Review Article

iMedPub Journals
http://www.imedpub.com

Journal of Intensive and Critical Care
ISSN 2471-8505

2016
Vol. 2 No. 2: 20

DOI: 10.21767/2471-8505.100029

# Accuracy of Modified Early Warning Scores for Predicting Mortality in Hospital: A Systematic Review and Meta-analysis

**Sunee Suwanpasu[1] and Youwanuch Sattayasomboon[2]**

1 Nursing Department, King Chulalongkorn Memorial Hospital, Bangkok 10330, Thailand
2 Faculty of Public Health, Mahidol University, Bangkok 10400, Thailand

**Corresponding author:**
Sunee Suwanpasu

suneesuwanpasu@gmail.com

Nursing Department, King Chulalongkorn Memorial Hospital, Bangkok 10330, Thailand.

Tel: 66819151210

**Citation:** Suwanpasu S, Sattayasomboon Y. Accuracy of Modified Early Warning Scores for Predicting Mortality in Hospital: A Systematic Review and Meta-analysis. J Intensive & Crit Care 2016, 2:2.

## Abstract

**Background:** Patients at risk of rapid deterioration and critical illness often have preceding changes in their physiological parameters. Use of Modified Early Warning Score (MEWS) allows distinguishing points-of-care of patients who are at increased risk of serious adverse outcomes

**Objectives:** To evaluate the prognostic accuracy of the MEWS to identify patients at risk of in-hospital death.

**Methods:** The review process conformed to the recommendation form five steps to conduct a systematic review. Relevant studies from January 2000 to December 2015 were obtained from electronic databases. Standards for Reporting of Diagnostic accuracy (STARD) and Quality Assessment of Diagnostic Accuracy Studies instrument (QUADAS-2) were used to assess quality and bias of individual studies. MedCalc statistic software was used to merge the diagnostic accuracy data of recruited studies. The prognostic accuracy of predicting for in-hospital death was pooled on data from the diagnostic odds ratio and area under The ROC Curve (AUC) analysis.

**Results:** A total of 402 citations were identified yielding 16 studies for inclusion in this systematic review. Studies were statistically significant heterogeneous in terms of age and sample size. For predicting in-hospital death, high risk group of MEWS that get the threshold equal 4 or more and equal 5 or more had the Diagnostic odds ratio (DOR) of 14.278 (95% Confidence Interval [CI] 12.185 to 16.730, I2=56.59%]) and 3.28 (95%CI:2.489 to 4.323, I2=48.64%). On pooled AUC analysis, there was a trend for MEWS to estimate fair at the discriminative power of test. AUC of MEWS > 4 was 0.778 (95% CI : 0.715 to 0.841, I2=89.54%) and of MEWS > 5 was 0.646 (95% CI : 0.611 to 0.682, I2=49.69%).

**Conclusions:** The result showed a robust positive trend to predict in-hospital death. MEWS equal 4 or greater may be a favored threshold to alert to call for immediate appropriate action in hospitalized patients.

**Keyword:** Sensitivity; Specificity; Morbidity; Prognostic accuracy

**Received:** April 26, 2016; **Accepted:** April 28, 2016; **Published:** May 05, 2016

## Introduction

Patients at risk of severe illness and unanticipated serious adverse outcomes often have tended to precede changes in their physiological parameters [1-3]. Delays in the recognition and treatment of these changes increase the risk of in-hospital death [4]. Early recognition of them is important not only in order to facilitate treatment, but also so that decisions can be made as to whether admission to critical care and cardiopulmonary resuscitation are appropriate. Calculation of early warning system scores is standard practice in many hospitals to predict clinical deterioration. Screening tools, such as the Modified Early Warning Score (MEWS), have been demonstrated to have some utility in identifying these patients particularly among general medical and surgical patients [3, 5].

The basis of a MEWS as being low risk (MEWS 0-1), intermediate risk (MEWS 2-3), high risk (4 to 5), or highest risk (MEWS > 6) of serious deterioration, nursing staff was instructed to alert appropriate medical staff if the MEWS threshold was 4 to 5 indicating deterioration and the need for greater concern and a score of 6 or higher, meaning the patient was experiencing serious changes in condition that called for immediate action [6]. A recent study found a statistically significant decrease in mortality after MEWS implementation [7]. They found that death per adult admission decreased from 1.4 to 1.2% (p<0.0001) in one hospital and 1.5 to 1.3% (p< 0.0001) in the other hospital. Moon A et al. also found that patients who had undergone cardiopulmonary resuscitation had significant decrease in in-hospital mortality at the two hospitals from 52 to 42% (p<0.05) and from 70 to 40% (p<0.0001), respectively [7].

MEWS was significantly higher in non-survivor than survivor group. In emergency medical patients, median score of MEWS in non-survivors was 2-5 and survivors were 1-2 [8-11]. Mean of MEWS in non-survivors was about 3.5±1.7 and survivors were 2.3 ± 1.7 [12]. Moreover, in general internal medicine department, mean of MEWS in non-survivors was 4.5-6.3 and survivors were 3.2-4.2. [13, 14]. In a poor outcome as death or admission to the ICU, mean of MEWS was 5.6 ± 2.5 and good outcomes were 3.3 ± 2.3 [15]. In predicting patient mortality a MEWS threshold being 4 or greater had a sensitivity of 70.6% and a specificity of 37.8%, and a MEWS score of 5 or more had a sensitivity of 58.8% and a specificity of 56.2% [16].

In an effort to improve patient safety, there have been a number of studies analyses the use of MEWS to recognize the potential for clinical deterioration in ED and ward patients. However, the screening scoring has also been demonstrated to have uncertain thresholds used to identify patients in at risk to deteriorating and alert for actions. The fact is little is known exactly point of threshold of MEWS tools to predict hospitalized death should be. The aim of the systematic review and meta-analysis was to evaluate and provide robust evident on the prognostic accuracy (sensitivity, specificity, AUROC) of MEWS to in-hospital death for hospitalized patients.

# Methods

This study followed five steps to conducting a systematic review [17]. The MedCalc statistic software was used for meta-analysis.

## Framing of questions

Our key study questions were devised on the basis of the acronym PICO (patient, intervention or exposure, comparator, and outcome). They met the following criteria: (a) patients: Adult patients (aged 18 years or over) requiring hospitalization because of acute medical or surgical reasons, (b) intervention or exposure: MEWS instrumentation, (c) comparator: high risk of in-hospital death (MEWS ≥ 4 or ≥ 5) compared with low risk of in-hospital death (MEWS < 4 or < 5) (d) outcome: unplanned in-hospital death as shown in Table 1 **(Table 1)**. The main research question was thus, how many is prognostic accuracy for MEWS tool to predict the deterioration in term of unplanned in-hospital death?

## Identifying relevant publications

Predefined criteria were applied to select the final list of articles to be included in the review. The articles had to describe a study that provided the prognostic accuracy of MEWS. We included studies in which researchers reported on unplanned in-hospital death were reported, and if diagnostic accuracy of tests for MEWS scoring system were available, or were derivable from the data presented. If several tests were studied simultaneously, data from each were extracted separately, and if patients who obstetric and psychiatric conditions could be excluded. We also excluded before surgical procedure and before discharge from operating room studies, editorials, single cases and case series, studies that published only abstracts, letters, or commentaries, or if they were part of duplicate populations.

## Selection of relevant databases and search terms

An electronic search of published reports was undertaken to identify studies published between January 2000 to December 2015 were obtained from electronic databases in English language journals. An electronic database search was conducted using the following databases: PubMed (121 papers), Clinical key (33 papers) Ovid MEDLINE (53 papers), CINAHL plus with full text, EBSCO host (50 papers), Scopus, Science Direct, and Web of Science (145 papers). A variety of key words were used to conduct multiple searches. We started by searching the terms in each column of the PICO question, linked by "OR," and then combined the results of each search and retrieved only articles that contained Medical Subject Headings (MeSH) terms (207), Cumulative Index to Nursing and Allied Health Literature (CINAHL) headings (35 papers), Title-Abstr-Key, and Topic (160 papers). Additional references were obtained from the bibliographies of review articles and original papers. The search was limited to manuscripts published from January 2010 to December 2015, English language and human participants **(Figure 1)**. The selected papers were then independently assessed for inclusion by Suwanpasu and Sattayasomboon (S.S. and S.Y). Discrepancies were discussed and agreement was achieved by consensus.

## Assessing the quality of studies

The methodological quality was assessed using The Standards for Reporting of Diagnostic Accuracy (STRAD) checklist [18]. This is a reporting guideline that includes a checklist of 25 items evaluated

**Table 1** PICO: eligibility criteria for considering studies for this review.

| Population | Adult patients (aged 18 years or over) requiring hospitalization because of acute medical or surgical reasons |
|---|---|
| Intervention or exposure | The Modified Early Warning Score (MEWS) instrumentation |
| Comparator group | Criteria of high risk of in-hospital death (MEWS ≥ 4 or≥ 5) compared with low risk of in-hospital death (MEWS <4 or <5) |
| Outcome or endpoint | Unplanned in-hospital death is the main outcomes of interest. |

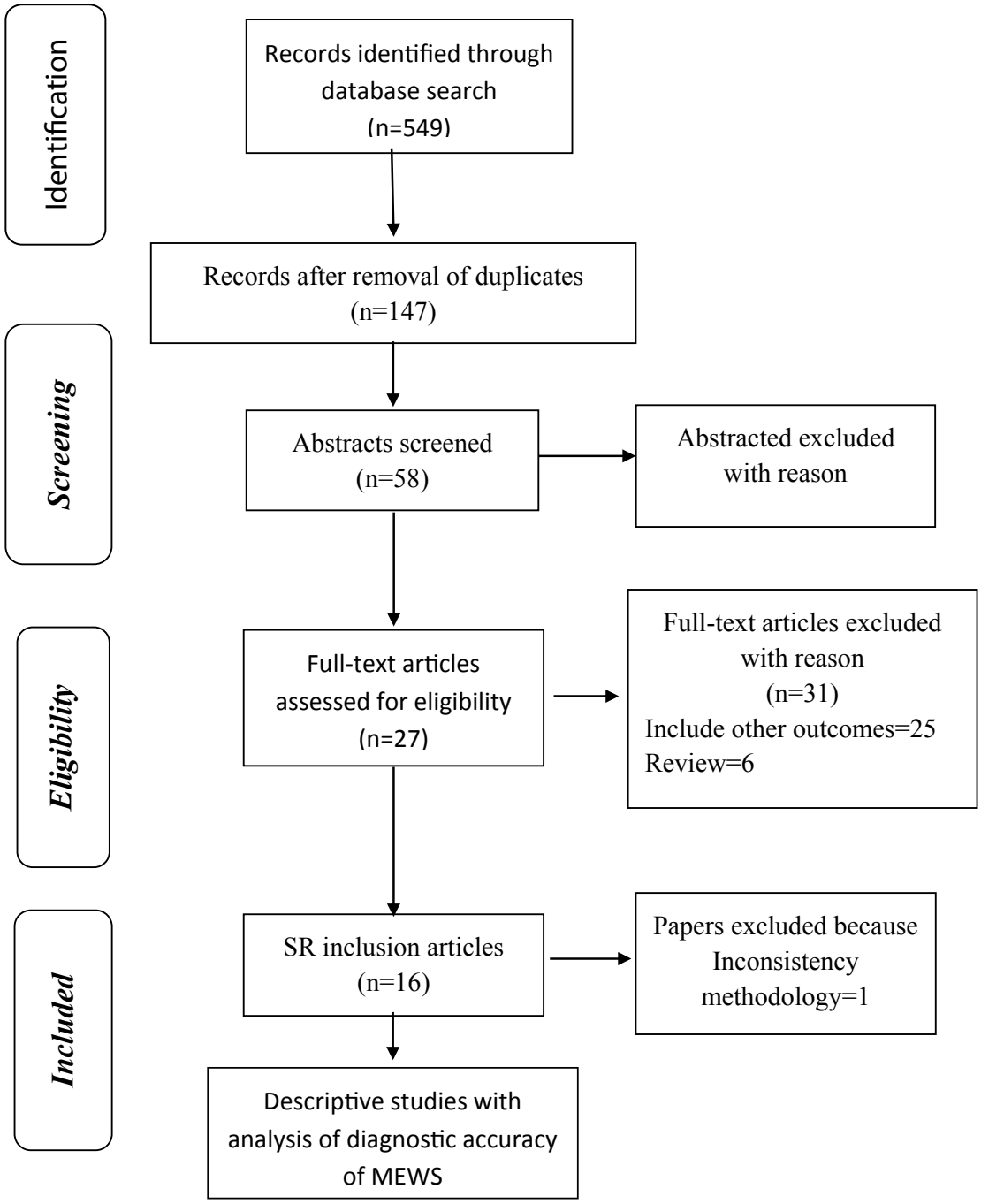



**Figure 1** PRISMA Flow Diagram.

the studies to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Each item was scored as completely reported (score=2), partly reported (score=1) or not reported (score=0). Equal weights were given to all items. A manuscript could score a maximum possible score of 50 if all items were fully reported; conversely a study not reporting any item would score 0. The final number of studies meeting our inclusion criteria was 16 reported more than 50% of the STARD items **(Table 2)**. Decisions to include or exclude a study and the information retrieved were compared between the two authors. Discrepancies were discussed and agreement was achieved by consensus.

## Data extraction strategy

Pertinent data from the selected studies were evaluated and extracted independently by two of us (SS and YS), using standardized spreadsheets for quality and extracted relevant data. Discrepancies were resolved by consensus. Information extracted included reference data (first author, journal, and institution), publication year, number of patients, mean age, proportion of

male patients, design, MEWS threshold, and prognostic accuracy (True Positive [TP], False Positive [FP], False Negative [FN], True Negative [TN], sensitivity, specificity, Positive Predictive Value [PPV], Negative Predictive Value [NPV], Positive Likelihood Ratio [LR+], Negative Likelihood Ratio [LR-], area under the receiver operating characteristic curve [AUROC], and Diagnostic Odds Ratio [DOR]. A number of patients with SSS, unplanned CA and in-hospital death were recorded. This information is shown in **Table 3**. Papers in languages other than English, case reports, commentaries, review papers, letters to editors were excluded.

## Risk of bias in individual studies

Outcome reporting bias within eligible studies was reported quantitatively using QUADAS-2 assessment [19]. Data were extracted by the lead author for patient selection, index test, reference standard, and flow and timing. The information was independently checked against full-text articles by the second author and agreement was reached for the accuracy of all studies. Study bias was assessed on a scale from 1 to 3 [1=low risk, 2=unclear risk, 3=High risk. Higher score with QUADAS-2 of

**Table 2** STARD checklist for reporting of studies of diagnostic accuracy of MEWS.

| Study | | Finlay et al. [23] | Ghanem-Zoubi et al. [13] | Cattermole et al. [15] | Ho et al. [26] | Adrijevic et al. [24] | Dundar et al. [11] | Vorwerk et al. [14] | Wheeler et al. [16] | Gardner-Thorpe et al. [27] | Subbe et al. [5] | Geier et al. [21] | Stark et al. [22] | Cooksley et al. [28] | Bulut et al. [10] | Eick et al. [12] | Armagan et al. [25] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Title/ Abstract/ Keywords** | 1. A study of diagnostic accuracy? | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| **Introduction** | 2. Study aims: estimating or comparing accuracy | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 0 | 0 | 2 | 2 | 2 |
| **Methods** | 3.Inclusion and exclusion criteria, setting and locations | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 0 | 1 | 2 | 2 | 0 | 0 |
| **Participants** | 4. Recruitment | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 |
| | 5.Sampling | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| | 6. Data collection | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **Test methods** | 7. Reference standard | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 |
| | 8. Technical specifications of material and methods | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 9. Definition of and rationale for the units, cut-offs and/or categories | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 10. Expertise | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 11. Blind to the results | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Statistical methods** | 12. Diagnostic accuracy and quantify uncertainty | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 13. Test reproducibility | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Results** | 14. Beginning and end dates of recruitment. | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **Participants** | 15. Clinical and demographic characteristics | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 16. Flow diagram | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Test results** | 17.Time interval | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 18. Distribution of severity of disease | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| | 19. A cross tabulation | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 |
| | 20. Any adverse events | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Estimates** | 21. Estimates of diagnostic accuracy and statistical uncertainty | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 22. Missing data were handled. | 2 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| | 23. Subgroups analysis | 0 | 2 | 2 | 2 | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 2 | 0 |
| | 24. Test reproduci bility | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Discussion** | 25. Discuss the clinical applicability | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **Total score** | | 32 | 30 | 33 | 33 | 31 | 36 | 36 | 32 | 33 | 32 | 31 | 29 | 28 | 30 | 29 | 25 |

the selected articles suggest more likely to represent bias than lower score **(Table 4)**.

## Summary measures and synthesis of results

We considered three issues: individual study setting number of patients, mean age, design, adverse outcomes prevalence, design, and prognostic test properties (sensitivity, specificity, PPV, NPV, LR+, LR-, DOR, and AUC). To compute meta-analysis summary estimates when more than one study assessed the same index test at the same threshold for the same or similar outcomes at the same follow-up interval, we combined eligible trials' data using MedCalc statistic software. A fixed-effects model was initially used in this systematic review because we found homogeneity across the study population. A random-effects model was applied only if statistical heterogeneity existed. We assessed statistical heterogeneity using the Cochran Q test and by calculating –squared and $I^2$values ($I^2$>75%considered high level of heterogeneity) [20]. When heterogeneity was substantial ($I^2$>75%), we investigated the sources of heterogeneity by determining the effect of important modifiers: sample details (type and quantity), study design and risk for bias, and the effect of the imputed data.

## Risk of bias across studies

Publication bias was assessed by the use of a funnel plot (Funnel plot Generator alpha), which is used primarily as a visual aid for detecting bias or systematic heterogeneity. A symmetric inverted funnel shape arises from a 'well-behaved' data set, in which publication bias is unlikely. An asymmetric funnel indicates a relationship between treatment effect and study size. This suggests the possibility of either publication bias or a systematic difference between smaller and larger studies.

# Results

## Search results

For among 549 titles and abstracts and 147 duplicates were identified and removed after exporting the results in to reference management database Endnote®. Studies were examined starting with an appraisal of titles and abstracts. A preliminary review of the titles and abstracts resulted in the identification of 170 articles were selected for full-text review (See **Figure 1**-systematic review flow diagram). In the next stage, printed copies of the remaining publications were read by either Suwanpasu or Sattayasomboon (S.S. or S.Y). A consensus was made on those that met the aforementioned criteria. Furthermore, 76 studies were excluded and 94 papers remained. An additional 36 studies were excluded after examining them carefully. Of these 58 papers, we included 27 providing on diagnostic accuracy of MEWS scoring system. The methodological quality of the 17 included articles was assessed using The Standards for Reporting of Diagnostic Accuracy (STRAD) checklist. The final number of studies meeting our inclusion criteria was 16. Decisions to include or exclude a study and the information retrieved were compared between the two authors.

**Table 3** Studies of diagnostic accuracy.

| Number | Author | Year | AUC | SE | DOR | Design | Population | Men | Age | Setting | CC | MEWS cut-off | TI | TP | FN | FP | TN | TC | Sensitivity | Specificity | PPV | NPV | LR+ | LR- |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Finlay et al. [23] | 2014 | 0.82 (0.82 to 0.83) | 0.005 | 14.4813 (12.2901 to17.0633) | Retrospective | 32472 | | | electronic medical records | general | > 4 | 2346 | 307 | 2039 | 310 | 29816 | 30126 | 0.498 | 0.936 | 13.07% | 98.97% | 7.8 | 0.54 |
| 2 | Cattermole et al. [15] | 2009 | 0.754 (0.703 to 0.799) | 0.023 | 6.3137 (3.1845 to12.5178 | Prospective | 330 | 195 | 61.3+ 20.6 | ED | General | > 4 | 105 | 31 | 74 | 14 | 211 | 225 | 0.69 | 0.74 | 29.52% | 93.77% | 2.65 | 0.4 |
| 3 | Ho et al. [26] | 2013 | 0.71 | 0.03 | 2.9109 (1.9161 to 4.4223 | Retrospective | 1024 | | 61.4+ 18.1 | ED | General | > 4 | 311 | 53 | 258 | 47 | 666 | 713 | 0.53 | 0.721 | 17.04% | 93.41% | 1.9 | 0.65 |
| 4 | Andrijevic et al. [24] | 2014 | | | 5.913 | Prospective | 101 | 76 | 63+ 11.8 | ED | CAP | > 5 | 35 | 17 | 18 | 9 | 57 | 66 | 0.654 | 0.76 | 48.57% | 86.36% | 2.72 | 0.46 |
| 5 | Dundar et al. [11] | 2015 | 0.891 (0.844 to 0.937) | 0.023 | 22.4824 (11.8409 to 42.6874) | Prospective | 671 | 375 | 75+11 | ED | Med-surg | > 4 | 110 | 42 | 68 | 15 | 546 | 561 | 0.74 | 0.89 | 39.25% | 97.34% | 6.76 | 0.29 |
| 6 | Vorwerk et al. [14] | 2008 | 0.72 (0.67 to 0.77) | 0.026 | 3.76 (2.11 to 6.71) | Retrospective | 307 | 158 | 69.7 (67.5 to 71.8) | ED | sepsis | > 5 | 148 | 52 | 96 | 20 | 139 | 159 | 0.722 | 0.592 | 35.37% | 87.42% | 1.77 | 0.47 |
| 7 | Wheeler et al. [16] | 2013 | 0.59 (0.51 to 0.68) | 0.046 | 1.8313 | Prospective | 302 | 155 | 39.5+15.9 | ED | General | > 5 | 140 | 30 | 110 | 21 | 141 | 162 | 0.588 | 0.562 | 21.43% | 87.00% | 1.343 | 0.733 |
| 8 | Geier et al. [21] | 2013 | 0.642(0.517 to 0.768) | 0.064 | | Prospective | 151 | 82 | 68.3+ 18 | ED | suspect sepsis | > 5 | | | | | | | | | | | | |
| 9 | Armagan et al. [25] | 2008 | | | 24 (9.093 to 63.3455) | Prospective | 309 | 183 | 57.1+ 15.3 | ED | Medicicine | > 4 | 106 | 40 | 66 | 5 | 198 | 203 | 0.889 | 0.75 | 37.74% | 97.54% | 3.56 | 0.15 |
| 10 | Subbe et al. [3] | 2003 | | | 4.2814 (2.8134 to 6.5154) | Prospective | 1695 | 45 | 64+ 19 | General | Medicine | >5 | 133 | 37 | 96 | 129 | 1433 | 1562 | 0.223 | 0.937 | 27.82% | 91.74% | 3.55 | 0.83 |
| 11 | Gardner-throepe et al. [27] | 2006 | | | | Prospective | 334 | 165 | 58.6+ 19.2 | Surgrery | Surgery | > 4 | 58 | 4 | 54 | 0 | 276 | 276 | 1 | 0.86 | 6.90% | ####### | 6.11 | 0 |
| 12 | Stark et al. [22] | 2015 | | | 9.9048 (2.4334 to 40.3164) | Retrospective | 62 | 40 | 62 | General | Surgical | > 4 | 46 | 32 | 14 | 3 | 13 | 16 | 0.91 | 0.48 | 71.00% | 80.00% | 1.75 | 0.19 |
| 13 | Cooksley et al. [28] | 2012 | 0.6 | 0.021 | | Retrospective | 840 | 430 | | Ward | Onco | > 4 | | | | | | | | | | | | |
| 14 | Bulut et al. [10] | 2015 | 0.630 (0.608 to 0.651) | 0.107 | 3.837 (2.358 to 6.243) | Prospective | 2000 | 1039 | 61.41+ 18.92 | ED | general | > 5 | | | | | | | | | | | | |
| 15 | Ghanem et al. [13] | 2011 | 0.6947 (0.65 to 0.73) | 0.021 | | Prospective | 1072 | | 74.7+16.1 | General | | > 4 | | | | | | | | | | | | |
| 16 | Eick et al. [12] | 2015 | 0.706 (0.667 to 0.750) | 0.022 | 1.28 (1.22 to 1.35) | Prospective | 5730 | 3125 | 61.2+ 17.7 | ED | general | | | | | | | | | | | | | |

R: Retrospective Study
P: Prospective Study
PPV: Positive Predictive Value
Npv: Negative Predictive Value
Lr+: Positive Likelihood Ratio
Lr- : Negative Likelihood Ratio
Auc: Area Under the Receiver Operating Characteristic Curve
Dor: Diagnostic Odds Ratio



Journal of Intensive and Critical Care
ISSN 2471-8505
2016
Vol. 2 No. 2: 29

Discrepancies were discussed and agreement was achieved by consensus.

## Characteristics of included studies

Papers include in the final review originated form six countries, Australia [14], Germany [12, 21], Israel [13], United States of America [22, 23], Hong Kong [15], Serbia [24], Malawi [16], Turkey [10, 11, 25], Singapore [26], and United Kingdom [5, 27, 28] and were published between 2010-2015. Twelve studies were prospective study and five studies were retrospective study. The mean score of STRAD was 31.25 ± 2.817 with a range 25-36 (out of a maximum of 50) **(Table 2)**. Only two studies fully reported at least 70% of the checklist items. The use of the MeSH heading 'sensitivity and specificity" were identified only 18.8% (3/16) of all studies. 11/16 (68%) state estimating diagnostic accuracy or comparing accuracy between tests or across participant groups in research question or study aims. Almost studies (15/16) state the reference standard and its rationale. Ten of sixteen (62.5%) fully described the inclusion and exclusion criteria, setting and locations where data were collected (Item 3). All of the articles mentioned data was collection planned before the index test and reference standard were performed or after (Item 6). Information about masking of the test readers was reported in only two of the included publications (12.5%). (Item 11) Methods for calculating test reproducibility and any adverse events were not reported of the publications (Item 13 and 20). Only 6.3% (2/16) of the manuscripts had a flow chart describing flow of patients within a study. Studies were significantly heterogeneous in terms of age and sample size.

In term of risk of bias and applicability of primary diagnostic accuracy studies, QUADAS-2 was applied. **(Table 4)** Ten of sixteen (62.5%) were low risk of bias with patient selection. Almost of studies were low risk of bias with index test (87.5%) and reference standard (68.8%). For, risk of bias of applicability concerns, 75% (12/16) were low risk of bias with patient selection and all studies were low risk of bias with Index test, reference standard and flow and time.

## Weighted pooled results

The diagnostic test performance for individual studies is outlined in table 3. In pooled data weighted by the number of patients of patients with in-hospital death and without in-hospital death in each study **(Table 5)**, MEWS using threshold ≥ 4 showed the Diagnostic odds ratio (DOR) of 14.278 (95% Confidence Interval [CI] 12.185 to 16.730, $I^2$=56.59%]) compared with MEWS using threshold ≤ 4 **(Figure 2).** In addition, MEWS using threshold ≥ 5 showed the Diagnostic odds ratio (DOR) of 3.28 (95% CI : 2.489 to 4.323, $I^2$=48.64%) for predicting in-hospital death **(Figure 3)**.

## Summary receiver operation characteristic analysis for each diagnostic test

In a pooled AUROC analysis for MEWS **(Table 5)** with using threshold ≥ 4, there was a trend of AUROC in-hospital death to estimate sufficient at the discriminative power of test for in-hospital death (AUROC 0.778 [95% CI : 0.715 to 0.841, $I^2$=89.54%], but there was statistically significant demonstrated heterogeneity **(Figure 4)**. For MEWS using threshold ≥ 5 to estimate at the discriminative power of test for in-hospital death was 0.646 (95% CI : 0.611 to 0.682, $I^2$=49.69%) **(Figure 5)**. Therefore, overall the MEWS using threshold ≥ 4 was generally more diagnostic accuracy than using threshold ≥ 5 for patients who had an in-hospital death.

# Discussion

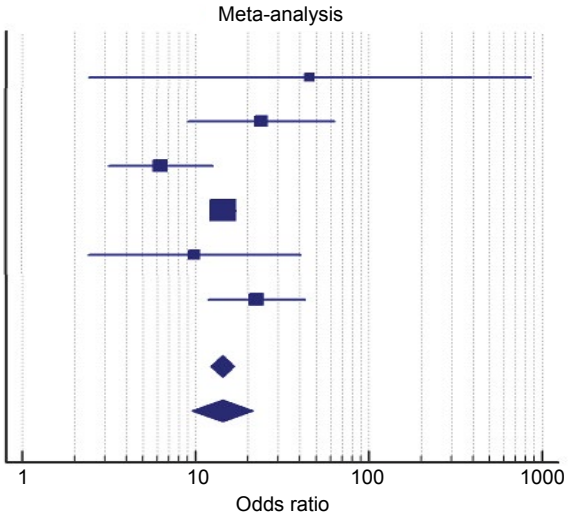
The systematic review aimed to identify the prognostic accuracy of screening instruments of the Modified Early Warning Score (MEWS) to detect risk of in-hospital death. Sixteen studies implementing using two triggering critical score approaches as 4 and greater or 5 and greater if there was in-hospital death. The first step in critical appraisal of a test is a comprehensive literature

**Table 4** QUADAS-2 results in diagnostic accuracy review of MEWS.

| Study | | Risk of Bias | | | | Applicability concerns | | |
|---|---|---|---|---|---|---|---|---|
| | | Patient selection | Index test | Reference standard | Flow and time | Patient selection | Index test | Reference standard |
| 1. Finlay et al. [23] | 2014 | 1 | 3 | 1 | 1 | 2 | 1 | 1 |
| 2. Ghanem-Zoubi et al. 13] | 2011 | 3 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3. Cattermole et al. [15] | 2009 | 1 | 3 | 1 | 1 | 1 | 1 | 1 |
| 4. Ho et al. [3] | 2013 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| 5. Andrijevic et al. [24] | 2014 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 6. Dundar et al. [11] | 2015 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7. Vorwerk et al. [14] | 2009 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 8. Wheeler et al. [16] | 2013 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 9. Gardner-Thorpe et al. [27] | 2006 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| 10. Subbe et al. [3] | 2003 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| 11. Geier et al. [21] | 2013 | 2 | 1 | 1 | 1 | 3 | 1 | 1 |
| 12. Stark et al. [22] | 2015 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| 13. Cooksley et al. [28] | 2012 | 1 | 1 | 1 | 1 | 3 | 1 | 1 |
| 14. Bulut et al. [10] | 2014 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 15. Eick et al. [12] | 2015 | 2 | 1 | 1 | 1 | 2 | 1 | 1 |
| 16. Armagan et al. [25] | 2008 | 3 | 1 | 2 | 1 | 1 | 1 | 1 |

1 low risk, 2 unclear risk, 3 High risk

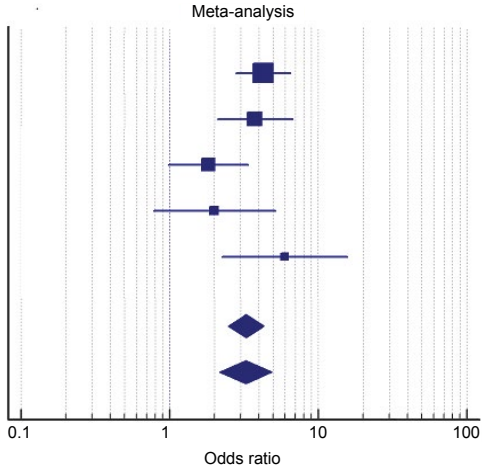
| Study | Odds ratio | 95% CI |
|---|---|---|
| Gardner-troepe et al. [27] | 45.661 | 2.423 to 860.432 |
| Armagan et al. [25] | 24.000 | 9.093 to 63.346 |
| Cattermore et al. [15] | 6.314 | 3.184 to 12.518 |
| Finlay et al. [23] | 14.481 | 12.290 to 17.063 |
| Stark et al. [22] | 9.905 | 2.433 to 40.316 |
| Dundar et al [11] | 22.482 | 11.841 to 42.687 |
| Total (fixed effects) | 14.332 | 12.234 to 16.789 |
| Total (random effects) | 14.310 | 9.558 to 21.425 |




Heterogeneity:
df=4
$I^2$=56.59%, p=0.056
Test for overall effect Z=32.879, p<0.001

**Figure 2** PRISMA Flow Diagram.

| Study | Odds ratio | 95% CI |
|---|---|---|
| Subbe et al. [5] | 4.281 | 2.813 to 6.515 |
| Vorwerk et al. [14] | 3.765 | 2.113 to 6.708 |
| Wheeler et al. [16] | 1.831 | 0.994 to 3.373 |
| Geier et al. [21] | 2.014 | 0.789 to 5.143 |
| Andrijevic et al. [24] | 5.981 | 2.276 to 15.720 |
| Total (fixed effects) | 3.280 | 2.489 to 4.323 |
| Total (random effects) | 3.262 | 2.171 to 4.901 |




Heterogeneity:
df=4
$I^2$=48.64%, p=0.0996
Test for overall effect Z=8.431, p<0.001

**Figure 3** Diagnostic odds ratio of MEWS threshold equal 5 or greater.

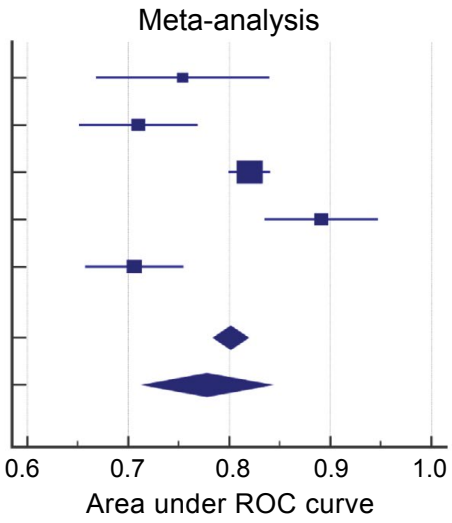
**Figure 3** Diagnostic odds ratio of MEWS threshold equal 5 or greater.

**Table 5** Prognostic accuracy of MEWS threshold.

| MEWS threshold | More than 4 or greater | | | More than 5or greater | | |
|---|---|---|---|---|---|---|
| | Fix effects (95% CI) | Random effects (95% CI) | p | Fix effects (95% CI) | Random effects (95% CI) | p |
| DOR | 14.287 (12.185-16.73) | 14.139 (9.169-21.804) | <0.001 | 3.280 (2.489-4.323) | 3.262 (2.171-4.901) | <0.001 |
| AUC | 0.801 (0.785-0.818) | 0.778 (0.715-0.841) | <0.001 | 0.801 (0.785-0.818) | 0.778 (0.715-0.841) | <0.001 |

search to identify relevant studies. Only three studies did not explicitly use the terms 'diagnostic accuracy' or 'sensitivity and specificity' in their titles or abstracts [16, 21, 27]. This is likely to limit a quick electronic search for evidence base of a diagnostic test. To evaluate the risk of bias and applicability of primary diagnostic accuracy studies, the selection of patients have introduced bias related to lack of exclusion criteria [25, 27]. Risk of bias of index test relates data did not contain the simplified alert/voice/pain/unresponsive (A/V/P/U), computation of MEWS used appropriate mapping of Glasgow Coma score [15, 23]. Moreover, Risk of bias of reference standard dues to the potential influence of prior knowledge on the interpretation of the reference standard, using

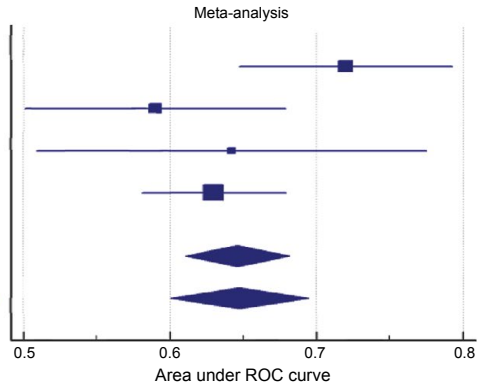
| Study | AUC | SE | 95% CI |
|---|---|---|---|
| Cattermole et al. [15] | 0.754 | 0.0437 | 0.668 to 0.840 |
| Ho et al. [26] | 0.710 | 0.0301 | 0.651 to 0.769 |
| Finlay et al. [23] | 0.820 | 0.0104 | 0.800 to 0.840 |
| Dundar et al. [11] | 0.891 | 0.0286 | 0.835 to 0.947 |
| Erick et al. [12] | 0.706 | 0.0248 | 0.657 to 0.755 |
| Total (fixed effects) | 0.801 | 0.00854 | 0.785 to 0.818 |
| Total (random effects) | 0.778 | 0.0320 | 0.715 to 0.841 |




Heterogeneity:
df=4
$I^2$=89.54%, p<0.0001
Test for overall effect Z=24.35, p <0.001

**Figure 4** Area under the receiver operating characteristic curve of MEWS threshold equal 4 or greater.

| Study | AUC | SE | 95% CI |
|---|---|---|---|
| Vorwerk et al. [14] | 0.720 | 0.0369 | 0.648 to 0.792 |
| Wheeler et al. [16] | 0.590 | 0.0452 | 0.501 to 0.679 |
| Geier et al. [21] | 0.642 | 0.0677 | 0.509 to 0.775 |
| Bulut et al. [10] | 0.630 | 0.0249 | 0.581 to 0.679 |
| Total (fixed effects) | 0.646 | 0.0181 | 0.611 to 0.682 |
| Total (random effects) | 0.647 | 0.0241 | 0.600 to 0.695 |




Heterogeneity:
df=3
$I^2$=79.69%, p=0.113
Test for overall effect Z=35.691, p<0.001

**Figure 5** Area under the receiver operating characteristic curve of MEWS threshold equal or greater.

difference threshold as reference standard [5, 22, 25-27].

However, a larger number of studies being identified for this review compares with previous systematic reviews in this subject area [1]. Studies reported that there is high discriminative performance for predicting in-hospital death. A MEWS threshold of 5 or more, a sensitivity of 58.8-68% was associated with specificity of 56.2-68% [16, 22] and accuracy 68% [22]. Lowering the MEWS triggering score for 5 to 4 (of possible 14) increased the sensitivity (70.6-91%) and accuracy (73%), but decreased specificity (37.8-48%) [16, 22]. The significant difference were found between patients who died and alive during hospital, the averages of the MEWS scores for patients who died during hospital was 3.5-4.5, and survival was 2.3-3.3 [12, 13, 15].

Our studies found that MEWS using threshold of 4 or more show the pooled diagnostic odds ratio (DOR) of 14.278 (95% Confidence Interval [CI] 12.185 to 16.730, $I^2$=56.59%]) compared with MEWS using threshold less than 4. In addition, MEWS using threshold of 5 or more showed the pooled diagnostic odds ratio (DOR) of 3.28 (95% CI : 2.489 to 4.323, $I^2$=48.64%) compared with MEWS using threshold less than 5. With the same sensitivity of the test, DOR increases with the increase of the test specificity. The value of a DOR ranges from 0 to infinity, with higher values indicating better discriminatory test performance [29].

The AUC for in-hospital death was 0.778 with MEWS using threshold of 4 or more as compared with 0.646 with the MEWS using threshold of 4 or more. The area under the curve (AUC) helps us estimate how high the discriminative power of a test is. The performance of a test for which several cutoffs are available can be expressed by means of ROC analysis [29]. The AUC takes values between 0 and 1, which higher values indicating better test performance. Thus, MEWS using threshold $\geq$ 4 was generally more diagnostic accuracy than using threshold $\geq$ 5, for patients who had an in-hospital death. There is a pair of diagnostic

sensitivity and specificity values for individual cut-off. The results support that patients were classified on the basic of MEWS as being low risk (MEWS 0-2), intermediate risk (MEWS 3-4) or high risk MEWS > 4 of serious deteriorations in association with a call out algorithm in a useful and appropriate risk-management that should be implemented for general in-patients.

## Limitations

This mate-analysis has limitations. First, this study was limited by the inclusion/exclusion criteria and the search terms employed. In addition, relevant studies may have been omitted due to applied search limitation and reviver decision-making for inclusion/exclusion for final review incorporation. Second, relevant studies may have been missed despite using comprehensive search strategies and the assistance of an information specialist to search multiple databases. Third, we identified a small number of studies with heterogeneous populations that measured different outcomes, which limited direct comparison of threshold of risk stratification tools. Finally, the meta-analysis of individual studies demonstrates significant statistical heterogeneity, even when assessing the same criteria of MEWS for the same outcome on similar patient populations. This heterogeneity is partly due to inconsistent definitions for outcomes and variable methods of measuring and/or obtaining the outcomes measures. In addition, some studies recruited patients solely from the ED, while others included ED patients after admission. Some articles lacked sufficient details to reconstruct 2×2 tables. QUADAS-2 assessment indicates several forms of potential bias, including spectrum bias and incorporation bias, since outcome assessors sometimes lacked blinding to the index test results. Another limitation is that a lack of sufficiently similar prognostic studies existed to perform meta-analysis for the instruments and outcomes. The relative importance of the outcome is undefined, but likely not equal.

## Conclusion

Accurate and reliable identification of early deterioration among hospitalized adult patients is essential to be used to risk stratify them do accurately distinguish high risk subsets and should be used by nurses. Independent risk stratification may provide clinicians with additional information to guide clinical decision-making, but further evaluation is required. Implementation of a risk stratification tool can improve processes and outcomes of care for patients. The tool needs to have sufficient predictive power to provide clinicians with confidence to use the results to guide decision-making. The first two evaluation criteria are satisfied to varying degrees by the tools identified in our review. However, prior to implementation, an impact analysis demonstrating evidence that risk stratification changes physician behavior and improves patient outcomes is needed

## References

1 Smith AF, Wood J (1998) Can some in-hospital cardio-respiratory arrests be prevented? A prospective survey. Resuscitation 37: 133-137.

2 Fairclough E, Cairns E, Hamilton J, Kelly C (2009) Evaluation of a modified early warning system for acute medical admissions and comparison with C-reactive protein/albumin ratio as a predictor of patient outcome. Clin Med (Lond) 9: 30-33.

3 Subbe CP, Kruger M, Rutherford P, Gemmel L (2001) Validation of a modified Early Warning Score in medical admissions. QJM 94: 521-526.

4 McQuillan P, Pilkington S, Allan A, Taylor B, Short A, et al. (1998) Confidential inquiry into quality of care before admission to intensive care. BMJ 316: 1853-1858.

5 Subbe CP, Davies RG, Williams E, Rutherford P, Gemmell L (2003) Effect of introducing the Modified Early Warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions. Anaesthesia 58: 797-802.

6 Mathukia C, Fan W, Vadyak K, Biege C, Krishnamurthy M (2015) Modified Early Warning System improves patient safety and clinical outcomes in an academic community hospital. Journal of Community Hospital Internal Medicine Perspectives 5: 10.3402.

7 Moon A, Cosgrove JF, Lea D, Fairs A, Cressey DM (2011) An eight year audit before and after the introduction of modified early warning score (MEWS) charts, of patients admitted to a tertiary referral intensive care unit after CPR. Resuscitation 82: 150-154.

8 Junhasavasdikul D, Theerawit P, Kiatboonsri S (2013) Association between admission delay and adverse outcome of emergency medical patients. Emerg Med J 30: 320-323.

9 Ong ME, Lee Ng CH, Goh K, Liu N, Koh ZX, et al. (2012) Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. Crit Care 16: R108.

10 Bulut M, Cebicci H, Sigirli D, Sak A, Durmus O, et al. (2014) The comparison of modified early warning score with rapid emergency medicine score: a prospective multicentre observational cohort study on medical and surgical patients presenting to emergency department. Emerg Med J 31: 476-481.

11 Dundar ZD, Ergin M, Karamercan MA, Ayranci K, Colak T, et al. (2015) Modified Early Warning Score and VitalPac Early Warning Score in geriatric patients admitted to emergency department. Eur J Emerg Med.

12 Erick C, Rizas KD, Meyer-Zurn CS, Groga-Bada P, Hamm W, et al. (2015) Autonomic nervous system activity as risk predictor in the medical emergency department: a prospective cohort study. Crit Care 43: 1079-1086.

13 Ghanem-Zoubi NO, Vardi M, Laor A, Weber G, Bitterman H (2011) Assessment of disease-severity scoring systems for patients with sepsis in general internal medicine departments. Crit Care 15: R95.

14 Vorwerk C, Loryman B, Coats TJ, Stephenson JA, Gray LD, et al. (2009) Prediction of mortality in adult emergency department patients with sepsis. Emerg Med J 26: 254-258.

15 Cattermole GN, Mak SK, Liow CH, Ho MF, Hung KY, et al. (2009) Derivation of a prognostic score for identifying critically ill patients in an emergency department resuscitation room. Resuscitation 80: 1000-1005.

16 Wheeler I, Price C, Sitch A, Banda P, Kellett J, et al. (2013) Early warning scores generated in developed healthcare settings are not sufficient at predicting early mortality in Blantyre, Malawi: a prospective cohort study. PLoS One 8: e59830.

17 Khan KS, Kunz R, Kleijnen J, Antes G (2003) Five steps to conducting a systematic review. Journal of the Royal Society of Medicine 96: 118-121.

18 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. (2003) The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern 138: W1-W12.

19 Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, et al. (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155: 529-536.

20 Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327: 557-560.

21 Geier F, Popp S, Greve Y, Achterberg A, Glockner E, et al. (2013) Severity illness scoring systems for early identification and prediction of in-hospital mortality in patients with suspected sepsis presenting to the emergency department. Wien Klin Wochenschr 125: 508-515.

22 Stark AP, Maciel RC, Sheppard W, Sacks G, Hines OJ (2015) An Early Warning Score Predicts Risk of Death after In-hospital Cardiopulmonary Arrest in Surgical Patients. Am Surg 81: 916-921.

23 Finlay GD, Rothman MJ, Smith RA (2014) Measuring the modified early warning score and the Rothman index: advantages of utilizing the electronic medical record in an early warning system. J Hosp Med 9: 116-119.

24 Andrijevic I, Matijasevic J, Andrijevic L, Kovacevic T, Zaric B (2014) Interleukin-6 and procalcitonin as biomarkers in mortality prediction of hospitalized patients with community acquired pneumonia. Ann Thorac Med 9: 162-167.

25 Armagan E, Yilmaz Y, Olmez OF, Simsek G, Gul CB (2008) Predictive value of the modified Early Warning Score in a Turkish emergency department. Eur J Emerg Med 15: 338-340.

26 Ho le O, Li H, Shahidah N, Koh ZX, Sultana P, et al. (2013) Poor performance of the modified early warning score for predicting mortality in critically ill patients presenting to an emergency department. World J Emerg Med 4: 273-278.

27 Gardner-Thorpe J, Love N, Wrightson J, Walsh S, Keeling N (2006) The value of Modified Early Warning Score (MEWS) in surgical in-patients: a prospective observational study. Ann R Coll Surg Engl 88: 571-575.

28 Cooksley T, Kitlowski E, Haji-Michael P (2012) Effectiveness of Modified Early Warning Score in predicting outcomes in oncology patients. QJM 105: 1083-1088.

29 Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM (2003) The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 56: 1129-1135.