This is an Open Access Article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0)

Review Article

PROGNOSTIC PERFORMANCE OF THE GLASGOW COMA SCALE VERSUS THE FOUR SCORE FOR PREDICTING MORTALITY IN ADULT INTENSIVE CARE PATIENTS: A SYSTEMATIC REVIEW AND META-ANALYSIS

Husniyatur Rizqiyah Dhomiri [1*], Taufiqur Rahman [1], Mohamad Nur [1], Endang Fauziyah [1]

[1] Jurusan Kesehatan, Politeknik Negeri Madura

Correspondence: Husniyatur Rizqiyah Dhomiri
Jurusan Kesehatan, Politeknik Negeri Madura
Email: husniyaturrizqi@gmail.com

© The Author(s). 2026.

ABSTRACT

Background: Level of consciousness is a core element of risk assessment in intensive care. The verbal component of the Glasgow Coma Scale (GCS) cannot be assessed in intubated patients, whereas the Full Outline of UnResponsiveness (FOUR) score replaces it with brainstem and respiratory items. Whether this design translates into better mortality prediction in adult intensive care unit (ICU) patients remains uncertain.

Purpose: This study aimed to compare the discrimination of the FOUR score and the GCS for mortality in adult ICU patients, and to quantify the difference in the area under the receiver operating characteristic curve (AUC) between the two scores when both were measured in the same patients.

Methods: This review was registered in PROSPERO (CRD420251229732) and reported according to PRISMA 2020. PubMed, Scopus and EBSCOhost (MEDLINE and CINAHL) were searched to 4 October 2026, with no supplementary searching. Observational studies of adult ICU patients in which both scores were recorded and mortality was reported were eligible; studies published more than ten years before the search and studies whose only outcome was functional status were excluded. Risk of bias was assessed with the QUIPS tool. Studies whose AUCs were reported inconsistently across the abstract, text and tables, or derived from repeated ratings of the same patients, were excluded from the primary analysis and examined separately. AUCs were pooled on the logit scale with a random-effects model (REML, Hartung–Knapp–Sidik–Jonkman confidence intervals). The within-study paired difference in AUC (FOUR minus GCS) was pooled using an assumed correlation between the two AUCs (r = 0.5, 0.7 and 0.9), because no study reported it. Heterogeneity was quantified with I-squared, tau-squared and a prediction interval.

Results: Eleven studies (2016–2022; 55 to 300 patients; 1,365 patients in total) from Iran (five), India (three), Taiwan, Poland and Nepal were included. Six studies (712 patients, 192 deaths) entered the primary analysis. The pooled AUC was 0.80 (95% CI 0.77 to 0.83) for the GCS and 0.85 (95% CI 0.81 to 0.88) for the FOUR score, with no detectable heterogeneity (I-squared 0% for both). The pooled within-study difference was +0.053 in favour of the FOUR score (95% CI +0.023 to +0.082; p = 0.006; assumed r = 0.7), with a 95% prediction interval of +0.021 to +0.084 that excludes zero. Leave-one-out analysis left the difference between +0.043 and +0.057 and significant throughout. One choice changed it: restoring the study that had been set aside for inconsistent reporting returned +0.036 (95% CI −0.008 to +0.080; p = 0.094). Five studies did not enter the primary analysis, five studies were at high risk of bias and six at moderate risk, and none was at low risk.
Conclusion: In adult ICU patients the FOU score discriminated mortality better than the GCS, by about 0.05 in AUC. The difference was consistent across studies and statistically significant in the primary analysis, but it disappeared when one study with inconsistent reporting was restored, and it comes from small studies at moderate to high risk of bias. The margin is too narrow, and the evidence too fragile, to justify replacing the GCS in routine practice.

Keywords: FOUR Score, Glasgow Coma Scale, Mortality, Prognosis, Intensive Care, Systematic Review, Meta-Analysis.

Article History:
Received: February 11, 2026 | Revised: October 27, 2026 | Accepted: November 2, 2026 | Published: November 6, 2026

INTRODUCTION

Mortality among adult ICU patients varies widely with case mix, health-system capacity and national income. In an international audit of 10,069 adult ICU patients from 84 countries, ICU and hospital mortality were 16.2% and 22.4%, respectively, with significant variation between countries and hospitals (Vincent et al., 2014); regional cohorts report different figures again (Kebede, Mosisa, & Yilma, 2024). Early and reliable prognostic assessment therefore supports triage, communication with families and planning of intensive care resources.

Level of consciousness is one of the clinical variables incorporated in ICU prediction models (Wijdicks et al., 2015). The Glasgow Coma Scale (GCS) has been used since 1974 and scores eye, verbal and motor responses (Teasdale & Jennett, 1974). Its simplicity and familiarity are strengths, but two limitations are relevant in the ICU. First, the verbal component cannot be assessed in intubated patients, so the verbal score must be recorded as untestable or estimated, and practice varies (Elkbuli et al., 2023; Mourão & Porto, 2024).

Second, the GCS contains no item for brainstem reflexes or respiration. Sedation is a separate problem from intubation, because it can depress all components of any coma scale.

The Full Outline of UnResponsiveness (FOUR) score was developed to address these limitations (Wijdicks, Bamlet, Maramattom, Manno, & McClelland, 2005). It scores eye response, motor response, brainstem reflexes and respiration (0–4 each), contains no verbal item, and can therefore be applied in intubated patients; it also distinguishes states such as locked-in syndrome and stages of herniation. A broader construct does not by itself guarantee better mortality prediction, and the respiratory item may be influenced by ventilatory support and non-neurological illness. Whether the FOUR score predicts mortality better than the GCS is therefore an empirical question.

Several syntheses have addressed this question. Foo, Loan, & Brennan, (2019) reviewed the relationship between the FOUR score and outcome in adults with impaired consciousness but did not compare it with the GCS. Anestis, Tsitsopoulos, Tsonidis, & Foroglou, (2020) compared the two scales across adult and pediatric populations and reported a difference in favour of the FOUR score for in-hospital mortality in adults. Schey, Schoch, & Kerr (2025) reviewed 20 studies in adult ICU patients and reported higher AUCs for the FOUR score, in studies of poor to moderate quality, whereas a review restricted to traumatic brain injury found no significant difference in any of the included studies (Nunes Dos Santos & Stayt, 2025). The largest single study, a prospective study of 1,695 consecutive admissions in 13 ICUs, reported an AUC for ICU mortality of 0.715 for the GCS and 0.742 for the FOUR score ($p = 0.001$), with no significant difference for hospital mortality (Wijdicks et al., 2015).

What these syntheses have in common is that the two scores are compared across studies rather than within them. Because both scores are recorded in the same patients, the two AUCs in one study are correlated, and comparing separately pooled AUCs ignores that correlation and can overstate the precision of the comparison. The present review therefore pools the within-study difference in AUC and examines how sensitive that difference is to the assumed correlation, to risk of bias and to individual studies. The present review differs from the most directly comparable of these syntheses in scope and in method. Schey et al. (2025) excluded studies that investigated patients with traumatic brain injury exclusively. Seven of the 11 studies included here, and three of the six studies pooled, do not appear in that review, six of them because they are confined to traumatic brain injury; conversely, 12 of their 20 studies fall outside the ten-year window applied here, and four more were not identified by our search or did not meet our outcome criterion; two of these, Kwamboka & Kerubo (2022) and Zhao et al.

(2021), fall inside our ten-year window and were not retrieved by our strategies, which is reported in the Limitations. Schey et al. also considered pooling the reported AUCs but judged the heterogeneity too great to do so (I-squared 71% for the FOUR score and 70% for the GCS), whereas the narrower and more recent set of studies assembled here was homogeneous enough to pool, which made it possible to estimate the paired difference rather than infer it from separately tabulated values. The objective was to compare the discrimination of the two scores for mortality in adult ICU patients and to quantify the paired difference in AUC.

METHODS

Review Design

This systematic review and meta-analysis followed the PRISMA 2020 reporting guideline (Page et al., 2021). The review was registered in PROSPERO on 29 December 2025 (CRD420251229732). Registration followed the start of screening and was therefore retrospective, which is acknowledged in the Limitations.

Protocol deviations. Three deviations from the registered protocol are reported here. First, after the search was completed, eligibility was restricted to studies published within the previous ten years, to reflect contemporary intensive care practice; this decision was taken after the records had been seen and therefore carries a risk of selection bias, and the seven studies excluded on this ground are listed in Supplementary Table S2. Second, studies whose only reported outcome was functional status rather than mortality were excluded at the full-text stage. Third, studies whose reported AUCs were internally inconsistent or derived from repeated ratings of the same patients were set aside from the primary analysis under a data-quality rule introduced during extraction (Section 2.7). The PROSPERO record has been updated to reflect these changes.

Eligibility Criteria

Eligibility was defined using a PICO framework. Population: adults admitted to an ICU of any type (neurological, trauma, medical, surgical or mixed). Studies enrolling only children were excluded. Studies with a predominantly adult sample that also admitted some adolescents were retained, and the age criterion of each is given in Table 1; the lowest was 12 years (Javvaji et al., 2022) and three studies enrolled from 16 years. Index test: the FOUR score. Comparator: the GCS, recorded in the same patients. Outcome: all-cause mortality during ICU or hospital admission. Studies reporting only functional outcome (for example the modified Rankin Scale

or the Glasgow Outcome Scale) were excluded. Eligible measures of prognostic performance were the AUC, sensitivity and specificity, calibration, and odds ratios. Eligible designs were observational studies. Case reports, reviews, editorials and conference abstracts were excluded. Studies published more than ten years before the search date were excluded (see Section 2.1). Only reports published in English were eligible.

Search Strategy and Information Sources

PubMed, Scopus and EBSCOhost (MEDLINE and CINAHL) were searched; the last search was run on 4 October 2026. The search combined terms for the index test (FOUR score, Full Outline of UnResponsiveness), the comparator (Glasgow Coma Scale, GCS), the setting (intensive care unit, critical care, critically ill) and the outcome (mortality, survival, prognosis, predictive value). The three strategies as run are reproduced in Supplementary Table S1. They were not equivalent across databases: the PubMed strategy combined the index test, the comparator and the setting using title or abstract fields together with the MeSH headings "Glasgow Coma Scale" and "Intensive Care Units"; the Scopus strategy added blocks for outcome and for accuracy or prognosis but used no field tags or controlled vocabulary; and the EBSCOhost strategy used the same five blocks searched across all text (TX), which is the widest of the three and accounts for the large number of records retrieved from that platform. No controlled vocabulary was used in Scopus or EBSCOhost. The reference lists of the included studies and of the previous reviews were not screened, no citation tracking was performed, and no grey literature was sought. The number of records exported from EBSCOhost was checked against the number of hits returned and the export was complete.

Study Selection Process

Records were de-duplicated by digital object identifier and by normalized title, year and first author; pairs that were similar but not identical were inspected individually. Titles and abstracts were then screened against the eligibility criteria, followed by full-text assessment. Two reviewers screened titles, abstracts and full texts independently, and disagreements were resolved by discussion. Reasons for exclusion at the full-text stage were recorded (Supplementary Table S2). One report could not be retrieved; it was sought through the publisher and through general web searching, without success, and no further attempt was made. The selection process is shown in Figure 1.

Data Extraction

The following items were extracted into a standardized form: first author, year, country and ICU type, study design, sample size, population, timing of score recording, handling of sedation and intubation, type of mortality outcome and its time horizon, number of deaths and survivors, AUC with 95% confidence interval for each score, standard error, optimal cut-off, sensitivity, specificity, calibration, any formal comparison of the two AUCs, and odds ratios. Items that were not reported are marked "NR". Where a study reported a standard error but no confidence interval, the interval was calculated as the AUC plus or minus 1.96 standard errors. Where a study reported neither, the standard error was calculated from the AUC and the numbers of deaths and survivors using the method of Hanley and McNeil (1982); these intervals are identified as calculated in Table 3. Two reviewers extracted data independently and resolved differences by discussion. The authors of the primary studies were not contacted for unreported data, and missing data were not imputed.

Quality Appraisal

Risk of bias was assessed with the Quality in Prognosis Studies (QUIPS) tool (Hayden, van der Windt, Cartwright, Cote, & Bombardier, 2013), which covers study participation, study attrition, prognostic-factor measurement, outcome measurement, study confounding, and statistical analysis and reporting. Because the two scores were compared within the same patients and were therefore subject to the same case mix, the confounding domain was rated low for studies that made an unadjusted within-patient comparison; the one study that used multivariable models with different covariates for the two scores is annotated in the rationale table, and its AUCs were not pooled because they describe the models rather than the scores. Domain judgments (low, moderate or high risk) are given in Table 2. An overall judgment of high risk was assigned if any domain was high, and of moderate risk if any domain was moderate. Risk of bias was used in interpretation and in sensitivity analysis, and not to exclude studies. Two reviewers assessed risk of bias independently and resolved differences by discussion.

Data Synthesis

AUCs were transformed to the logit scale before pooling, so that the pooled estimate and its interval remain within the interval from 0 to 1; standard errors on the logit scale were derived from the reported or calculated 95% confidence limits. AUCs were pooled with a random-effects model with the between-study variance estimated by restricted maximum likelihood, and confidence intervals were obtained with the Hartung–Knapp–Sidik–Jonkman adjustment,

which is preferred when the number of studies is small. Results were back-transformed for presentation.

The main comparative analysis was based on the within-study difference in AUC (FOUR minus GCS). Because both scores were measured in the same patients, the variance of this difference depends on the correlation between the two AUCs. No study reported this correlation, so it was assumed, and the analysis was repeated for r = 0.5, 0.7 and 0.9; r = 0.7 was taken as the primary value and the others as sensitivity analyses. The assumption is a limitation, and Section 4 returns to it.

Heterogeneity was quantified by tau-squared, by I-squared derived from Cochran's Q, and by a 95% prediction interval. Because restricted maximum likelihood can return a between-study variance of zero when Q is only moderately elevated, both quantities are reported. Two studies were set aside from the primary analysis under a data-quality rule applied during extraction and before any pooled estimate was computed: one reported AUC values that differ between its abstract, its text and its tables, and one appears to have derived its AUCs from repeated daily ratings of the same patients rather than from one observation per patient. Both are shown in the forest plots and both are restored in sensitivity analyses, so the effect of the rule on the result is visible. The other sensitivity analyses were: varying the assumed correlation; using the alternative mortality horizon reported by the study with inconsistent values; removing the study at high overall risk of bias; recalculating the one implausible published confidence interval; leave-one-out analysis; and cross-checks with the DerSimonian–Laird and fixed-effect estimators. Tests for small-study effects were not performed because fewer than ten studies were pooled (Page et al., 2021). Studies that could not be pooled are described narratively following the SWiM guideline (Campbell et al., 2020). Analyses were performed in Python 3 with the scipy library; the analysis script is provided as a supplementary file so that the estimates can be reproduced.

RESULTS

Study Selection

The search identified 258 records (PubMed 28, EBSCOhost 146, Scopus 84). After 31 duplicates were removed, 227 records were screened by title and abstract and 207 were excluded, of which 128 made no reference to the FOUR score and 79 were excluded for other reasons, mainly a pediatric population, a review or editorial format, a reliability study without a mortality outcome, or an abstract with insufficient information. Twenty reports were sought

for retrieval; one could not be obtained (Section 2.4). Of the 19 reports assessed in full text, eight were excluded: seven because they had been published more than ten years before the search, and one because its only outcome was functional status rather than mortality. Eleven studies were included (Figure 1).

Characteristics of Included Studies

The 11 included studies were published between 2016 and 2022 and enrolled 55 to 300 patients each, 1,365 patients in total, among whom 374 died (27.4%) (Table 1). Five were conducted in Iran (Hosseini, Ayyasi, Akbari, & Heidari Gorji, 2017; Khoshfetrat, Yaghoubi, Khodadadi Hosseini, & Farahmandrad, 2020; Momenyan, Mojtaba, & Tahmineh, 2017; Ramazani & Hosseini, 2019; Sepahvand et al., 2016), three in India (Javvaji et al., 2022; Mishra, Mahajan, Kapoor, Prabhakar, & Bithal, 2019; Suresh, Yaddanapudi, & Podder, 2019), and one each in Taiwan (Nyam et al., 2017), Poland (Kasprowicz, Burzynska, Melcer, & Kubler, 2016) and Nepal (Khanal, Bhandari, Shrestha, & Acharya, 2016). All were single-centre hospital-based observational studies; most were described by their authors as prospective, and one was described as retrospective in its methods section although its abstract described enrolment (Nyam et al., 2017).

Six studies enrolled patients with traumatic brain injury or head trauma (Hosseini et al., 2017; Kasprowicz et al., 2016; Khoshfetrat et al., 2020; Momenyan et al., 2017; Nyam et al., 2017; Sepahvand et al., 2016), two enrolled neurological or neurosurgical ICU patients (Mishra et al., 2019), aneurysmal subarachnoid haemorrhage (Khanal et al., 2016), and three enrolled mixed or medical ICU populations with altered consciousness (Javvaji et al., 2022; Ramazani & Hosseini, 2019; Suresh et al., 2019). The mortality endpoint was in-hospital mortality or mortality at discharge in six studies, one of which reported early (before 14 days) and delayed mortality separately (Hosseini et al., 2017) and one of which did not state a time horizon (Ramazani & Hosseini, 2019); ICU mortality in three studies; 28-day mortality in one; and mortality up to one month in one. Scores were recorded within the first 24 hours of ICU admission in seven studies. One recorded them early after admission without specifying when (Suresh et al., 2019), one used the mean of two raters obtained within seven days (Momenyan et al., 2017), one assessed patients daily (Sepahvand et al., 2016), and in one the timing of the analysed scores was not stated (Khoshfetrat et al., 2020). Eight studies described how sedation was handled, by excluding sedated patients, by reducing or stopping sedation before scoring, or by scoring at the first spontaneous awakening trial; three did not report it. No study reported results separately for intubated and non-intubated patients.

Synthesis of Results

The 11 included studies were published between 2016 and 2022 and enrolled 55 to 300 patients each, 1,365 patients in total, among whom 374 died (27.4%) (Table 1). Five were conducted in Iran (Hosseini, Ayyasi, Akbari, & Heidari Gorji, 2017; Khoshfetrat, Yaghoubi, Khodadadi Hosseini, & Farahmandrad, 2020; Momenyan, Mojtaba, & Tahmineh, 2017; Ramazani & Hosseini, 2019; Sepahvand et al., 2016), three in India (Javvaji et al., 2022; Mishra, Mahajan, Kapoor, Prabhakar, & Bithal, 2019; Suresh, Yaddanapudi, & Podder, 2019), and one each in Taiwan (Nyam et al., 2017), Poland (Kasprowicz, Burzynska, Melcer, & Kubler, 2016) and Nepal (Khanal, Bhandari, Shrestha, & Acharya, 2016). All were single-centre hospital-based observational studies; most were described by their authors as prospective, and one was described as retrospective in its methods section although its abstract described enrolment (Nyam et al., 2017).

Six studies enrolled patients with traumatic brain injury or head trauma (Hosseini et al., 2017; Kasprowicz et al., 2016; Khoshfetrat et al., 2020; Momenyan et al., 2017; Nyam et al., 2017; Sepahvand et al., 2016), two enrolled neurological or neurosurgical ICU patients (Mishra et al., 2019), aneurysmal subarachnoid haemorrhage (Khanal et al., 2016), and three enrolled mixed or medical ICU populations with altered consciousness (Javvaji et al., 2022; Ramazani & Hosseini, 2019; Suresh et al., 2019). The mortality endpoint was in-hospital mortality or mortality at discharge in six studies, one of which reported early (before 14 days) and delayed mortality separately (Hosseini et al., 2017) and one of which did not state a time horizon (Ramazani & Hosseini, 2019); ICU mortality in three studies; 28-day mortality in one; and mortality up to one month in one. Scores were recorded within the first 24 hours of ICU admission in seven studies. One recorded them early after admission without specifying when (Suresh et al., 2019), one used the mean of two raters obtained within seven days (Momenyan et al., 2017), one assessed patients daily (Sepahvand et al., 2016), and in one the timing of the analysed scores was not stated (Khoshfetrat et al., 2020). Eight studies described how sedation was handled, by excluding sedated patients, by reducing or stopping sedation before scoring, or by scoring at the first spontaneous awakening trial; three did not report it. No study reported results separately for intubated and non-intubated patients.

Of the 11 included studies, six entered the primary analysis (712 patients, 192 deaths). Two reported no AUC, one reported AUCs for multivariable models rather than for the scores themselves, and two were set aside under the data-quality rule described in Section 2.7; all five

are described in Section 3.6 and the two-set aside are plotted alongside the pooled studies in Figures 2 and 3.

The pooled AUC for mortality was 0.80 (95% CI 0.77 to 0.83) for the GCS and 0.85 (95% CI 0.81 to 0.88) for the FOUR score (Figure 2, Table 3). Neither comparison showed detectable between-study variation: I-squared was 0% for both scores, Cochran's Q was non-significant (p = 0.84 and p = 0.57), the restricted maximum likelihood estimate of the between-study variance was zero, and the prediction intervals (0.77 to 0.83 and 0.81 to 0.89) were accordingly narrow. Both scores separated survivors from non-survivors well, and did so consistently across the six settings.

The within-study difference in AUC favoured the FOUR score in five of the six studies and was essentially zero in the sixth (Nyam et al., 2017, −0.003). The pooled difference was +0.053 (95% CI +0.023 to +0.082; p = 0.006) with the primary assumed correlation of r = 0.7, with I-squared 0% and a 95% prediction interval of +0.021 to +0.084 that excludes zero (Figure 3). The estimate was unchanged at r = 0.5 (+0.053, 95% CI +0.023 to +0.082) and almost unchanged at r = 0.9 (+0.050, 95% CI +0.015 to +0.086), although the apparent heterogeneity rose to 58% at the highest assumed correlation, because a higher assumed correlation shrinks the within-study variance against which between-study variation is judged.

The pooled difference was stable against every analytical choice except one (Table 4). Leave-one-out analysis gave values between +0.043 and +0.057, significant in all six analyses (p = 0.003 to 0.034), so no single study drives the result. Removing the study at high risk of bias gave +0.054 (95% CI +0.020 to +0.088), recalculating the one implausible published confidence interval gave +0.053, and the DerSimonian–Laird and fixed-effect estimators both gave +0.053 (95% CI +0.029 to +0.076).

The exception is the data-quality rule itself. Restoring Hosseini et al. (2017), whose AUC values differ between its abstract, its text and its tables, returned a pooled difference of +0.036 (95% CI −0.008 to +0.080; p = 0.094) and raised I-squared to 54%; using the delayed rather than the early mortality estimate from that study gave +0.040 (95% CI +0.003 to +0.076). Restoring Sepahvand et al. (2016) instead gave +0.045 (95% CI +0.023 to +0.068), and restoring both gave +0.036 (95% CI +0.001 to +0.071). Only one of these analyses removes statistical significance, and it is the one that restores Hosseini et al. alone; restoring both studies set aside leaves the difference significant at +0.036. Section 4 sets out why the rule was applied.

Five of the 11 studies did not contribute to the primary analysis. Suresh et al., (2019) and Khoshfetrat et al., (2020) reported no AUC; Suresh et al. reported that both scores were

associated with ICU survival in logistic regression, with a marginally lower Akaike information criterion for the FOUR score (105.5 versus 109.5), and Khoshfetrat et al. reported that mean scores on both scales were lower in patients who died. Kasprowicz et al., (2016) reported AUCs of 0.906 for a multivariable model containing the FOUR score and 0.913 for a model containing the GCS together with pupil reactivity (p = 0.55); these describe the models rather than the scores, and the two models did not contain the same covariates.

Two further studies were set aside under the data-quality rule. In Hosseini et al. (2017) the AUC for the FOUR score in predicting delayed mortality is 0.89 in the abstract and in Table 2 but 0.94 in the results text, and the APACHE II value for early mortality is 0.90 in the abstract and 0.92 in the text. The confidence limits throughout are printed as "CI = 0.95. 0.81 - 0.97", which conflates the confidence level with the limits and cannot be read unambiguously. The study also contributed only 11 early deaths. Sepahvand et al. (2016) reports AUCs of 0.961 and 0.928 but appears to have computed them from repeated daily ratings rather than from one observation per patient, and states that the difference in GCS between survivors and non-survivors was not significant (p = 0.542) although the reported group means differ by more than six points. Both studies are displayed in Figures 2 and 3 and restored in Table 4, so their effect on the pooled estimate can be read directly. None of the three studies without a usable AUC contradicts the pooled result, and none points to a larger difference than the one estimated.

Three of the pooled studies reported calibration, with inconsistent results: an acceptable Hosmer–Lemeshow fit for the FOUR score and a poor fit for the GCS (Ramazani & Hosseini, 2019), an acceptable fit for both scores (Khanal et al., 2016), and a poor fit for the FOUR score model (Javvaji et al., 2022). Optimal cut-offs varied across studies from 4 to 8 for the GCS and from 4 to 11 for the FOUR score, so no single threshold can be recommended. Odds ratios per one-point change in score were reported by several studies but cannot be compared between the scales, which have different ranges (GCS 3 to 15, FOUR 0 to 16). No study reported discrimination separately for intubated or sedated patients, the group in which the FOUR score is expected to do best.

DISCUSSION

In six studies that recorded both scores in the same adult ICU patients, the FOUR score discriminated mortality better than the GCS by 0.053 in AUC (95% CI +0.023 to +0.082). The estimate held up well. It was consistent across studies, it survived every leave-one-out analysis

and both alternative estimators, and its prediction interval excluded zero. One thing unsettles it: restoring the study that had been set aside for inconsistent reporting cuts the difference by a third and removes its significance. The FOUR score is therefore probably the better discriminator by a small margin, and the size of that margin is on firmer ground than its statistical significance.

The size and direction agree closely with the largest primary study, in which the FOUR AUC for ICU mortality exceeded the GCS AUC by 0.027 (Wijdicks et al., 2015), and with earlier reviews that tabulated the two AUCs separately (Anestis et al., 2020; Schey et al., 2025). Schey et al. (2025) considered pooling and judged the heterogeneity too great (I-squared 71% and 70%); in the narrower and more recent set assembled here, heterogeneity was undetectable, which is what made a paired analysis possible. The difference in scope matters in both directions: their review excluded studies confined to traumatic brain injury, and ours excludes studies published more than ten years ago, so neither set is a subset of the other.

Two caveats apply even where the estimate is stable. The first is that discrimination is only one of the properties that decide whether a score should replace another. Calibration and incremental value over severity-of-illness scores such as APACHE II were examined only sporadically here; two studies reported acceptable calibration for the FOUR score and one reported a poor fit. The second caveat weighs more. Nobody knows what a 0.05 difference in AUC means at the bedside. No included study tested whether a difference of that size changes a triage decision or what a clinician tells a family, and on its own it would not justify retraining staff in a new scale.

Setting two studies aside is the decision on which statistical significance turns, so it needs defending. The rule was about data quality, it was applied during extraction, and it is recorded in the protocol amendment. A study whose AUC values conflict between its own abstract, text and tables cannot be assigned a reliable point estimate. A study whose AUC comes from repeated ratings of the same patients breaks the independence that the pooled variance assumes. Both studies appear in the forest plots and both are restored in Table 4, so a reader who rejects the rule can read the alternatives directly. Restoring Hosseini et al. alone gives +0.036 (95% CI −0.008 to +0.080), which is the only analysis in which the difference loses significance; restoring Sepahvand et al. alone gives +0.045 (95% CI +0.023 to +0.068), and restoring both gives +0.036 (95% CI +0.001 to +0.071). We think the six-study estimate is the better one and the eight-study estimate the more cautious one. The conclusion below holds under any of them.

The central argument for the FOUR score concerns intubated and sedated patients, in whom the GCS verbal score cannot be obtained. None of the 11 studies reported discrimination separately in that subgroup, and most did not describe how the verbal score was handled in intubated patients. The question that motivated the development of the FOUR score therefore remains unanswered by this literature. Intubation and sedation are also distinct problems: intubation prevents testing of one item, whereas sedation depresses all items of both scales, and the respiratory item of the FOUR score is itself altered by mechanical ventilation.

The limitations are of three kinds: in the evidence, in the analysis and in the review process itself. The evidence is 11 small single-centre studies, none at low risk of bias and five at high risk, concentrated in a few countries, with mortality horizons that had to be combined (ICU, in-hospital, 28-day and one-month) and with scoring done at different times. The analysis assumes a within-study correlation between the two AUCs that no study reported, and it uses confidence intervals that had to be calculated in two studies. With six studies it could not test for small-study effects. It was run in Python rather than in established meta-analysis software, although the script is supplied and two estimators agree. The review process is the weakest part. Registration in PROSPERO followed the start of screening and was therefore retrospective. The ten-year restriction was applied after the records had been seen, and it removed several larger and widely cited studies, among them Wijdicks et al. (2015). The search covered three databases with strategies that were not equivalent, used controlled vocabulary in only one of them, screened no reference lists, tracked no citations, sought no grey literature and was limited to English. It also missed two studies that fall inside its own date window and that a previous review retrieved (Kwamboka & Kerubo, 2022; Zhao et al., 2021), which says plainly enough how sensitive it was. Both report AUCs for the two scores and would have been eligible here. As tabulated by Schey et al. (2025) they favour the FOUR score more strongly than any study in our set (FOUR 0.76 versus GCS 0.63, and FOUR 0.76 versus GCS 0.64), so their omission is unlikely to have inflated the pooled difference, but it does mean the search cannot be described as comprehensive. Publication and language bias cannot be excluded. Overlap between study populations was checked and none was found.

IMPLICATION AND LIMITATIONS

Future studies should record both scores in the same patients at a prespecified time, report the paired AUC comparison with a confidence interval and the correlation between the two curves, separate ICU from hospital mortality, report calibration as well as discrimination, and

prespecify analyses in intubated and sedated patients and in non-neurological ICU populations. Individual-participant-data analysis of the existing cohorts would answer the paired-comparison question far more reliably than any further aggregate synthesis.

CONCLUSION

In 11 observational studies of adult ICU patients, of which six entered the primary analysis, the FOUR score discriminated mortality better than the GCS (pooled AUC 0.85 versus 0.80; pooled paired difference +0.053, 95% CI +0.023 to +0.082). The advantage was consistent but small. It fell to +0.036 and lost significance when a study with inconsistent reporting was restored, and it rests on a literature at moderate to high risk of bias in which no study examined the intubated and sedated patients for whom the FOUR score was designed. The evidence supports the FOUR score as a reasonable alternative to the GCS in the ICU, but not as a demonstrated improvement that warrants replacing it. Adequately powered studies that report the paired comparison properly, and that address intubated and sedated patients specifically, are needed.

ACKNOWLEDGMENTS

The authors would like to express their sincere appreciation to all researchers whose published studies contributed to this systematic review and meta-analysis. Their valuable research provided the evidence necessary for the completion of this study. The authors also acknowledge the institutional support provided by the Jurusan Kesehatan, Politeknik Negeri Madura, throughout the preparation of this manuscript. Finally, the authors thank all individuals who indirectly supported the completion of this work.

AI ASSISTANCE STATEMENT

The author did not use artificial intelligence (AI) tools in preparing this manuscript.

REFERENCES

Anestis, D. M., Tsitsopoulos, P. P., Tsonidis, C. A., & Foroglou, N. (2020). The current significance of the FOUR score: A systematic review and critical analysis of the literature. Journal of the Neurological Sciences, 409, 116600. https://doi.org/10.1016/j.jns.2019.116600

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., … Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ, 368, l6890. https://doi.org/10.1136/bmj.l6890

Elkbuli, A., Breeding, T., Ngatuvai, M., Patel, H., Andrade, R., Rosander, A., … Ang, D. (2023). Glasgow Coma Scale Intubation Thresholds and Outcomes of Patients With Traumatic Brain Injury: The Need for Tailored Practice Management Guidelines. The American Surgeon, 89(12), 6098–6113. https://doi.org/10.1177/00031348231192062

Foo, C. C., Loan, J. J. M., & Brennan, P. M. (2019). The Relationship of the FOUR Score to Patient Outcome: A Systematic Review. Journal of Neurotrauma, 36(17), 2469–2483. https://doi.org/10.1089/neu.2018.6243

Hayden, J. A., van der Windt, D. A., Cartwright, J. L., Cote, P., & Bombardier, C. (2013). Assessing bias in studies of prognostic factors. Annals of Internal Medicine, 158(4), 280–286. https://doi.org/10.7326/0003-4819-158-4-201302190-00009

Hosseini, S. H., Ayyasi, M., Akbari, H., & Heidari Gorji, M. A. (2017). Comparison of Glasgow Coma Scale, Full Outline of Unresponsiveness and Acute Physiology and Chronic Health Evaluation in prediction of mortality rate among patients with traumatic brain injury admitted to intensive care unit. Anesthesiology and Pain Medicine, 7(5), e33653. https://doi.org/10.5812/aapm.33653

Javvaji, P. K., Nagatham, P., Venkata, R. R. I., Puttam, H., John, S. K., Karavalla, H., & Pulivarthi, T. (2022). A comparison of Full Outline of UnResponsiveness score with Glasgow Coma Scale score in predicting outcomes among patients with altered mental status admitted to the critical care unit. Indian Journal of Critical Care Medicine, 26(2), 210–215. https://doi.org/10.5005/jp-journals-10071-23921

Kasprowicz, M., Burzynska, M., Melcer, T., & Kubler, A. (2016). A comparison of the Full Outline of UnResponsiveness (FOUR) score and Glasgow Coma Score (GCS) in predictive modelling in traumatic brain injury. British Journal of Neurosurgery, 30(2), 211–220. https://doi.org/10.3109/02688697.2016.1161173

Kebede, F., Mosisa, G., & Yilma, M. (2024). Incidence and predictors of mortality among patients admitted to adult intensive care unit at public hospitals in Western Ethiopia: a retrospective cohort study. Frontiers in Medicine, 11. https://doi.org/10.3389/fmed.2024.1370729

Khanal, K., Bhandari, S. S., Shrestha, N., & Acharya, S. P. (2016). Comparison of outcome predictions by the Glasgow coma scale and the Full Outline of UnResponsiveness score

in the neurological and neurosurgical patients in the Intensive Care Unit, (41), 473–476. https://doi.org/10.4103/0972-5229.188199

Khoshfetrat, M., Yaghoubi, M. A., Khodadadi Hosseini, B. M., & Farahmandrad, R. (2020). The ability of GCS, FOUR, and APACHE II in predicting the outcome of patients with traumatic brain injury: A comparative study. Biomedical Research and Therapy, 7(2), 3614–3621. https://doi.org/10.15419/bmrat.v7i2.588

Kwamboka, D., & Kerubo, L. (2022). A comparison of the Glasgow Coma Scale with Full Outline of Unresponsiveness scale in prediction of patient outcomes in the critical care unit at Kenyatta National Hospital. International Journal of Caring Sciences, 15(2), 1023–1033.

Mishra, R. K., Mahajan, C., Kapoor, I., Prabhakar, H., & Bithal, P. K. (2019). Comparison of Full Outline of UnResponsiveness (FOUR) score and the conventional scores in predicting outcome in aneurysmal subarachnoid haemorrhage patients. Indian Journal of Anaesthesia, 63(4), 295–299. https://doi.org/10.4103/ija.IJA_786_18

Momenyan, S., Mojtaba, S., & Tahmineh, M. (2017). Predictive Validity and Inter-Rater Reliability of the Persian Version of Full Outline of Unresponsiveness Among Unconscious Patients with Traumatic Brain Injury in an Intensive Care Unit, 229–236. https://doi.org/10.1007/s12028-016-0324-0

Mourão, F. A., & Porto, F. (2024). Glasgow Coma Scale: Assessing Consciousness for 50 Years. Contribuciones a Las Ciencias Sociales, 17(2), e5040. https://doi.org/10.55905/revconv.17n.2-066

Nunes Dos Santos, M. S., & Stayt, L. C. (2025). The effectiveness of FOUR score versus GCS scale in predicting mortality and morbidity in traumatic brain injured patients in intensive care: A systematic review. Intensive and Critical Care Nursing, 89, 104048. https://doi.org/10.1016/j.iccn.2025.104048

Nyam, T.-T. E., Ao, K.-H., Hung, S.-Y., Shen, M.-L., Yu, T.-C., & Kuo, J.-R. (2017). FOUR score predicts early outcome in patients after traumatic brain injury. Neurocritical Care, 26(2), 225–231. https://doi.org/10.1007/s12028-016-0326-y

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Ramazani, J., & Hosseini, M. (2019). Comparison of Full Outline of Unresponsiveness Score and Glasgow Coma Scale in Medical Intensive Care Unit. Annals of Cardiac Anaesthesia,

22(2), 143–148. https://doi.org/10.4103/aca.ACA

Schey, J. E., Schoch, M., & Kerr, D. (2025). The predictive validity of the Full Outline of UnResponsiveness score compared to the Glasgow Coma Scale in the intensive care unit: A systematic review. Neurocritical Care, 43(2), 645–658. https://doi.org/10.1007/s12028-024-02150-8

Sepahvand, E., Jalali, R., Mirzaei, M., Ebrahimzadeh, F., Ahmadi, M., & Amraii, E. (2016). Glasgow Coma Scale versus Full Outline of UnResponsiveness scale for prediction of outcomes in patients with traumatic brain injury in the intensive care unit. Turkish Neurosurgery, 26(5), 720–724. https://doi.org/10.5137/1019-5149.JTN.13536-14.0

Suresh, V., Yaddanapudi, L. N., & Podder, S. (2019). Full Outline of UnResponsiveness score versus Glasgow Coma Scale in critically ill patients with altered sensorium: A comparison of inter-observer variability and outcomes. Indian Journal of Anaesthesia, 63(8), 640–647. https://doi.org/10.4103/ija.IJA_377_19

Vincent, J.-L., Marshall, J. C., Namendys-Silva, S. A., Francois, B., Martin-Loeches, I., Lipman, J., … Sakr, Y. (2014). Assessment of the worldwide burden of critical illness: The Intensive Care Over Nations (ICON) audit. The Lancet Respiratory Medicine, 2(5), 380–386. https://doi.org/10.1016/S2213-2600(14)70061-X

Wijdicks, E. F. M., Bamlet, W. R., Maramattom, B. V, Manno, E. M., & McClelland, R. L. (2005). Validation of a new coma scale: The FOUR score. Annals of Neurology, 58(4), 585–593. https://doi.org/10.1002/ana.20611

Wijdicks, E. F. M., Kramer, A. A., Rohs, T., Hanna, S., Sadaka, F., O'Brien, J., … Foss, M. M. (2015). Comparison of the Full Outline of UnResponsiveness Score and the Glasgow Coma Scale in Predicting Mortality in Critically Ill Patients*. Critical Care Medicine, 43(2), 439–444. https://doi.org/10.1097/ccm.0000000000000707

Zhao, Z., Zhang, X., Song, C., Zhao, J., Gao, Q., & Jiang, W. (2021). A novel INCNS score for prediction of mortality and functional outcome of comatose patients. Frontiers in Neurology, 11, 585818. https://doi.org/10.3389/fneur.2020.585818

Cite this article as: Dhomiri, H. R., Rahman, T., Nur, M., Fauziyah, E. (2026). Prognostic Performance of the Glasgow Coma Scale Versus the FOUR Score for Predicting Mortality in Adult Intensive Care Patients: A Systematic Review and Meta-Analysis. *Nurse and Holistic Care, 6* (2), 215-240. https://doi.org/10.33086/nhc.v6i2.8605

Figure Captions

Figure 1. PRISMA 2020 flow diagram of study selection.

Figure 2. Area under the receiver operating characteristic curve for mortality, by score. Squares are the six pooled studies, sized by weight; horizontal lines are 95% confidence intervals; the diamond is the random-effects pooled estimate with its Hartung–Knapp–Sidik–Jonkman confidence interval; the dashed line below it is the 95% prediction interval. Open circles with dashed lines are the two studies set aside under the data-quality rule (Section 2.7); they are shown for reference and do not contribute to the pooled estimate.

Figure 3. Within-study difference in the area under the curve (FOUR minus GCS); values above zero favour the FOUR score. The within-study correlation between the two curves was assumed to be 0.7 because no study reported it. Open circles are the two studies set aside under the data-quality rule.

Table Captions

Table 1. Characteristics of the 11 included studies.

Table 2. Risk of bias of the included studies (QUIPS).

Table 3. Area under the curve for mortality, and pooled estimates.

Table 4. Sensitivity analyses for the pooled difference in AUC (FOUR − GCS)

Supplementary Table Captions

Supplementary Table S1. Search strategies as run.

Supplementary Table S2. Reports excluded at the full-text stage, and studies no longer included.

Figure 1. PRISMA 2020 flow diagram of study selection.

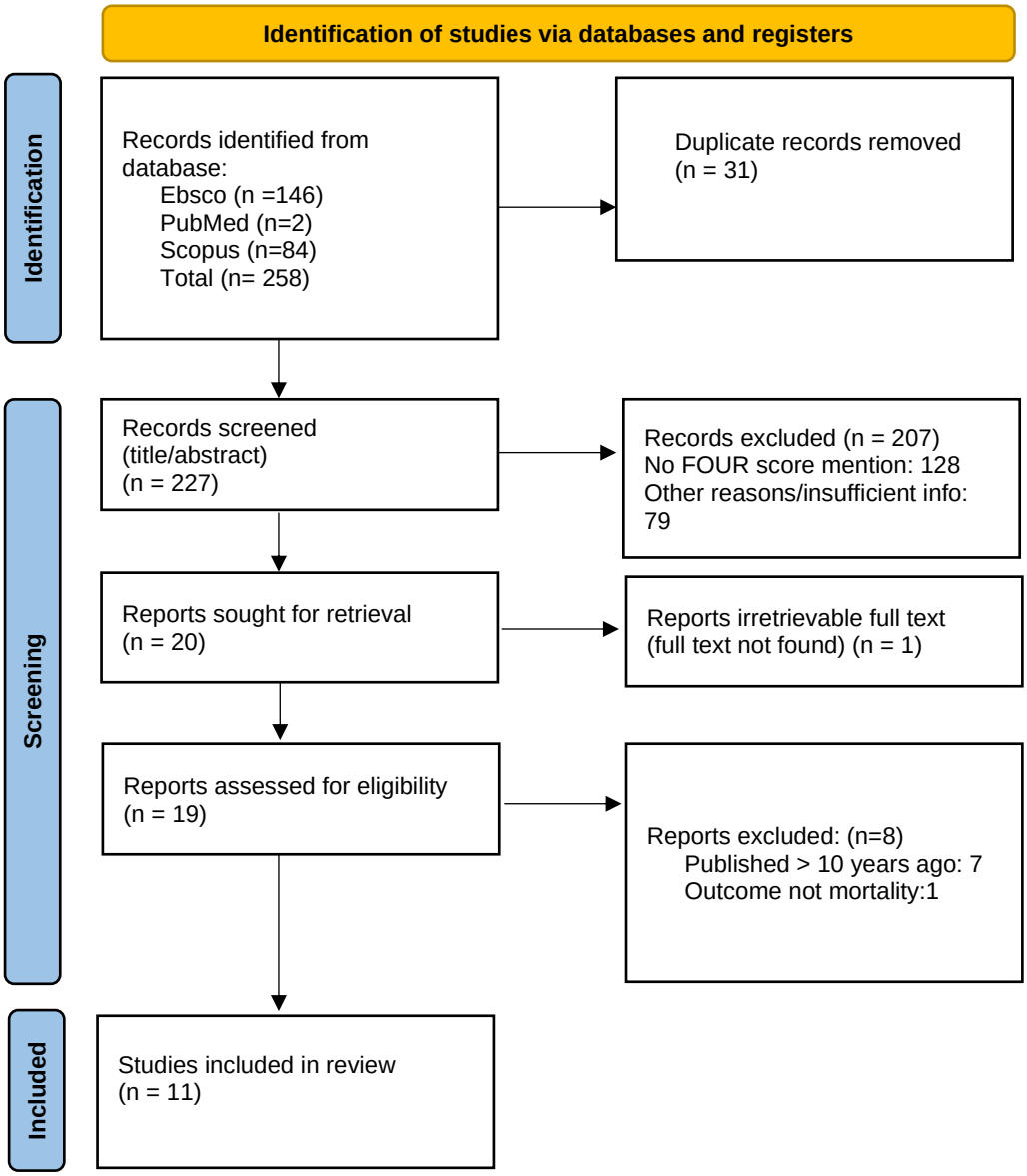


Figure 2. Area under the receiver operating characteristic curve for mortality, by score. Squares are the six pooled studies, sized by weight; horizontal lines are 95% confidence intervals; the diamond is the random-effects pooled estimate with its Hartung–Knapp–Sidik–Jonkman confidence interval; the dashed line below it is the 95% prediction interval. Open circles with dashed lines are the two studies set aside under the data-quality rule (Section 2.7); they are shown for reference and do not contribute to the pooled estimate.

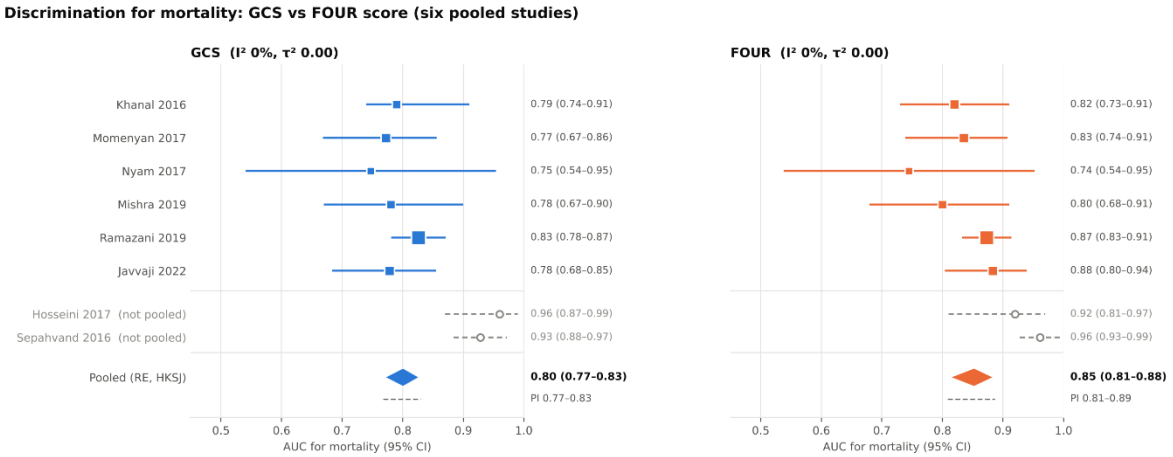


Figure 3. Within-study difference in the area under the curve (FOUR minus GCS); values above zero favour the FOUR score. The within-study correlation between the two curves was assumed to be 0.7 because no study reported it. Open circles are the two studies set aside under the data-quality rule.

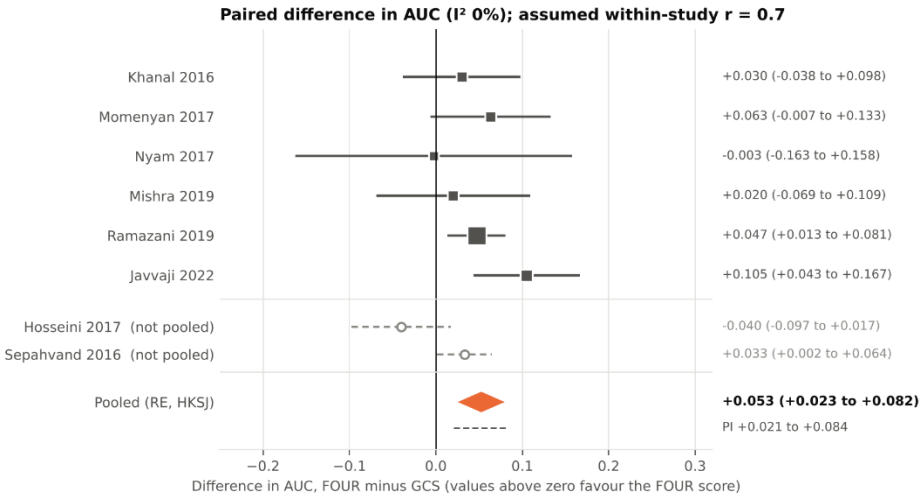


Table 1. Characteristics of the 11 included studies.

Study	Country / ICU type	Design	N	Population	Timing of scoring	Mortality outcome	Deaths, n (%)	In primary analysis?
Kasprowicz et al. (2016)	Poland / mixed ICU	Prospective cohort	162	Traumatic brain injury; 95.7% sedated	Within 24 h of admission	In-ICU mortality	50 (30.9)	No — model AUC only
Khanal et al. (2016)	Nepal / neuro–neurosurgical ICU	Prospective observational	97	Neurological or neurosurgical disease, age >16 y	Within 24 h of admission	In-hospital mortality	29 (29.9)	Yes
Sepahvand et al. (2016)	Iran / ICU	Prospective	198	Traumatic brain injury surviving >24 h; age 16–80 y	Daily from the second 24 h	In-hospital mortality	69 (34.8)	No — repeated ratings
Hosseini et al. (2017)	Iran / ICU	Diagnostic study	80	Traumatic brain injury, age 16–60 y	First 24 h	Early (<14 d) and delayed (≥15 d) mortality	15 (18.8); 11 early	No — inconsistent values
Momenyan et al. (2017)	Iran / ICU	Prospective validation	85	Unconscious traumatic brain injury; 71.8% intubated	Within 7 d; mean of two raters	In-hospital mortality (mRS 6)	16 (18.8)	Yes
Nyam et al. (2017)	Taiwan / neurosurgical ICU	Cohort	55	Isolated traumatic brain injury; sedated patients excluded	At ICU admission	ICU mortality	8 (14.5)	Yes
Mishra et al. (2019)	India / neurosurgical ICU	Prospective cohort	75	Aneurysmal subarachnoid haemorrhage	Within 24 h of admission	28-day mortality (23 of 24 in hospital)	24 (32.0)	Yes
Ramazani & Hosseini (2019)	Iran / medical ICU	Prospective observational	300	Consecutive medical ICU admissions	First 24 h	In-hospital mortality (horizon not stated)	84 (28.0)	Yes
Suresh et al. (2019)	India / multidisciplinary ICU	Prospective observational	111	Altered sensorium; 99.1% ventilated	Early after admission	ICU mortality	38 (34.2)	No — no AUC
Khoshfetrat et al. (2020)	Iran / ICU	Cross-sectional	102	Ventilated head trauma, age 18–65 y	Not stated for analysed scores	Mortality at discharge	10 (9.8)	No — no AUC
Javvaji et al. (2022)	India / critical care unit	Prospective observational	100	Non-traumatic altered sensorium, age ≥12 y	At admission	Mortality up to 1 month	31 (31.0)	Yes
Total	—	—	1,365	—	—	—	374 (27.4)	6 pooled

ICU, intensive care unit; mRS, modified Rankin Scale. Nyam et al. describe their cohort as enrolled in the abstract and as studied retrospectively in the methods; the design is reported here as a cohort study. Momenyan et al. state N = 84 in the text while their percentages imply N = 85; N = 85 is used. The data-quality rule that set two studies aside is described in Section 2.7.

Table 2. Risk of bias of the included studies (QUIPS).

Study	Study participation	Study attrition	Prognostic factor measurement	Outcome measurement	Study confounding	Statistical analysis and reporting	Overall
Kasprowicz et al. (2016)	Moderate	Low	Moderate	Low	Low	Moderate	Moderate
Khanal et al. (2016)	Moderate	Low	Moderate	Moderate	Low	Moderate	Moderate
Sepahvand et al. (2016)	Moderate	Low	Moderate	Low	Low	High	High
Hosseini et al. (2017)	Moderate	Low	Moderate	Moderate	Low	High	High
Momenyan et al. (2017)	Moderate	Moderate	Moderate	Low	Low	Moderate	Moderate
Nyam et al. (2017)	Moderate	Low	Moderate	Low	Low	High	High
Mishra et al. (2019)	Low	Low	Low	Low	Low	Moderate	Moderate
Ramazani & Hosseini (2019)	Low	Low	Moderate	Moderate	Low	Low	Moderate
Suresh et al. (2019)	Moderate	Low	Moderate	Low	Low	High	High
Khoshfetrat et al. (2020)	Moderate	Low	Moderate	Low	Low	High	High
Javvaji et al. (2022)	Moderate	Moderate	Moderate	Moderate	Low	Moderate	Moderate

Overall judgment: high if any domain was high, moderate if any domain was moderate, otherwise low. No study was at low risk of bias overall. Assessment by two reviewers independently, with differences resolved by discussion.

Table 3. Area under the curve for mortality, and pooled estimates.

Study	Mortality endpoint	GCS AUC (95% CI)	FOUR AUC (95% CI)	Difference (FOUR − GCS)	Source of the confidence interval	Paired comparison reported by the authors
Khanal et al. (2016)	In-hospital	0.790 (0.740–0.910)	0.820 (0.730–0.910)	+0.030	Reported; the published GCS interval is not symmetrical about its point estimate	No formal test
Momenyan et al. (2017)	In-hospital	0.772 (0.668–0.856)	0.835 (0.739–0.907)	+0.063	Reported	p = 0.01 (significant)
Nyam et al. (2017)	ICU	0.747 (0.541–0.954)	0.745 (0.538–0.952)	−0.003	Calculated (Hanley & McNeil, 1982)	No formal test
Mishra et al. (2019)	28-day	0.780 (0.670–0.900)	0.800 (0.680–0.910)	+0.020	Reported	p = 0.61 (not significant)
Ramazani & Hosseini (2019)	In-hospital	0.826 (0.781–0.871)	0.873 (0.832–0.914)	+0.047	Calculated from the reported standard error	No formal test
Javvaji et al. (2022)	Up to 1 month	0.778 (0.684–0.855)	0.883 (0.804–0.939)	+0.105	Reported	No formal test; authors note overlapping intervals
Pooled (random effects, HKSJ)	—	0.801 (0.770–0.828)	0.852 (0.812–0.885)	+0.053 (+0.023 to +0.082); p = 0.006	—	—
Heterogeneity	—	I^2 0%, Q p = 0.84, PI 0.768–0.830	I^2 0%, Q p = 0.57, PI 0.809–0.887	I^2 0%, PI +0.021 to +0.084	—	—
Set aside: Hosseini et al. (2017)	Early (<14 d)	0.960 (0.870–0.990)	0.920 (0.810–0.970)	−0.040	Reported; values differ between the abstract, the text and the tables	No formal test
Set aside: Sepahvand et al. (2016)	In-hospital	0.928 (0.884–0.972)	0.961 (0.928–0.994)	+0.033	Calculated; AUCs appear to derive from repeated daily ratings	No formal test

AUC, area under the receiver operating characteristic curve; CI, confidence interval; HKSJ, Hartung–Knapp–Sidik–Jonkman; PI, 95% prediction interval. Pooling was performed on the logit scale. The pooled difference assumes a within-study correlation of 0.7 between the two curves. The two studies set aside are restored in Table 4.

Table 4. Sensitivity analyses for the pooled difference in AUC (FOUR − GCS).

Analysis	Studies (k)	Pooled difference	95% CI	p	I^2
Primary analysis (assumed r = 0.7)	6	+0.053	+0.023 to +0.082	0.006	0%
Assumed r = 0.5	6	+0.053	+0.023 to +0.082	0.006	0%
Assumed r = 0.9	6	+0.050	+0.015 to +0.086	0.015	58%
Restoring Hosseini et al. (2017), early mortality	7	+0.036	−0.008 to +0.080	0.094	54%
Restoring Hosseini et al. (2017), delayed mortality	7	+0.040	+0.003 to +0.076	0.037	39%
Restoring Sepahvand et al. (2016)	7	+0.045	+0.023 to +0.068	0.003	0%
Restoring both studies set aside	8	+0.036	+0.001 to +0.071	0.047	46%
Excluding Nyam et al. (2017), the pooled study at high risk of bias	5	+0.054	+0.020 to +0.088	0.012	0%
Khanal et al. (2016) with recalculated rather than published CI	6	+0.053	+0.024 to +0.083	0.006	0%
Leave-one-out: range across the six analyses	5	+0.043 to +0.057	—	0.003 to 0.034	0–6%
Cross-check: DerSimonian–Laird estimator	6	+0.053	+0.029 to +0.076	—	—
Cross-check: fixed-effect model	6	+0.053	+0.029 to +0.076	—	—

A positive difference favours the FOUR score. Tests for small-study effects were not performed because fewer than ten studies were pooled.

Supplementary Table S1. Search strategies as run.

Database (platform)	Date run	Records	Strategy as run
PubMed	4 October 2026	28	("FOUR Score"[Title/Abstract] OR "FOUR scale"[Title/Abstract] OR "Full Outline of UnResponsiveness"[Title/Abstract]) AND ("Glasgow Coma Scale"[MeSH Terms] OR "Glasgow Coma Scale"[Title/Abstract] OR GCS[Title/Abstract]) AND ("Intensive Care Units"[MeSH Terms] OR ICU[Title/Abstract] OR ICUs[Title/Abstract] OR "intensive care"[Title/Abstract] OR "critical care"[Title/Abstract])
Scopus	4 October 2026	84	(("FOUR Score" OR "Full Outline of Unresponsiveness" OR "FOUR scale") AND ("Glasgow Coma Scale" OR GCS OR "Glasgow score") AND ("Intensive Care Unit*" OR ICU OR "critical care" OR "critically ill") AND (mortality OR "in-hospital mortality" OR "ICU mortality" OR outcome* OR "length of stay" OR "neurological outcome") AND (accurac* OR predict* OR prognos* OR sensitivity OR specificity OR AUC OR "area under the curve"))
EBSCOhost (MEDLINE and CINAHL)	4 October 2026	146	TX ("FOUR Score" OR "Full Outline of Unresponsiveness" OR "FOUR scale") AND TX ("Glasgow Coma Scale" OR GCS OR "Glasgow score") AND TX ("Intensive Care Unit*" OR ICU OR "critical care" OR "critically ill") AND TX (mortality OR "in-hospital mortality" OR "ICU mortality" OR outcome* OR "length of stay" OR "neurological outcome") AND TX (accurac* OR predict* OR prognos* OR sensitivity OR specificity OR AUC OR "area under the curve")

TX denotes a search of all text fields in EBSCOhost. The three strategies are not equivalent: only the PubMed strategy used controlled vocabulary, and only the Scopus and EBSCOhost strategies included outcome and accuracy terms. Eligibility was restricted to reports published in English. No reference-list screening, citation tracking or grey-literature searching was performed.

Supplementary Table S2. Reports excluded at the full-text stage, and studies no longer included.

Report	Stage	Reason
Gorji et al. (2014); Jalali & Rezaei (2014); Okasha et al. (2014); McNett et al. (2014); Wijdicks et al. (2015); Kocak et al. (2012); Sadaka et al. (2012)	Full text	Published more than ten years before the search (before 2016)
Nazari Ostad et al. (2017)	Full text	Outcome was poor functional outcome (modified Rankin Scale 3–6) at two weeks, not mortality; AUC values appear only in figures
Batran et al. (2026)	Retrieval	Full text could not be obtained; sought through the publisher and by general web searching
Ghelichkhani et al. (2018); Mahmoud et al. (2022); Omar et al. (2024)	Title and abstract	Abstract did not provide enough information to establish eligibility in the re-run search
Bayraktar et al. (2024)	Title and abstract	Study of variation between evaluators; no mortality discrimination statistic for either score
Kwamboka & Kerubo (2022); Zhao et al. (2021)	Not retrieved by the search	Both are eligible by the criteria applied here and were identified in Schey et al. (2025); neither was returned by any of the three strategies. Reported in the Limitations

The four studies in the last two rows of the upper section were included in the earlier version of this review and did not pass title and abstract screening of the re-run search. Bibliographic details for Kwamboka & Kerubo (2022) were taken from the reference list of Schey et al. (2025) because the original could not be obtained.