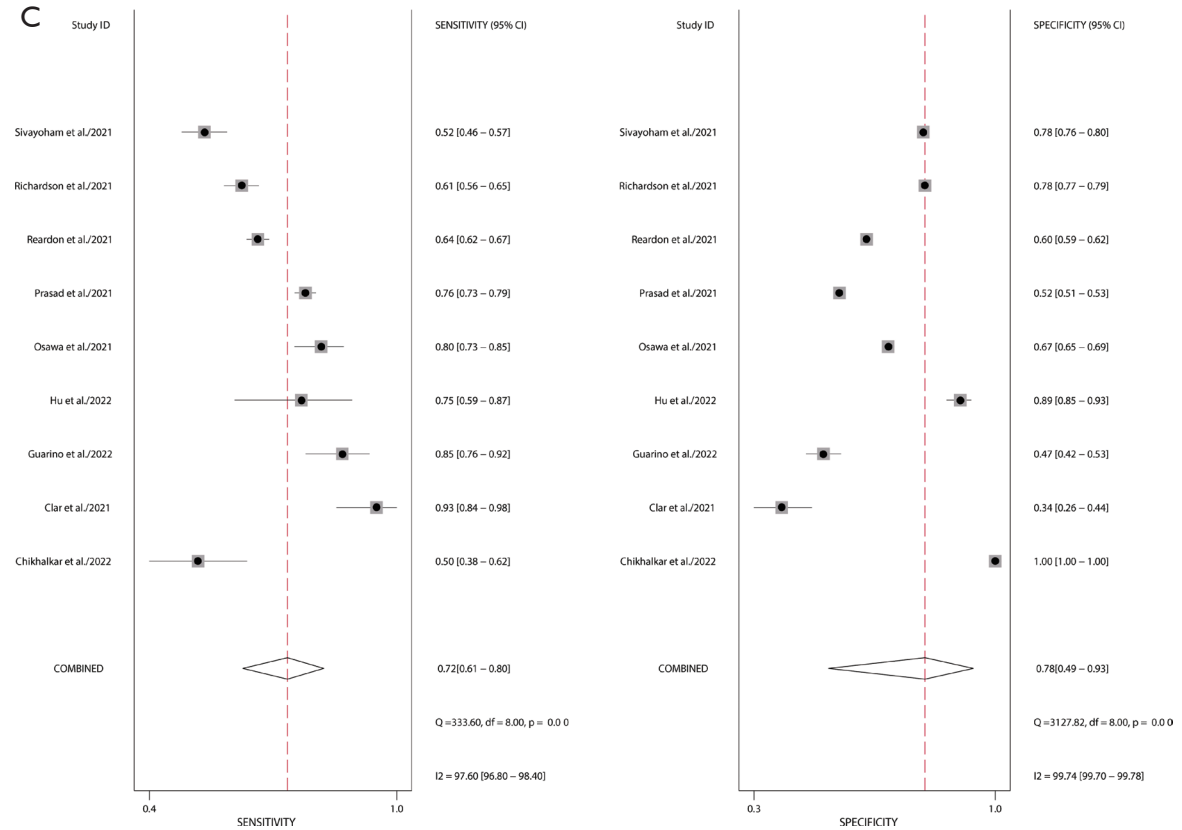
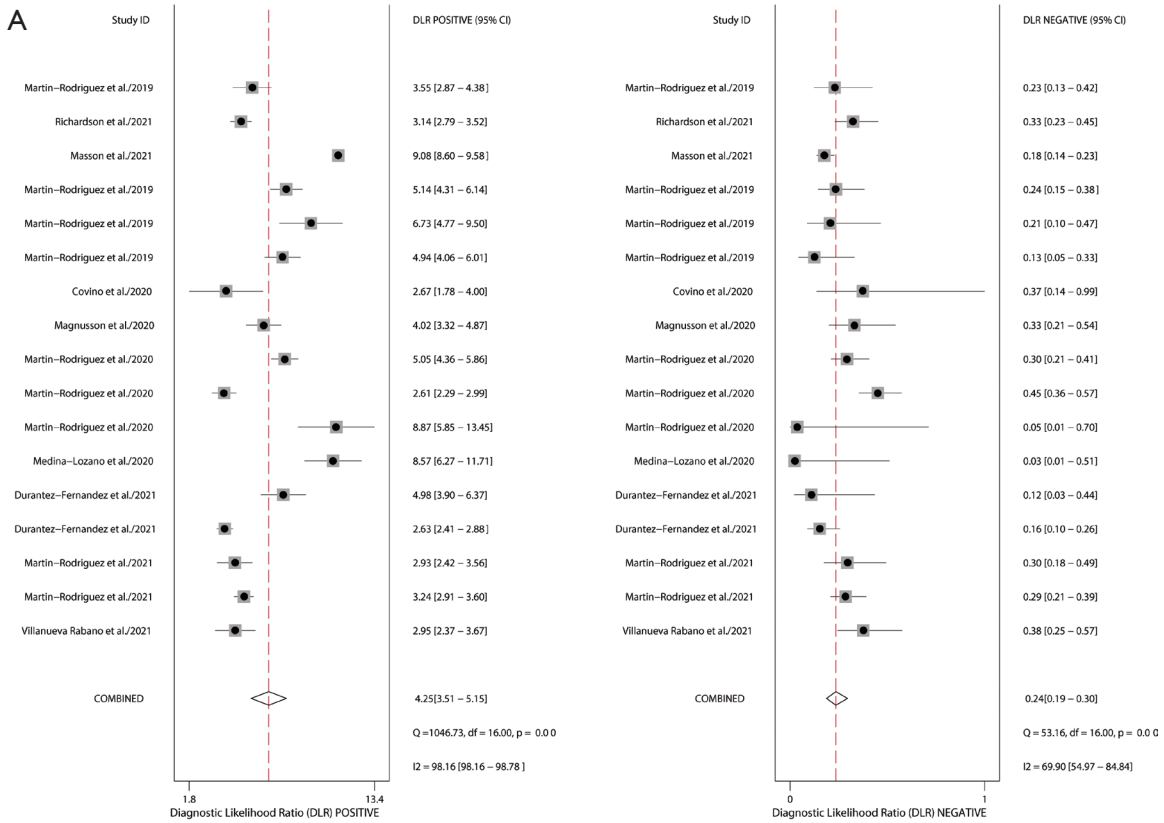
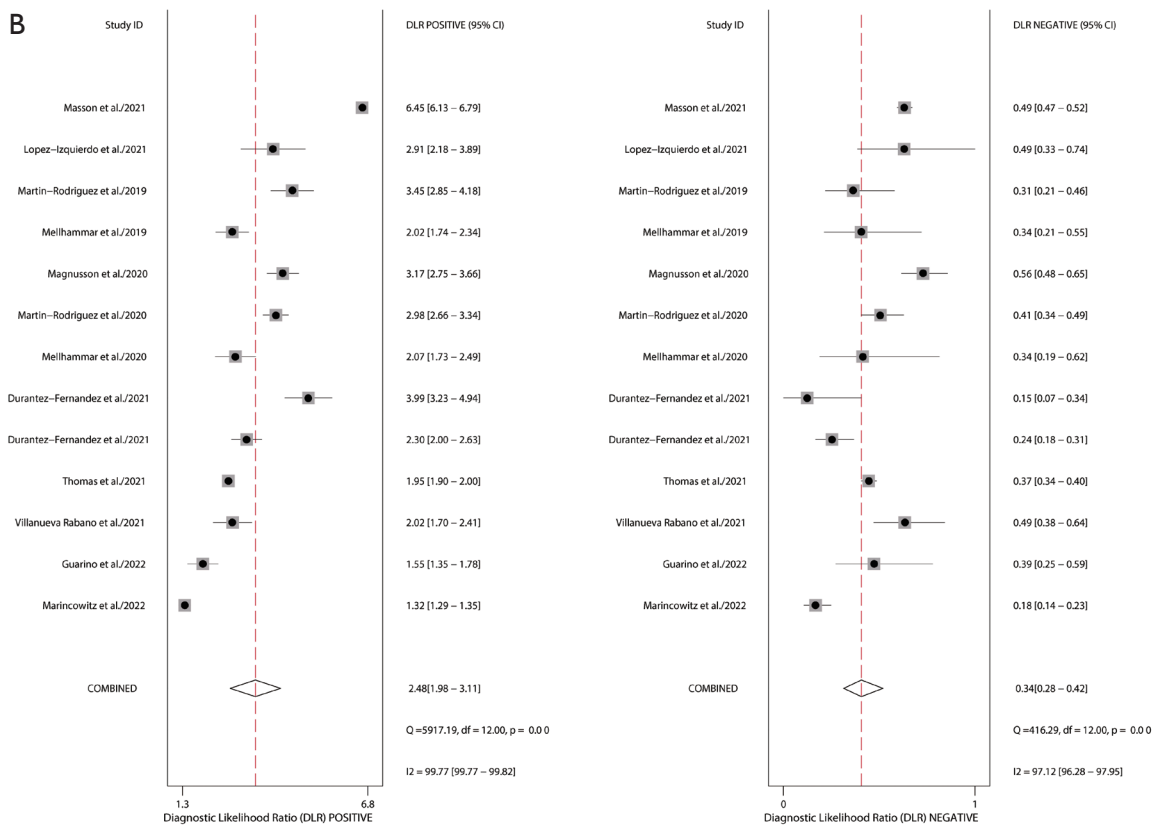


Figure 2 Forest plots of sensitivity and specificity for NEWS2. (A) 2-day mortality; (B) 30-day mortality; (C) in-hospital mortality. NEWS2, National Early Warning Score 2.

in identifying early mortality (2-day mortality) in patients in the emergency department and prehospital areas. Analysis of the data shows that the accuracy of NEWS2 in predicting the abovementioned adverse outcomes is acceptable or even excellent. Thus, our results support the use of the NEWS2 as a tracking and triggering aid in the assessment of conditions and the allocation of emergency resources when prescreening and triaging patients in prehospital and emergency settings, especially in crowded emergency rooms (53). In addition, our results also show that NEWS2 is highly sensitive (0.82) in predicting 2-day mortality for results with a threshold ≥4, while the sensitivity of NEWS2 (0.78) decreases in predicting 2-day mortality for results with a threshold ≥9. This means that for patients with a NEWS2 score ≥4, we should increase clinical attention, identify patients with high-risk factors in the population, and provide early intervention as soon as possible to improve the prognosis. NEWS2 is rather stable in predicting the in-hospital mortality rate across different continents, and there is no obvious difference in accuracy.

The pooled results showed significant heterogeneity among the included studies, where I^2>50% represented significant heterogeneity (54). The large sample size gap, the different study designs (prospective and retrospective), and the methods of registering the population may be sources of heterogeneity. In addition, we considered that the study location might be a confounding factor for various health care systems, which could influence clinical outcomes. Heterogeneity may also arise due to the various time windows between score calculation and outcome measurement. Since early warning score systems have been introduced in many United Kingdom (UK) hospitals, they have been used for a wide variety of patients clinically and associated with relevant clinical responses (55). Some urgent patients are likely to receive rapid medical care after triggering the alert. The actual death rate tends to be lower





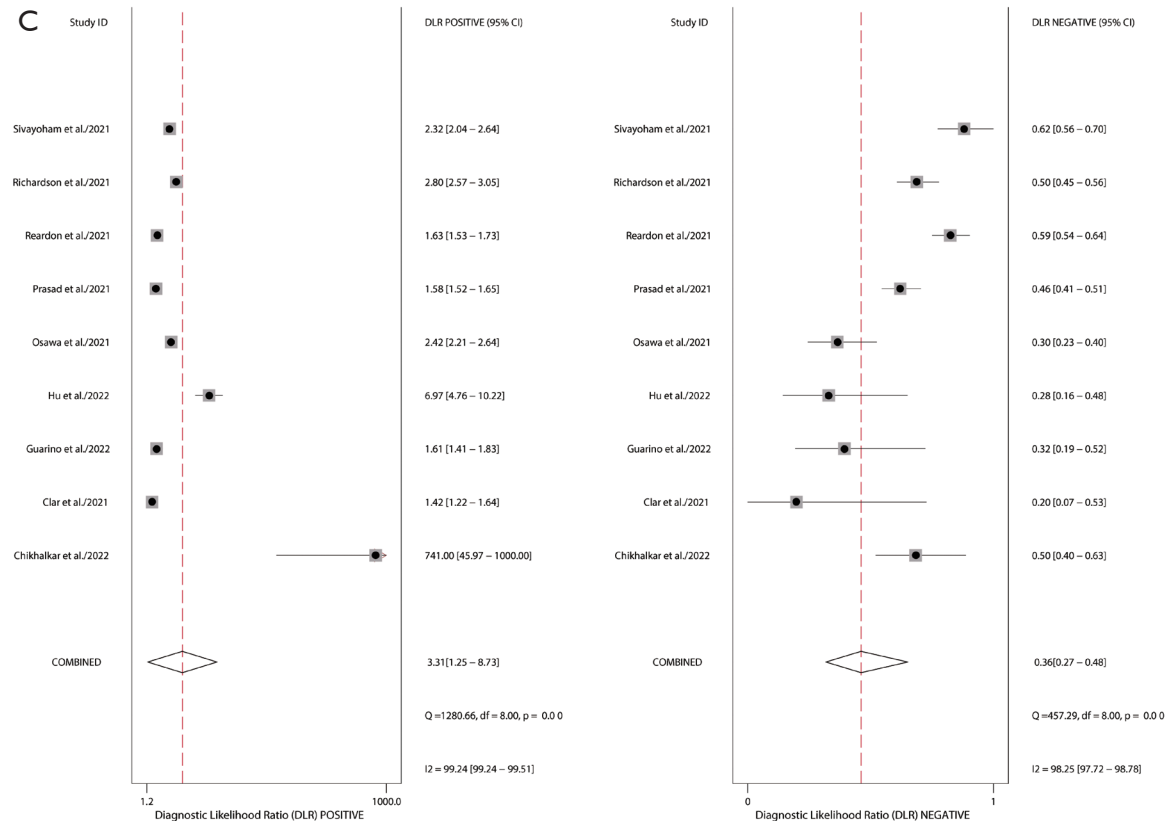
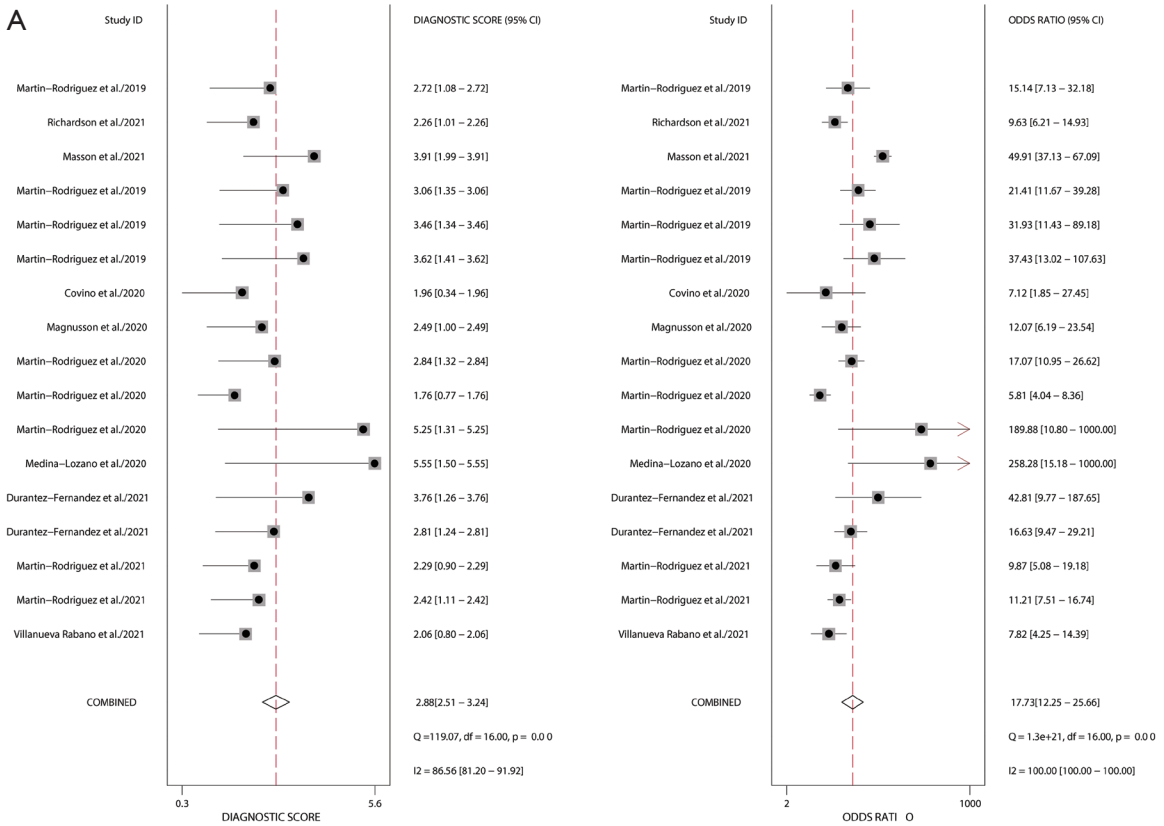


Figure 3 The positive and negative likelihood ratios for NEWS2. (A) 2-day mortality; (B) 30-day mortality; (C) in-hospital mortality. NEWS2, National Early Warning Score 2.

than the predicted rate, which may bias our estimate of accuracy and lead to heterogeneity. Furthermore, our study included prehospital and emergency settings, which are different, as well as various outcome measures, such as 2-day mortality, 30-day mortality, and in-hospital mortality. The difference in setting and outcome measures may also explain the source of heterogeneity.

Patients with novel coronavirus-infected pneumonia are usually characterized by solitary respiratory failure (56). Moreover, supplemental oxygen has been confirmed as an independent risk factor for the progression of novel coronavirus pneumonia to critical illness (57). Compared to the original NEWS, NEWS2 has similar sensitivity and specificity to NEWS in predicting non-hypercapnic respiratory failure. In predicting hypercapnic respiratory failure, based on the SPO_2 scoring scale specially developed for hypercapnia (58), NEWS2 is better than NEWS. In addition, compared with other scoring systems, such as Early Warning Score (EWS), MEWS, and quick Sepsis related Organ Failure Assessment (qSOFA), the main advantage of NEWS2 is that both hypoxemia and supportive oxygen therapy are included in the scoring parameters. Therefore, although other scoring systems and NEWS2 have good discrimination, sensitivity, and specificity, NEWS2 might be more reliable in prehospital and emergency department settings, especially during the COVID-19 pandemic. In addition, our research suggests that we should activate early medical care for patients with a NEWS2 threshold ≥4 in predicting early mortality (2-day mortality). According to the guidelines of the Royal College of Physicians (59), patients with a NEWS2 score of fewer than 5 points still have the possibility of rapid deterioration, leading to severe respiratory failure. Thus, we need to continuously monitor this subset of patients with a NEWS2 threshold less than 5. Notably, NEWS2 should be utilized to support clinical decision-making by providing objective data, but it should not be an alternative to the clinical judgment of experienced clinicians. Hence, NEWS2 could



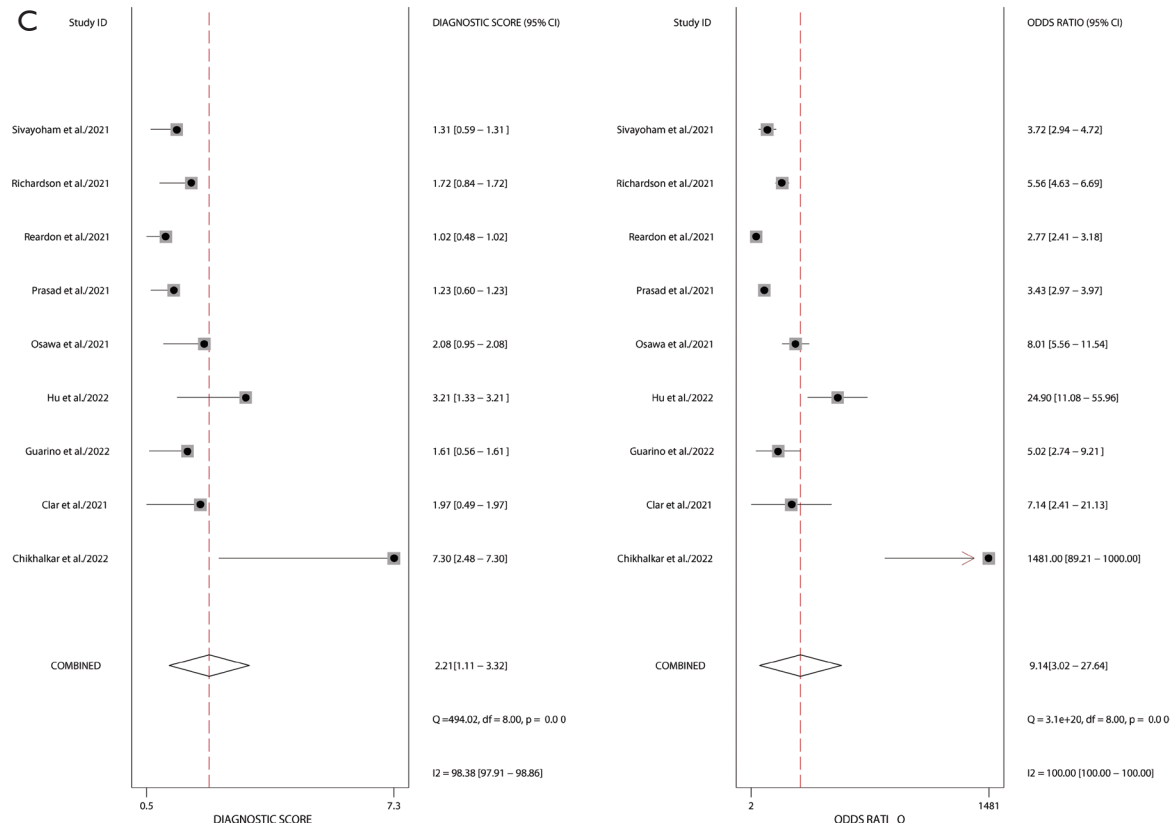


Figure 4 The diagnostic odds ratio for NEWS2. (A) 2-day mortality; (B) 30-day mortality; (C) in-hospital mortality. NEWS2, National Early Warning Score 2.

be used to evaluate a possible deterioration in the patient's condition throughout their hospital stay.

Strengths and limitations of the review

The current meta-analysis has several strengths. First, we included the most recent cohort study data available in this fast-moving domain, including findings during the COVID-19 pandemic from 2019 to 2022. Second, we focused on the triage performance of the NEWS2 scoring system in the prehospital setting and emergency departments, as both locations have the characteristics of diverse patients and diseases. Thus, our conclusion could more typically represent the accuracy of the scoring system.

Nevertheless, there are some crucial limitations in our research. First, there was significant heterogeneity in our study. One-fifth of the included studies had small sample sizes (<400), and the quality of the included studies was not as high as the reliability of large samples. Second, the meta-analysis did not have sufficient data to explore the performance of NEWS2 in patients younger than 18 years. The age of the study subjects was concentrated in adults. Therefore, using NEWS2 on individuals younger than 18 years of age might lead to inaccurate results. Third, more than two-thirds of the included studies in our research were from Europe. We still need more evidence from non-European countries to improve the accuracy and clinical applicability of NEWS2.

Conclusions

This is the first meta-analysis to assess the accuracy of the NEWS2 in predicting in-hospital, 2-day, and 30-day mortality for patients in the prehospital setting and emergency departments. Based on the AUC, sensitivity, and specificity results greater than 0.8 as an excellent prediction

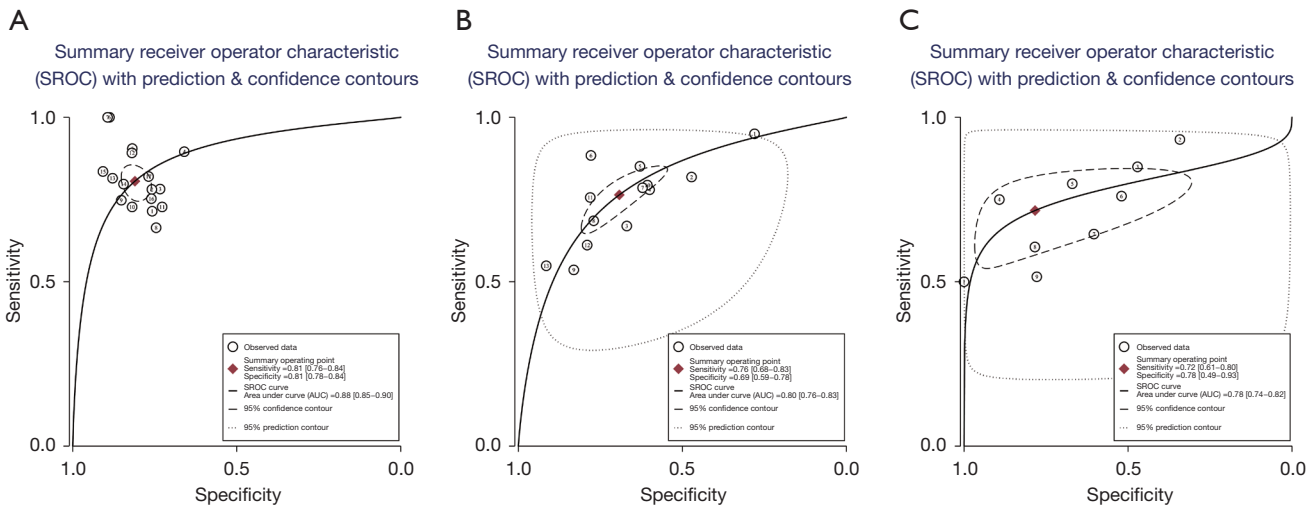


Figure 5 The summary ROC curve for NEWS2 for predicting patients in (A) 2-day, (B) 30-day and (C) in-hospital mortality. ROC curve, receiver operator characteristic curve; NEWS2, National Early Warning Score 2.

Table 3 Results of meta-analysis

Results	N	Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR	AUC (95% CI)
In-hospital mortality	9	0.72 (0.61, 0.80)	0.78 (0.49, 0.93)	3.3 (1.3, 8.7)	0.36 (0.27, 0.48)	9 (3, 28)	0.78 (0.74, 0.82)
2-day mortality	17	0.81 (0.76, 0.84)	0.81 (0.78, 0.84)	4.3 (3.5, 5.2)	0.24 (0.19, 0.30)	18 (12, 26)	0.88 (0.85, 0.90)
30-day mortality	13	0.76 (0.68, 0.83)	0.69 (0.59, 0.78)	2.5 (2.0, 3.1)	0.34 (0.28, 0.42)	7 (6, 9)	0.80 (0.76, 0.83)
Subgroup analysis							
Threshold value (2-day mortality)							
NEWS2 ≥4	9	0.82 (0.77, 0.86)	0.80 (0.74, 0.85)	4.0 (3.1, 5.3)	0.23 (0.17, 0.29)	18 (11, 29)	0.88 (0.84, 0.90)
NEWS2 ≥9	8	0.78 (0.71, 0.84)	0.83 (0.79, 0.86)	4.5 (3.4, 5.8)	0.27 (0.19, 0.37)	17 (10, 29)	0.87 (0.84, 0.90)
Continent (in-hospital mortality)							
Europe	4	0.77 (0.55, 0.90)	0.61 (0.40, 0.78)	2.0 (1.5, 2.7)	0.38 (0.24, 0.61)	5 (4, 7)	0.74 (0.70, 0.78)
Other continents	5	0.69 (0.59, 0.78)	0.90 (0.39, 0.99)	7.1 (0.7, 70.2)	0.34 (0.28, 0.43)	20 (2, 217)	0.76 (0.72, 0.79)

CI, confidence interval; PLR, positive likelihood ratio; NLR, negative likelihood ratio; DOR, diagnostic odds ratio; AUC, area under the curve; NEWS2, National Early Warning Score 2.

threshold, we comprehensively analyzed the above outcomes. Thus, NEWS2 has excellent sensitivity and specificity in predicting early mortality (2-day mortality) and can reliably identify the patients requiring emergency preparing and response. Our findings underpin the use of NEWS2 as a pre-examination and triage tool to predict early death in the prehospital settings and emergency departments. However, it shows poor performance in predicting in-hospital mortality and 30-day mortality. The predictive performance of NEWS2 is more reliable when the cut-off value is ≥4. Nevertheless, with the increase of score, the predictive performance decrease. Ultimately, ongoing clinical attention is warranted, even though patients with a low NEWS2 score have a reduced risk of death for several days. Besides, to improve the predictive accuracy, NEWS2 should be used to monitor patients continuously rather than at a single point-in-time.

In the future, we hope that there will be more large-scale

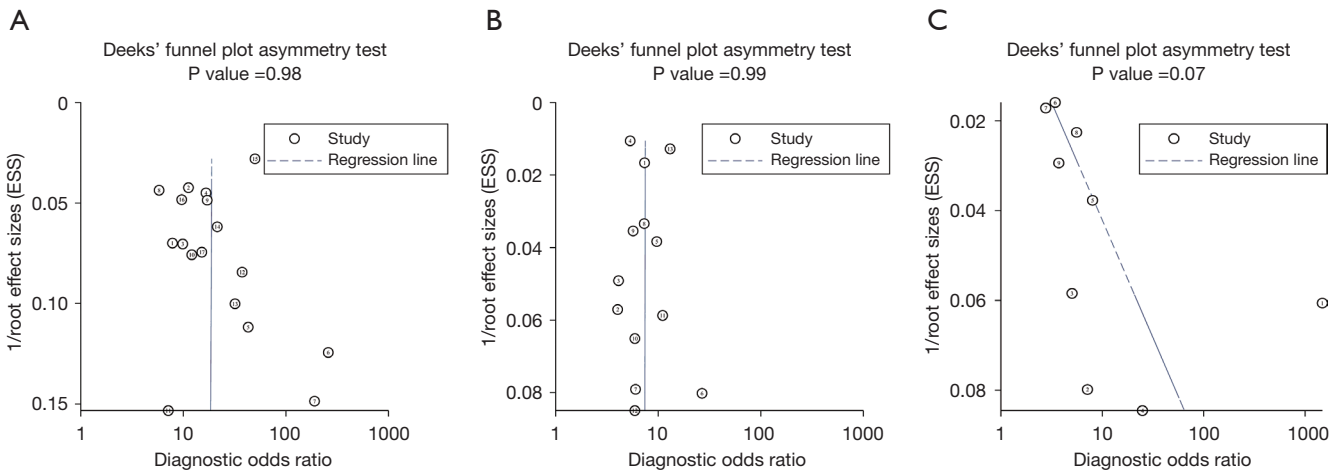


Figure 6 The results of publication bias. (A) 2-day mortality; (B) 30-day mortality; (C) in-hospital mortality.

and high-quality studies on this topic to further inform our results.

Acknowledgments

We would like to thank the researchers and study participants for their contributions.
Funding: None.

Footnote

Reporting Checklist: The authors have completed the PRISMA-DTA reporting checklist. Available at https://atm.amegroups.com/article/view/10.21037/atm-22-6587/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-22-6587/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Considine J, Jones D, Bellomo R. Emergency department rapid response systems: the case for a standardized approach to deteriorating patients. Eur J Emerg Med 2013;20:375-81.
2. Canto JG, Zalenski RJ, Ornato JP, et al. Use of emergency medical services in acute myocardial infarction and subsequent quality of care: observations from the National Registry of Myocardial Infarction 2. Circulation 2002;106:3018-23.
3. Burkholder TW, Hill K, Calvello Hynes EJ. Developing emergency care systems: a human rights-based approach. Bull World Health Organ 2019;97:612-9.
4. Pinto C, Cameron PA, Gabbe B, et al. Trauma case review: A quality and safety feature of the Victorian State Trauma System. Emerg Med Australas 2018;30:125-9.
5. Burrell AR, McLaws ML, Fullick M, et al. SEPSIS KILLS: early intervention saves lives. Med J Aust 2016;204:73.
6. National Stroke Foundation. Clinical guidelines for stroke management 2017. Chapter 1 of 8: Pre-hospital care. Retrieved 05 November 2017. Available online: https://informme.org.au/en/Guidelines/Clinical-Guidelines-for-Stroke-Management-2017. Melbourne: National Stroke Foundation, 2017.

7. Considine J, Curtis K, Shaban RZ, et al. Consensus-based clinical research priorities for emergency nursing in Australia. Australas Emerg Care 2018;21:43-50.
8. Lecky F, Benger J, Mason S, et al. The International Federation for Emergency Medicine framework for quality and safety in the emergency department. Emerg Med J 2014;31:926-9.
9. Keijzers G, Thom O, Taylor D, et al. Clinical research priorities in emergency medicine. Emerg Med Australas 2014;26:19-27.
10. Hayward RA, Asch SM, Hogan MM, et al. Sins of omission: getting too little medical care may be the greatest threat to patient safety. J Gen Intern Med 2005;20:686-91.
11. Stang AS, Wingert AS, Hartling L, et al. Adverse events related to emergency department care: a systematic review. PLoS One 2013;8:e74214.
12. Hagiwara MA, Magnusson C, Herlitz J, et al. Adverse events in prehospital emergency care: a trigger tool study. BMC Emerg Med 2019;19:14.
13. Considine J, Jones D, Pilcher D, et al. Patient physiological status during emergency care and rapid response team or cardiac arrest team activation during early hospital admission. Eur J Emerg Med 2017;24:359-65.
14. Forster AJ, Rose NG, van Walraven C, et al. Adverse events following an emergency department visit. Qual Saf Health Care 2007;16:17-22.
15. Guan G, Lee CMY, Begg S, et al. The use of early warning system scores in prehospital and emergency department settings to predict clinical deterioration: A systematic review and meta-analysis. PLoS One 2022;17:e0265559.
16. National Early Warning Score (NEWS) 2. Standardising the Assessment of Acute-Illness Severity in the NHS: Updated Report of a Working Party. London: Royal College of Physicians, 2017.
17. Smith GB, Redfern OC, Pimentel MA, et al. The National Early Warning Score 2 (NEWS2). Clin Med (Lond) 2019;19:260.
18. Medina-Lozano E, Martín-Rodríguez F, Castro-Villamor MÁ, et al. Accuracy of early warning scores for predicting serious adverse events in pre-hospital traumatic injury. Injury 2020;51:1554-60.
19. Martín-Rodríguez F, López-Izquierdo R, Delgado Benito JF, et al. Prehospital Point-Of-Care Lactate Increases the Prognostic Accuracy of National Early Warning Score 2 for Early Risk Stratification of Mortality: Results of a Multicenter, Observational Study. J Clin Med 2020;9:1156.
20. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA 2018;319:388-96.
21. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010;5:1315-6.
22. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005;58:882-93.
23. Marincowitz C, Sutton L, Stone T, et al. Prognostic accuracy of triage tools for adults with suspected COVID-19 in a prehospital setting: an observational cohort study. Emerg Med J 2022;39:317-24.
24. Hu H, Yao N, Qiu Y. Predictive Value of 5 Early Warning Scores for Critical COVID-19 Patients. Disaster Med Public Health Prep 2022;16:232-9.
25. Guarino M, Perna B, Remelli F, et al. A New Early Predictor of Fatal Outcome for COVID-19 in an Italian Emergency Department: The Modified Quick-SOFA. Microorganisms 2022;10:806.
26. Chikhalkar B, Gosain D, Gaikwad S, et al. Assessment of National Early Warning Score 2 as a Tool to Predict the Outcome of COVID-19 Patients on Admission. Cureus 2022;14:e21164.
27. Villanueva Rábano R, Martín-Rodríguez F, López-Izquierdo R. National Early Warning Score 2 Lactate (NEWS2-L) in Predicting Early Clinical Deterioration in Patients with Dyspnoea in Prehospital Care. Invest Educ Enferm 2021;39:e05.
28. Thomas B, Goodacre S, Lee E, et al. Prognostic accuracy of emergency department triage tools for adults with suspected COVID-19: the PRIEST observational cohort study. Emerg Med J 2021;38:587-93.
29. Sivayoham N, Hussain AN, Shabbo L, et al. An observational cohort study of the performance of the REDS score compared to the SIRS criteria, NEWS2, CURB65, SOFA, MEDS and PIRO scores to risk-stratify emergency department suspected sepsis. Ann Med 2021;53:1863-74.
30. Richardson D, Faisal M, Fiori M, et al. Use of the first National Early Warning Score recorded within 24 hours of admission to estimate the risk of in-hospital mortality in unplanned COVID-19 patients: a retrospective cohort study. BMJ Open 2021;11:e043721.
31. Reardon PM, Seely AJE, Fernando SM, et al. Can Early Warning Systems Enhance Detection of High Risk Patients by Rapid Response Teams? J Intensive Care Med 2021;36:542-9.

32. Prasad PA, Fang MC, Martinez SP, et al. Identifying the Sickest During Triage: Using Point-of-Care Severity Scores to Predict Prognosis in Emergency Department Patients With Suspected Sepsis. J Hosp Med 2021;16:453-61.
33. Osawa I, Sonoo T, Soeno S, et al. Clinical performance of early warning scoring systems for identifying sepsis among anti-hypertensive agent users. Am J Emerg Med 2021;48:120-7.
34. Masson H, Stephenson J. Investigation into the predictive capability for mortality and the trigger points of the National Early Warning Score 2 (NEWS2) in emergency department patients. Emerg Med J 2022;39:685-90.
35. Martín-Rodríguez F, Sanz-García A, Medina-Lozano E, et al. The Value of Prehospital Early Warning Scores to Predict in - Hospital Clinical Deterioration: A Multicenter, Observational Base-Ambulance Study. Prehosp Emerg Care 2021;25:597-606.
36. Martín-Rodríguez F, Martín-Conty JL, Sanz-García A, et al. Early Warning Scores in Patients with Suspected COVID-19 Infection in Emergency Departments. J Pers Med 2021;11:170.
37. López-Izquierdo R, Martín-Rodríguez F, Santos Pastor JC, et al. Can capillary lactate improve early warning scores in emergency department? An observational, prospective, multicentre study. Int J Clin Pract 2021;75:e13779.
38. Durantez-Fernández C, Martín-Conty JL, Polonio-López B, et al. Lactate improves the predictive ability of the National Early Warning Score 2 in the emergency department. Aust Crit Care 2022;35:677-83.
39. Durantez-Fernández C, Martín-Conty JL, Medina-Lozano E, et al. Early detection of intensive care needs and mortality risk by use of five early warning scores in patients with traumatic injuries: An observational study. Intensive Crit Care Nurs 2021;67:103095.
40. Clar J, Oltra MR, Benavent R, et al. Prognostic value of diagnostic scales in community-acquired sepsis mortality at an emergency service. Prognosis in community-adquired sepsis. BMC Emerg Med 2021;21:161.
41. Mellhammar L, Linder A, Tverring J, et al. Scores for sepsis detection and risk stratification - construction of a novel score using a statistical approach and validation of RETTS. PLoS One 2020;15:e0229210.
42. Martín-Rodríguez F, López-Izquierdo R, Mohedano-Moriano A, et al. Identification of Serious Adverse Events in Patients with Traumatic Brain Injuries, from Prehospital Care to Intensive-Care Unit, Using Early Warning Scores. Int J Environ Res Public Health 2020;17:1504.
43. Martín-Rodríguez F, López-Izquierdo R, Del Pozo Vegas C, et al. Can the prehospital National Early Warning Score 2 identify patients at risk of in-hospital early mortality? A prospective, multicenter cohort study. Heart Lung 2020;49:585-91.
44. Magnusson C, Herlitz J, Axelsson C. Pre-hospital triage performance and emergency medical services nurse's field assessment in an unselected patient population attended to by the emergency medical services: a prospective observational study. Scand J Trauma Resusc Emerg Med 2020;28:81.
45. Covino M, Sandroni C, Santoro M, et al. Predicting intensive care unit admission and death for COVID-19 patients in the emergency department using early warning scores. Resuscitation 2020;156:84-91.
46. Mellhammar L, Linder A, Tverring J, et al. NEWS2 is Superior to qSOFA in Detecting Sepsis with Organ Dysfunction in the Emergency Department. J Clin Med 2019;8:1128.
47. Martín-Rodríguez F, López-Izquierdo R, Del Pozo Vegas C, et al. A Multicenter Observational Prospective Cohort Study of Association of the Prehospital National Early Warning Score 2 and Hospital Triage with Early Mortality. Emerg Med Int 2019;2019:5147808.
48. Martín-Rodríguez F, López-Izquierdo R, Del Pozo Vegas C, et al. Predictive value of the prehospital NEWS2-L —National Early Warning Score 2 Lactate— for detecting early death after an emergency. Emergencias 2019;31:173-9.
49. Martín-Rodríguez F, López-Izquierdo R, Del Pozo Vegas C, et al. Accuracy of National Early Warning Score 2 (NEWS2) in Prehospital Triage on In-Hospital Early Mortality: A Multi-Center Observational Prospective Cohort Study. Prehosp Disaster Med 2019;34:610-8.
50. Martín-Rodríguez F, Castro-Villamor MÁ, Del Pozo Vegas C, et al. Analysis of the early warning score to detect critical or high-risk patients in the prehospital setting. Intern Emerg Med 2019;14:581-9.
51. Sedgwick P. Statistical question. What is recall bias? Br Med J 2012;344. doi: 10.1136/bmj.e3519.
52. Sedgwick P. Retrospective cohort studies: advantages and disadvantages. BMJ-British Medical Journal 2014;348. doi: 10.1136/bmj.g1072.
53. Kenny JF, Chang BC, Hemmert KC. Factors Affecting Emergency Department Crowding. Emerg Med Clin North Am 2020;38:573-87.
54. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539-58.

55. Scott LJ, Redmond NM, Garrett J, et al. Distributions of the National Early Warning Score (NEWS) across a healthcare system following a large-scale roll-out. Emerg Med J 2019;36:287-92.
56. Su Y, Ju MJ, Xie RC, et al. Prognostic Accuracy of Early Warning Scores for Clinical Deterioration in Patients With COVID-19. Front Med (Lausanne) 2020;7:624255.
57. Sun Q, Qiu H, Huang M, et al. Lower mortality of COVID-19 by early recognition and intervention: experience from Jiangsu Province. Ann Intensive Care 2020;10:33.
58. Williams B. The National Early Warning Score 2 (NEWS2) in patients with hypercapnic respiratory failure. Clin Med (Lond) 2019;19:94-5.
59. NEWS2 and deterioration in COVID-19 2021. Available online: https://www.rcplondon.ac.uk/news/news2-and-deterioration-covid-19

Cite this article as: Wei S, Xiong D, Wang J, Liang X, Wang J, Chen Y. The accuracy of the National Early Warning Score 2 in predicting early death in prehospital and emergency department settings: a systematic review and meta-analysis. Ann Transl Med 2023;11(2):95. doi: 10.21037/atm-22-6587

Appendix 1 Searching strategies, inclusion and exclusion criteria, and quality assessment criteria

Searching strategies

Pubmed
#1 "Mortality"[Mesh]
#2 "Death"[Mesh]
#3 ((Mortalit*[Title/Abstract]) OR (Fatalit*[Title/Abstract])) OR (Death*[Title/Abstract])
#4 (("Mortality"[Mesh]) OR ("Death"[Mesh])) OR (((Mortalit*[Title/Abstract]) OR (Fatalit*[Title/Abstract])) OR (Death*[Title/Abstract]))
#5 (NEWS2[Title/Abstract]) OR (National Early Warning Score 2[Title/Abstract])
#6 (National Early Warning Score 2) OR (NEWS2)
#7 ((("Mortality"[Mesh]) OR ("Death"[Mesh])) OR (((Mortalit*[Title/Abstract]) OR (Fatalit*[Title/Abstract])) OR (Death*[Title/Abstract]))) AND ((National Early Warning Score 2) OR (NEWS2))

Embase
#1 'mortality'/exp
#2 'death'/exp
#3 'fatality'/exp
#4 mortalit*:ab,ti OR fatalit*:ab,ti OR death*:ab,ti
#5 #1 OR #2 OR #3 OR #4
#6 'national early warning score 2'/exp
#7 'national early warning score 2' OR news2
#8 #6 OR #7
#9 #5 AND #8

Cochrane Library
#1 MeSH descriptor: [Mortality] explode all trees
#2 MeSH descriptor: [Death] explode all trees
#3 (Mortalit*):ti,ab,kw OR (Fatalit*):ti,ab,kw OR (Death*):ti,ab,kw
#4 (National Early Warning Score 2) OR (NEWS2)
#5 #1 OR #2 OR #3
#6 #4 AND #5

Web of Science
#1 National Early Warning Score 2 (Topic) or NEWS2 (Topic)
#2 TS=(Mortalit*) OR TS=(Fatalit*) OR TS=(Death*)
#3 #2 AND #1

CNKI
#1 ((旧版主题=中英文扩展(NEWS2)+中英文扩展('National Early Warning Score 2') 或者 keyword=NEWS2+'National Early Warning Score 2' 或者 title=NEWS2+'National Early Warning Score 2' 或者 abstract=NEWS2+'National Early Warning Score 2') 并且 (旧版主题=死亡率 或者 keyword=中英文扩展(死亡率) 或者 title=中英文扩展(死亡率) 或者 abstract=中英文扩展(死亡率))) (模糊匹配)，专辑导航：全部；数据库：文献 跨库检索

WanFang Data
#1 全部:(NEWS2 or 'National Early Warning Score 2') and 全部:(死亡率)

Vip Database
#1 任意字段=NEWS2 OR 'National Early Warning Score 2' AND 任意字段=死亡率

SinoMed
#1 "NEWS2"[全部字段:智能] AND "死亡率"[全部字段:智能]

Quality Assessment Criteria

Two authors used the Prediction model Risk Of Bias Assessment Tool (PROBAST) independently to assessed the risk of bias of included trials. The PROBAST consists assessment of four key domains to judge the quality of studies: participants, predictors, outcome, analysis. The answer to each item was "+", "-" or "?" ("+" indicates low risk of bias; "-" indicates high risk of bias; and "?" indicates unclear risk of bias). If a study was judged as "low" on all domains relating to bias, then it was assigned an overall judgment of "low risk of bias" or "low concern regarding applicability", and had high quality. If a study was judged "high" in one or more domains, then it may have been judged as "at risk of bias" or "concerns regarding applicability". Disagreements were resolved by the third author.

https://dx.doi.org/10.21037/atm-22-6587

Appendix 2 Details of the quality assessment

Authors	Risk of Bias																						Applicability			
	1. Participants			2. Predictors				3. Outcome							4. Analysis										1. Participants	2. Predictors	3. Outcome	
	1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	1.2 Were all inclusions and exclusions of participants appropriate?	overall	2.1 Were predictors defined and assessed in a similar way for all participants?	2.2 Were predictor assessments made without knowledge of outcome data?	2.3 Are all predictors available at the time the model is intended to be used?	overall	3.1 Was the outcome determined appropriately?	3.2 Was a pre-specified or standard outcome definition used?	3.3 Were predictors excluded from the outcome definition?	3.4 Was the outcome defined and determined in a similar way for all participants?	3.5 Was the outcome determined without knowledge of predictor information?	3.6 Was the time interval between predictor assessment and outcome determination appropriate?	overall	4.1 Were there a reasonable number of participants with the outcome??	4.2 Were continuous and categorical predictors handled appropriately?	4.3 Were all enrolled participants included in the analysis?	4.4 Were participants with missing data handled appropriately?	★ 4.5 Was selection of preditors based on univariable analysis avoided?	4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?	4.7 Were relevant model performance measures evaluated appropriately?	★ 4.8 Were model overfitting and optimism in model performance accounted for?	★ 4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?	overalll	(1, 2, 3, 4)Risk of Bias	1. Concern that the included participants and setting do not match the review question.	2. Concern that the definition, assessment or timing of predictors in the model do not match the review question.	3. Concern that the outcome, its definition, timing or determination do not match the review question.	(1, 2, 3) Applicability
Medina-Lozano *et al.* 2020 (18)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Martin-Rodriguez *et al.* 2020 (19)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Marincowitz *et al.* 2022 (23)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Hu *et al.* 2022 (24)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Guarino *et al.* 2022 (25)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Chikhalkar *et al.* 2022 (26)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Villanueva Rabano *et al.* 2021 (27)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Thomas *et al.* 2021 (28)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Sivayoham *et al.* 2021 (29)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Richardson *et al.* 2021 (30)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Reardon *et al.* 2021 (31)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Prasad *et al.* 2021 (32)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Osawa *et al.* 2021 (33)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Masson *et al.* 2021 (34)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Martin-Rodriguez *et al.* 2021 (35)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Martin-Rodriguez *et al.* 2021 (36)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+

Appendix 2 (*continued*)

https://dx.doi.org/10.21037/atm-22-6587

Appendix 2 (*continued*)

Authors	Risk of Bias																						Applicability			
	1. Participants			2. Predictors				3. Outcome							4. Analysis										1. Participants	2. Predictors	3. Outcome	
	1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	1.2 Were all inclusions and exclusions of participants appropriate?	overall	2.1 Were predictors defined and assessed in a similar way for all participants?	2.2 Were predictor assessments made without knowledge of outcome data?	2.3 Are all predictors available at the time the model is intended to be used?	overall	3.1 Was the outcome determined appropriately?	3.2 Was a pre-specified or standard outcome definition used?	3.3 Were predictors excluded from the outcome definition?	3.4 Was the outcome defined and determined in a similar way for all participants?	3.5 Was the outcome determined without knowledge of predictor information?	3.6 Was the time interval between predictor assessment and outcome determination appropriate?	overall	4.1 Were there a reasonable number of participants with the outcome??	4.2 Were continuous and categorical predictors handled appropriately?	4.3 Were all enrolled participants included in the analysis?	4.4 Were participants with missing data handled appropriately?	★ 4.5 Was selection of preditors based on univariable analysis avoided?	4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?	4.7 Were relevant model performance measures evaluated appropriately?	★ 4.8 Were model overfitting and optimism in model performance accounted for?	★ 4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?	overalll	(1, 2, 3, 4)Risk of Bias	1. Concern that the included participants and setting do not match the review question.	2. Concern that the definition, assessment or timing of predictors in the model do not match the review question.	3. Concern that the outcome, its definition, timing or determination do not match the review question.	(1, 2, 3) Applicability
Lopez-Izquierdo *et al.* 2021 (37)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Durantez-Fernandez *et al.* 2022 (38)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Durantez-Fernandez *et al.* 2021 (39)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Clar *et al.* 2021 (40)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Mellhammar *et al.* 2020 (41)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Martin-Rodriguez *et al.* 2020 (42)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Martin-Rodriguez *et al.* 2020 (43)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Magnusson *et al.* 2020 (44)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Covino *et al.* 2020 (45)	N	Y	–	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	–	Y	Y	Y	+
Mellhammar *et al.* 2019 (46)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Martin-Rodriguez *et al.* 2019 (47)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Martin-Rodriguez *et al.* 2019 (48)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Martin-Rodriguez *et al.* 2019 (49)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+
Martin-Rodriguez *et al.* 2019 (50)	Y	Y	+	Y	Y	Y	+	Y	Y	Y	Y	Y	Y	+	Y	Y	Y	Y	Y	Y			+	+	Y	Y	Y	+

★ : This question is limited to model development studies; "Y" indicates positive of the question; "N" indicates negative of the question ; "+" indicates low risk of bias; "-" indicates high risk of bias.

https://dx.doi.org/10.21037/atm-22-6587