SRIWIJAYA JOURNAL OF SURGERY
Vol. 9 No. 2 (2026): pp. 275-289 · sriwijayasurgery.com
e-ISSN 2722-3558

ORIGINAL ARTICLE

Only the Boey Score Discriminates In-Hospital Mortality After Gastric Perforation: A STARD-Compliant Comparison with the Jabalpur Score

Samuel Bertua Halomoan Manurung[1*], Efman EU Manawan[1], Erial Bahar[2]

[1] Department of Surgery, Faculty of Medicine, Universitas Sriwijaya, Palembang, Indonesia
[2] Department of Anatomy, Faculty of Medicine, Universitas Sriwijaya, Palembang, Indonesia

ARTICLE INFO

Article history:
Received: August 19, 2026
Revised: September 17, 2026
Accepted: October 12, 2026
Available Online: October 18, 2026
Published: October 20, 2026

Keywords:
Boey score; Diagnostic accuracy; Gastric perforation; Jabalpur score; Mortality prediction

Corresponding author:
Samuel Bertua Halomoan Manurung
Email:
manurung.surgeon@gmail.com

DOI:
https://doi.org/10.37275/sjs.v9i2.158

Access licence:
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0).

Copyright:
© 2026 The Authors. Authors retain copyright and grant the journal the right of first publication. Published by HM Publisher.

ABSTRACT

Background: Gastric perforation remains a highly lethal surgical emergency, and competing preoperative mortality scores have not been compared directly in an Indonesian cohort.

Objective: To compare the discrimination, calibration, and clinical utility of the Boey and Jabalpur scores for in-hospital mortality after emergency surgery for gastric perforation.

Methods: This retrospective diagnostic-accuracy study, reported according to STARD 2015, included 55 consecutive adults who underwent emergency laparotomy for non-traumatic gastric perforation at a tertiary referral centre between January 2023 and September 2025. Both scores were evaluated against in-hospital death. Analyses included areas under the receiver operating characteristic curve, operating characteristics, calibration, and decision-curve analysis; quantities not identifiable from the aggregate source output were bounded rather than estimated.

Result: Twenty-nine patients (52.7%) died in hospital. The Boey score discriminated mortality (AUC 0.747, 95% CI 0.630–0.863), whereas the Jabalpur score did not (AUC 0.623, 95% CI 0.470–0.777; p = 0.118; power 34.1%). The AUC difference was 0.123, but its variance was not identifiable. At the optimal cut-offs, Boey ≥2 yielded 58.6% sensitivity and 73.1% specificity, while Jabalpur ≥10 yielded 69.0% sensitivity and 61.5% specificity. The published Boey risk table underpredicted mortality (observed-to-expected ratio 1.70), and decision-curve benefit was limited.

Conclusion: Only the Boey score separated survivors from decedents in this cohort. Formal superiority over the Jabalpur score remains unresolved, and the published Boey risk estimates require local recalibration before individual clinical use.

1. Introduction

Gastric perforation is among the most lethal of the abdominal surgical emergencies. Free communication between the gastric lumen and the peritoneal cavity produces a chemical peritonitis within minutes, a bacterial peritonitis within hours, and a systemic inflammatory response that is fatal in a substantial minority of patients even where critical care is unrestricted. Contemporary operative series place mortality at about 11% in a 500-patient Indian cohort and 14% in an Ethiopian tertiary series, against a background figure of roughly 30% mortality and 50% morbidity for perforated peptic ulcer as a whole[1-4]. The global burden of peptic ulcer disease has fallen with the retreat of *Helicobacter pylori* and the wider use of proton-pump inhibition[5-7], but the fall has been uneven: incidence and case fatality remain concentrated in low- and middle-income settings where patients present late and where operative capacity is stretched, and smoking-attributable mortality is still rising in several regions[8-10]. Indonesia sits squarely in that group, and the cohort described here sits at the severe end even of that distribution.

The prepyloric and antral segments of the anterior gastric wall are the classical site of perforation, and large operative series bear that out: among 500 consecutive perforations at one Indian tertiary centre 328 were gastric and 160 duodenal with a mean defect of 8.6 mm, while an Ethiopian series of 106 placed 81.1% of defects in the first part of the duodenum[1,2]. A breach of the anterior prepyloric wall opens directly into the supracolic compartment of the greater sac, so spillage is not contained by the lesser sac and tracks along the right paracolic gutter; a breach of the posterior wall of the body, by contrast, may be temporarily walled off within the lesser sac. The anatomical distribution observed in this cohort is set out in Figure 5, and the cascade that converts a local breach into multi-organ failure is traced in the same figure alongside the point at which each score variable enters it.

Because the operation itself is simple and largely invariant – primary repair with an omental patch, whether open or laparoscopic[11-13] – prognosis is largely determined before the incision is made, and preoperative risk stratification carries most of the clinical weight[3,14]. The Boey score, introduced in its validated form in 1987, uses three dichotomous variables: a coexisting medical illness, preoperative shock, and a delay of more than twenty-four hours between perforation and operation[15]. Its appeal is that it can be completed at the bedside in under a minute and that each component maps onto a distinct step of the pathophysiological cascade. Its published mortality table – 0%, 10%, 45% and 100% for scores of 0 to 3 – has been reproduced with varying fidelity: a prospective Nepalese series of 94 patients recorded 0%, 4.7%, 9.1% and 60.0% across the four strata and an Indian series of 103 recorded 0%, 0%, 9.1% and 90.9%, so the ordering survives everywhere while the absolute risks do not[16-20].

The Jabalpur score was developed in central India from a retrospective series of 140 patients with peritonitis and was intended to improve on that simplicity by grading rather than dichotomising the underlying physiology[21]. It scores the perforation-to-operation interval in five strata and adds weighted terms for systolic pressure, heart rate, serum creatinine, age and comorbidity, giving a range far wider than the Boey score's four levels. Two decades of use have produced markedly discordant results. A 77-patient series reported an area under the curve of 0.884 (95% CI 0.785-0.981) for mortality, statistically indistinguishable from POSSUM; a 44-patient West African series reported 0.75 with a sensitivity of 69%; and a 235-patient prospective Indian series reported 0.665 and concluded that the index is not suited to predicting mortality in secondary peritonitis[22-26]. The intuition that more information should yield better prediction is a natural one, but that spread of results is not what a well-behaved instrument produces.

That intuition has not been tested against the Boey score in an Indonesian population, and it is not self-evidently correct. A score with more terms has more opportunities to be miscoded from a retrospective record, more parameters to estimate from a small sample, and more scope for a variable that is prognostic in one case mix to be uninformative in another[27,28]. Whether the additional resolution buys additional discrimination is an empirical question about a particular population, and only a head-to-head comparison in that population can answer it. At this centre the Boey score has been compared with the Mannheim Peritonitis Index in 31 patients with gastric perforation, with a reported sensitivity of 75.0% and specificity of 53.3% at a cut-off of 2[29]; the Jabalpur score has not been evaluated here in any indication.

This study therefore set out to: (1) compare the discrimination of the Boey and Jabalpur scores for in-hospital mortality after emergency surgery for gastric perforation; (2) determine the optimal cut-off for each score by the Youden index and report the operating characteristics at every observed cut-off; (3) assess how well the published Boey mortality table calibrates to an Indonesian tertiary referral case mix; and (4) state which of the analyses conventionally requested of a diagnostic accuracy study are identified by the available data and which are not.

2. Methods

2.1 Study design and reporting

This was a retrospective diagnostic accuracy study of two competing preoperative risk scores against a single reference standard. It is reported in accordance with the Standards for Reporting of Diagnostic Accuracy Studies (STARD 2015)[30], supplemented where relevant by current guidance on the external evaluation of clinical prediction models[27]. The completed participant flow is shown in Figure 1. The study addresses a prognostic question posed in a diagnostic accuracy framework: the index tests are computed before the operation and the reference standard is ascertained at discharge, so the two are separated in time rather than measured concurrently[27].

2.2 Ethical approval

This study received ethical exemption from the Research Ethics Committee of Dr. Mohammad Hoesin Central General Hospital Palembang, Universitas Sriwijaya (No. DP.04.03/D.XVII.9.4/ETIK/184/2026). The study was conducted in accordance with the Declaration of Helsinki and 7 WHO 2011 Standards (Social Values, Scientific Values, Equitable Assessment and Benefits, Risks, Persuasion/Exploitation, Confidentiality and Privacy, and Informed Consent), referring to the 2016 CIOMS Guidelines.

Patient identifiers were stripped at the point of extraction and no patient was approached, no additional investigation was performed and no clinical decision was influenced by the study. The committee's exemption carries with it a waiver of individual written informed consent, which is the only tenable arrangement for a record review in which more than half of the cohort died during the index admission. Approval was granted after the close of the accrual window, as is customary for a retrospective chart review.

2.3 Setting, participants and data source

The study was conducted in the Department of Surgery of Dr. Mohammad Hoesin Central General Hospital, Palembang, the tertiary surgical referral centre for the province of South Sumatera, Indonesia. All adults who underwent exploratory laparotomy for a non-traumatic gastric perforation between January 2023 and September 2025 were screened. Sampling was total rather than random: every consecutive record satisfying the criteria within the accrual window entered the analysis, which removes selection at the level of the investigator but leaves the referral filter of a tertiary centre in place.

Patients were eligible if they were older than eighteen years, carried an intraoperative diagnosis of gastric perforation, and underwent exploratory laparotomy at the study centre. Patients were excluded if the perforation was traumatic in origin, if it arose from a malignancy, or if the record was incomplete for either index test or for the reference standard. The exclusion of malignant perforation is consequential and is revisited in the Discussion, since it removes precisely the subgroup in which the Jabalpur score's age and comorbidity terms would be expected to carry most weight. A second feature of the case definition needs stating plainly: ten of the 55 perforations are recorded in the operative notes as postpyloric, which is a duodenal rather than a gastric site. They are retained because they formed part of the analysed cohort and cannot be removed reproducibly from aggregate output, but the cohort is therefore better described as perforated peptic ulcer with a gastric predominance than as pure gastric perforation, and a sensitivity analysis excluding them could not be performed.

2.4 Index tests

Boey score. Three variables were each scored 0 or 1 and summed to give a total of 0 to 3: a coexisting medical illness (cardiac, hepatic, renal or diabetic), preoperative shock defined as a systolic pressure below 100 mmHg, and an interval of more than twenty-four hours between the onset of perforation and definitive treatment[15]. For orientation, 55 patients had a recorded systolic pressure, of whom 12 were below 100 mmHg and 4 below 90 mmHg.

Jabalpur score. Six domains were scored and summed[21]: the perforation-to-operation interval in five strata (<24 h, 25-72 h, 73-96 h, 97-120 h, >120 h, scoring 0 to 4); systolic pressure (0 for 70-109 mmHg, 2 for 50-69 or 110-129 mmHg, 3 for 130-159 mmHg, 4 for <49 or ≥160 mmHg); heart rate on the same four-level scheme; serum creatinine (0 for 0.6-1.4 mg/dL, 2 for 1.5-1.9, 3 for 2.0-3.4, 4 for ≥3.5); age (0 for <45 years, 2 for 45-54, 3 for 55-64, 5 for 65-74, 6 for ≥75); and 5 points for any of a defined list of comorbidities. Observed totals ranged from 1 to 21. The two scores are not independent: they share the delay and comorbidity domains outright, and the Boey shock criterion coincides with the boundary of the Jabalpur systolic-pressure band. That overlap matters for the comparison and is quantified in section 3.3.

Both index tests were derived from data recorded in the medical record before operation, and both were computed after discharge from the extracted dataset. The reviewers who computed them were therefore not blinded to the outcome in the strict STARD sense, because the outcome was already present in the record they were reading. This is an unavoidable feature of retrospective scoring and is treated as a limitation rather than glossed over[30]. The components are, however, objective measurements recorded contemporaneously for clinical purposes, which limits the scope for differential misclassification.

2.5 Reference standard

The reference standard was in-hospital all-cause death following emergency laparotomy, ascertained from the discharge summary. It was available for every patient; no patient was lost to follow-up and no outcome was missing. Death is an unambiguous endpoint that requires no adjudication, which removes the reference-standard misclassification that troubles most diagnostic accuracy studies. It is, however, an in-hospital rather than a thirty-day endpoint, and the two are not interchangeable: an in-hospital endpoint misses deaths shortly after discharge and can be lengthened by a prolonged admission.

2.6 Statistical analysis

Continuous variables are summarised as the median with the interquartile range and compared with the Mann-Whitney U test; categorical variables are summarised as counts with percentages and compared with Fisher's exact test or Pearson's chi-square as the expected counts allowed. Proportions carry Wilson score confidence intervals, which behave correctly at the boundary where a Wald interval does not[30].

Discrimination was quantified by the area under the receiver operating characteristic curve, computed as the Mann-Whitney statistic with mid-rank handling of ties and reported with a DeLong confidence interval[31]. Because the DeLong variance is anti-conservative at this sample size, each area was also tested against the null of 0.5 by a 200,000-replicate **two-sided** permutation test, in which a replicate counts against the null when its area is at least as far from 0.5 in either direction as the observed value. Operating characteristics were computed at every observed cut-off: sensitivity, specificity, positive and negative predictive value with Wilson intervals, and positive and negative likelihood ratios with Katz log-normal intervals. The optimal cut-off was identified by the maximum Youden index and is reported as a plateau where more than one cut-off attains the maximum, because an argmax over a discrete score is biased toward the smallest tied value.

Calibration of the Boey score was assessed against its own published mortality table by the ratio of observed to expected deaths, by exact binomial tests within each score level, by the calibration slope and by calibration-in-the-large with the slope fixed at unity, and globally by the exact convolution of the four stratum-specific binomial distributions[27,28]. A Hosmer-Lemeshow statistic was computed because it is conventionally requested, and is reported alongside an explanation of why it is not a valid goodness-of-fit test for an externally fixed model evaluated on its own strata. Overall accuracy was summarised by the Brier score and by the scaled Brier score referenced to the cohort prevalence. Clinical usefulness was assessed by decision-curve analysis of the two **prescribed** decision rules – Boey ≥2 and Jabalpur ≥10 – rather than of the envelope over all cut-offs, since an envelope selected in the same data is optimistic[27]. Both scores were additionally refitted as logistic models and their apparent performance corrected by fifty repetitions of ten-fold cross-validation, with the dispersion of the cross-validated statistics across repetitions reported alongside their mean. Analyses were performed in Python 3.11 with NumPy and SciPy; the analysis script and the verification harness are supplied as supplementary material. Two-sided p values are reported to three decimal places and significance was set at 0.05. No adjustment for multiplicity was made: the two areas under the curve are the pre-specified primary comparison, and every other p value in this paper is descriptive of a single covariate and is interpreted as such rather than as a screening decision.

Three quantities conventionally requested of a head-to-head comparison are not identified by aggregate output and are bounded or refused rather than fabricated: the variance of the difference between the two areas, for which the DeLong test was evaluated at both extremes of the transportation polytope of admissible pairings and under independence and is reported as an interval of p values; the net reclassification improvement, which is not reported at all; and accuracy within any subgroup.

2.7 Sample size and precision

The primary study estimated its target from a single-proportion sensitivity formula, $n = Z^2_{alpha/2} \times Sn(1 - Sn) / (d^2 \times P)$, with Z = 1.96, an anticipated sensitivity of 92%, an absolute precision of 10% and an expected prevalence of 99%, which returns 29 and was reported as 28. That formula sizes a study to estimate one sensitivity to a stated precision; it does not size a comparison of two areas under the curve[28].

Precision is therefore reported directly. Under the null of an area of 0.5 the standard error of a non-parametric area with 29 events and 26 non-events is 0.079, so the smallest area detectable at 80% power and a two-sided alpha of 0.05 is 0.720. At the area actually observed for the Boey score the power was 92.0%, so the Boey result was well powered. At the area observed for the Jabalpur score, however, the power was only 34.1%, and the Jabalpur null must therefore be read as an inconclusive result rather than as evidence that the score does not discriminate. Under the most favourable pairing of the two scores the observed difference of 0.123 would require about 46 patients for 80% power; under less favourable pairings it would require considerably more.

3. Results

3.1 Cohort and participant flow

Figure 1 sets out the participant flow. 55 adults underwent emergency laparotomy for a non-traumatic gastric perforation at Dr. Mohammad Hoesin General Hospital, Palembang between January 2023 and September 2025 and had complete data on both index tests and on the reference standard. 29 patients (52.7%) died before discharge and 26 (47.3%) were discharged alive. No outcome was missing and no patient was lost to follow-up. Two variables rest on 54 rather than 55 patients: one serum creatinine value was absent, and one heart rate was recorded as 0 per minute and is treated as missing.

The characteristics of the cohort are presented in Table 1. The median age was 62 years among those who died and 63 years among survivors (p=0.736), and 38 of 55 patients (69.1%) were men. Two variables separated the outcome groups. Systolic pressure was lower in those who died (median 110 versus 125 mmHg, p=0.010), and the corresponding dichotomy at 100 mmHg carried an odds ratio of 6.32 (95% CI 1.23-32.34, Fisher p=0.022). The interval between the onset of symptoms and arrival at hospital was strongly associated with death across its five ordered strata (chi-square 13.795, df 4, p=0.008; Cochran-Armitage trend z=3.541, p<0.001): 12 of the 13 patients who reached hospital more than 120 hours after onset died. Sex, comorbidity, creatinine, heart rate and the site of the perforation were not associated with the outcome.

Table 1. Demographic, physiological and operative characteristics of the 55 patients, stratified by in-hospital outcome.

Characteristic	Died (n = 29)	Survived (n = 26)	All (n = 55)	p
Age, years, median (IQR)	62.0 (50.0-71.0)	63.0 (55.2-66.8)	63.0 (51.5-68.0)	0.736
Sex, n (%)				0.260
Male	18 (62.1)	20 (76.9)	38 (69.1)	
Female	11 (37.9)	6 (23.1)	17 (30.9)	
Systolic pressure, mmHg, median (IQR)	110.0 (90.0-130.0)	125.0 (112.0-138.8)	117.0 (103.0-135.0)	0.010
Heart rate, /min[b], median (IQR)	101.0 (97.0-115.5)	102.0 (93.0-105.0)		0.279
Serum creatinine, mg/dL[a], median (IQR)	2.17 (1.36-3.53)	1.52 (1.04-2.31)		0.161
Preoperative shock (SBP <100 mmHg), n (%)	10 (34.5)	2 (7.7)	12 (21.8)	0.022
Coexisting medical illness, n (%)	15 (51.7)	10 (38.5)	25 (45.5)	0.419
Delay to treatment >24 h, n (%)	27 (93.1)	22 (84.6)	49 (89.1)	0.406
Serum creatinine ≥1 mg/dL, n (%)	25 (86.2)	20 (76.9)	45 (81.8)	0.490
Heart rate >100/min, n (%)	14 (48.3)	15 (57.7)	29 (52.7)	0.591
Pre-hospital interval, n (%)				0.008
<24 h	1 (3.4)	2 (7.7)	3 (5.5)	
25-72 h	11 (37.9)	19 (73.1)	30 (54.5)	
73-96 h	3 (10.3)	4 (15.4)	7 (12.7)	
97-120 h	2 (6.9)	0 (0.0)	2 (3.6)	
>120 h	12 (41.4)	1 (3.8)	13 (23.6)	
Site of perforation, n (%)				0.325
Pylorus	1 (3.4)	0 (0.0)	1 (1.8)	
Prepyloric	21 (72.4)	22 (84.6)	43 (78.2)	
Postpyloric	7 (24.1)	3 (11.5)	10 (18.2)	
Corpus	0 (0.0)	1 (3.8)	1 (1.8)	
Boey score, median (range)	2 (1-3)	1 (0-2)	1 (0-3)	0.001
Jabalpur score, median (range)	11 (1-21)	9 (1-20)	10 (1-21)	0.118

Notes: IQR, interquartile range; SBP, systolic blood pressure. Continuous variables compared with the Mann-Whitney U test; categorical variables with Fisher's exact test, except the pre-hospital interval and the perforation site, which were compared with Pearson's chi-square. [a]One creatinine value was missing, so that row is based on 54 patients and no cohort summary is given; the categorical creatinine variable was complete for all 55. [b]One heart rate was recorded as 0 per minute and is treated as missing, so that row is also based on 54 patients.

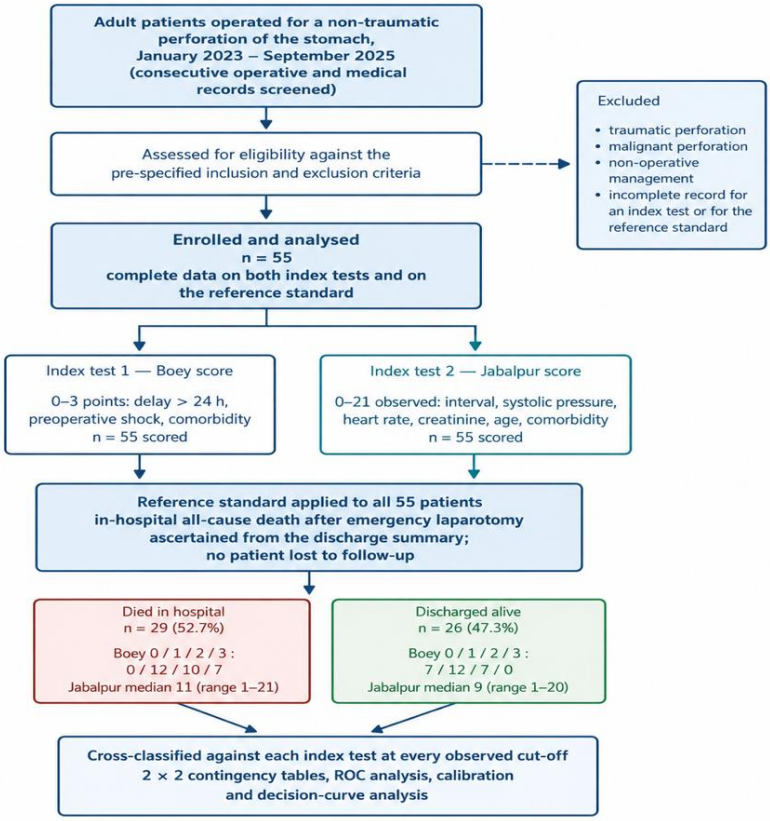

Figure 1. STARD 2015 flow diagram of patient selection, index-test computation and application of the reference standard. Both index tests were computed for every enrolled patient and the reference standard was available for all of them.

3.2 Distribution of the two scores

Table 2 gives the distribution of both scores and the mortality observed within each stratum. The Boey score was higher among those who died (mean 1.83 ± 0.80 versus 1.00 ± 0.75, Mann-Whitney p=0.001). The Jabalpur score was also higher among those who died, but not significantly so (mean 11.66 ± 5.31 versus 9.58 ± 5.62, p=0.118).

Mortality rose monotonically across the four Boey strata, as displayed in Figure 3A: none of the 7 patients scoring 0 died, 12 of 24 (50.0%, 95% CI 31.4-68.6) scoring 1 died, 10 of 17 (58.8%, 36.0-78.4) scoring 2 died, and all 7 scoring 3 died (64.6-100.0). The Cochran-Armitage test for trend gave z=3.526 (p<0.001). The gradient across the Jabalpur bands, shown in Figure 3B, was shallower and non-monotonic at the lower end: 3 of 8 (37.5%) in the 0-4 band, 6 of 17 (35.3%) in 5-9, 9 of 14 (64.3%) in 10-14 and 11 of 16 (68.8%) in 15-21, with a trend statistic of z=2.064 (p=0.039). The two extreme Boey strata separate the cohort completely; no Jabalpur band does.

Table 2. Distribution of both index tests and the in-hospital mortality observed within each stratum, with the mortality each stratum is expected to carry.

Stratum	n	Deaths	Mortality (%)	95% CI (%)	Published mortality (%)
Boey score					
Boey 0	7	0	0.0	0.0-35.4	0
Boey 1	24	12	50.0	31.4-68.6	10
Boey 2	17	10	58.8	36.0-78.4	45
Boey 3	7	7	100.0	64.6-100.0	100
Jabalpur score band					
0-4	8	3	37.5	13.7-69.4	not published
5-9	17	6	35.3	17.3-58.7	not published
10-14	14	9	64.3	38.8-83.7	not published
15-21	16	11	68.8	44.4-85.8	not published

Notes: Confidence intervals are Wilson score intervals. Published mortality for the Boey score is the original four-level table; no mortality table has been published for the Jabalpur bands. Cochran-Armitage test for trend: Boey z = 3.526 (p<0.001); Jabalpur z = 2.064 (p=0.039).

3.3 Discrimination and the head-to-head comparison

The receiver operating characteristic curves of both scores are drawn on common axes in Figure 2 and the head-to-head comparison is tabulated in Table 3. The Boey score achieved an area of 0.747 (95% CI 0.630-0.863), significantly above chance both by the DeLong z statistic (z=4.140, p<0.001) and by a two-sided permutation test (p=0.001). The Jabalpur score achieved 0.623 (95% CI 0.470-0.777), which was not distinguishable from

chance (z=1.578, p=0.115; permutation p=0.118). Its confidence interval includes 0.5. At the area observed the study had only 34.1% power, so this is an inconclusive result and not a demonstration that the Jabalpur score fails to discriminate.

The difference between the two areas is 0.123 in favour of the Boey score, and that difference is identified exactly. Its standard error is not, for the reason set out in section 4.5. Evaluated across the pairings of the two scores the DeLong test returns z=3.089 (p=0.002) when the two scores are as concordant as their marginal distributions allow, z=1.255 (p=0.210) under independence, and z=0.935 (p=0.350) at maximal discordance, so the p value lies somewhere in 0.002 to 0.350. That interval is an outer set rather than a sharp bound, and the reason is worth stating precisely: the rearrangement inequality maximises and minimises the covariance of the two placement vectors without regard to whether the pairing it produces is clinically possible. It is not always possible here. A Boey score of 3 requires all three components including a comorbidity, which alone contributes 5 Jabalpur points, so a patient with a Boey score of 3 must have a Jabalpur score of at least 5; yet 3 of the decedents have Jabalpur totals below 5, and the discordant extreme assigns some of them to the 7 decedents scoring 3 on the Boey scale. The true admissible set is therefore narrower than the interval quoted, and the true p value closer to its lower end, but no sharp value can be recovered from aggregate output. What is not in doubt is the direction of the difference and the fact that only one of the two areas is separable from chance.

Table 3. Head-to-head diagnostic accuracy of the Boey and Jabalpur scores for in-hospital mortality.

Parameter	Boey score	Jabalpur score	Comparison
Area under the ROC curve (95% CI)	0.747 (0.630-0.863)	0.623 (0.470-0.777)	see final row
p versus an area of 0.5, DeLong	<0.001	0.115	
p versus an area of 0.5, two-sided permutation	0.001	0.118	
Power at the observed area	92.0%	34.1%	
Youden-optimal cut-off	2	10	
Youden index at that cut-off	0.317	0.305	
Sensitivity, % (95% CI)	58.6 (40.7-74.5)	69.0 (50.8-82.7)	
Specificity, % (95% CI)	73.1 (53.9-86.3)	61.5 (42.5-77.6)	
Positive predictive value, % (95% CI)	70.8 (50.8-85.1)	66.7 (48.8-80.8)	
Negative predictive value, % (95% CI)	61.3 (43.8-76.3)	64.0 (44.5-79.8)	
Positive likelihood ratio (95% CI)	2.18 (1.08-4.40)	1.79 (1.04-3.09)	
Negative likelihood ratio (95% CI)	0.57 (0.35-0.93)	0.50 (0.27-0.94)	
Overall accuracy, %	65.5	65.5	
Odds ratio per point (95% CI)	3.972 (1.687-9.353)	1.074 (0.971-1.188)	
Cross-validated area under the curve	0.659	0.538	
Cross-validated calibration slope (SD over 50 repetitions)	0.824 (0.056)	-0.048 (0.241)	
Brier score	0.2597	0.2400	
Scaled Brier score	-0.042	0.037	
Hosmer-Lemeshow-type statistic[a]	43.979 on 2 df, p < 0.001	not applicable	
Difference in areas (Boey minus Jabalpur)	**0.123**		**DeLong p within 0.002 to 0.350**

Notes: Operating characteristics are given at each score's Youden-optimal cut-off. Confidence intervals for proportions are Wilson score intervals and those for likelihood ratios are Katz log-normal intervals. The difference between the two areas is exactly identified; its variance is not, because it depends on the joint distribution of the two scores within a patient, which aggregate output does not fix. The DeLong p value is therefore given as an interval evaluated over the pairings of the two placement vectors (concordant p=0.002, independent p=0.210, discordant p=0.350). That interval is an OUTER set, not a sharp bound: the discordant extreme includes pairings the two scores' shared components forbid (section 3.3). [a]Reported because it is conventionally requested; it is not a valid goodness-of-fit test for an externally specified model evaluated on its own strata (see section 3.5).

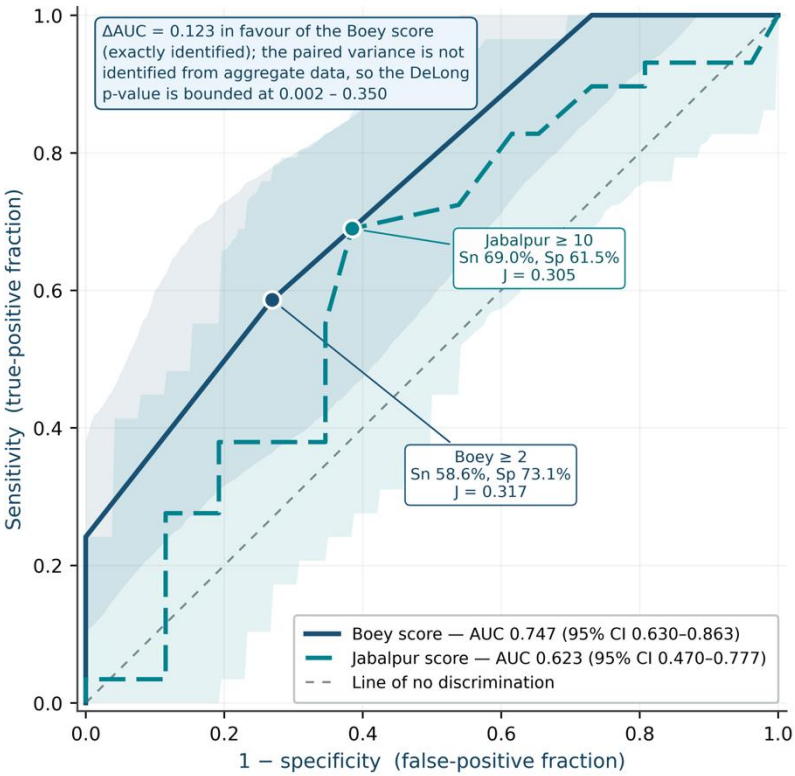

Figure 2. Receiver operating characteristic curves of the Boey and Jabalpur scores for in-hospital mortality after emergency surgery for gastric perforation, drawn on common axes. Filled circles mark the Youden-optimal cut-off of each score. The shaded bands are pointwise 95% bootstrap confidence bands on sensitivity at a fixed false-positive fraction, from 2,000 replicates in which survivors and decedents were resampled independently with replacement.

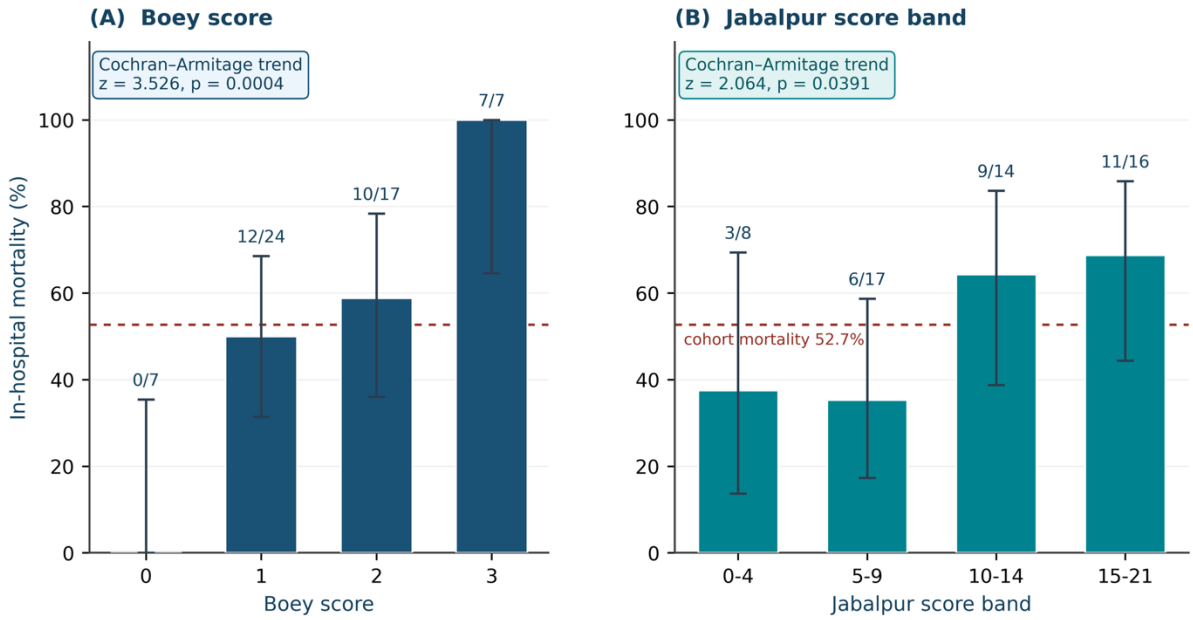

Figure 3. In-hospital mortality by (A) Boey score and (B) Jabalpur score band. Error bars are Wilson score 95% confidence intervals, the fraction above each bar gives deaths over patients in that stratum, and the dashed line marks the overall cohort mortality of 52.7%.

3.4 Operating characteristics at every cut-off

Table 4 gives sensitivity, specificity, predictive values and likelihood ratios at every observed cut-off of both scores. Three points on the Boey scale are of clinical interest. At a cut-off of 1 – the threshold applied in the source study – sensitivity was 100.0% (95% CI 88.3-100.0%) but specificity only 26.9% (13.7-46.1%): every death occurred in a patient with at least one Boey point, so a score of 0 excluded death with a negative predictive value of 100%, but 48 of 55 patients met the threshold and the positive predictive value was only 60.4%. At the Youden-optimal cut-off of 2 (Youden index 0.317), sensitivity was 58.6% (40.7-74.5%) and specificity 73.1% (53.9-86.3%), with a positive likelihood ratio of 2.18. At a cut-off of 3, specificity and positive predictive value were both 100%: all 7 patients with the maximum score died. Both statements about the ends of the scale rest on 7 patients each and their intervals are correspondingly wide – 64.6% to 100% for the mortality of a score of 3 and 0% to 35.4% for a score of

0 – so the extremes are informative without being certain, and the sensitivity at a cut-off of 3 is only 24.1%.

The Jabalpur score reached its maximum Youden index of 0.305 at a cut-off of 10, which coincides with the boundary between the second and third of the four bands set out in the study protocol; those bands are the protocol's own and no mortality has ever been published against them. There sensitivity was 69.0% (50.8-82.7%) and specificity 61.5% (42.5-77.6%), giving a positive likelihood ratio of 1.79 and a negative likelihood ratio of 0.50. No cut-off of the Jabalpur score achieved a Youden index above 0.305, and none achieved a likelihood ratio far enough from unity to change management. Each score's Youden maximum was attained at a single cut-off, so no plateau needed to be reported.

Table 4. Operating characteristics of both index tests at every observed cut-off.

Cut-off	TP / FP / FN / TN	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	LR+	LR-	Youden
Boey								
≥ 1	29 / 19 / 0 / 7	100.0 (88.3-100.0)	26.9 (13.7-46.1)	60.4 (46.3-73.0)	100.0 (64.6-100.0)	1.37	0.00	0.269
≥ 2	17 / 7 / 12 / 19	58.6 (40.7-74.5)	73.1 (53.9-86.3)	70.8 (50.8-85.1)	61.3 (43.8-76.3)	2.18	0.57	0.317
≥ 3	7 / 0 / 22 / 26	24.1 (12.2-42.1)	100.0 (87.1-100.0)	100.0 (64.6-100.0)	54.2 (40.3-67.4)	∞	0.76	0.241
Jabalpur								
≥ 3	27 / 25 / 2 / 1	93.1 (78.0-98.1)	3.8 (0.7-18.9)	51.9 (38.7-64.9)	33.3 (6.1-79.2)	0.97	1.79	-0.031
≥ 4	27 / 21 / 2 / 5	93.1 (78.0-98.1)	19.2 (8.5-37.9)	56.2 (42.3-69.3)	71.4 (35.9-91.8)	1.15	0.36	0.123
≥ 5	26 / 21 / 3 / 5	89.7 (73.6-96.4)	19.2 (8.5-37.9)	55.3 (41.2-68.6)	62.5 (30.6-86.3)	1.11	0.54	0.089
≥ 6	26 / 19 / 3 / 7	89.7 (73.6-96.4)	26.9 (13.7-46.1)	57.8 (43.3-71.0)	70.0 (39.7-89.2)	1.23	0.38	0.166
≥ 7	24 / 17 / 5 / 9	82.8 (65.5-92.4)	34.6 (19.4-53.8)	58.5 (43.4-72.2)	64.3 (38.8-83.7)	1.27	0.50	0.174
≥ 8	24 / 16 / 5 / 10	82.8 (65.5-92.4)	38.5 (22.4-57.5)	60.0 (44.6-73.7)	66.7 (41.7-84.8)	1.34	0.45	0.212
≥ 9	21 / 14 / 8 / 12	72.4 (54.3-85.3)	46.2 (28.8-64.5)	60.0 (43.6-74.4)	60.0 (38.7-78.1)	1.34	0.60	0.186
≥ 10	20 / 10 / 9 / 16	69.0 (50.8-82.7)	61.5 (42.5-77.6)	66.7 (48.8-80.8)	64.0 (44.5-79.8)	1.79	0.50	0.305
≥ 11	16 / 9 / 13 / 17	55.2 (37.5-71.6)	65.4 (46.2-80.6)	64.0 (44.5-79.8)	56.7 (39.2-72.6)	1.59	0.69	0.206
≥ 12	13 / 9 / 16 / 17	44.8 (28.4-62.5)	65.4 (46.2-80.6)	59.1 (38.7-76.7)	51.5 (35.2-67.5)	1.30	0.84	0.102
≥ 13	11 / 9 / 18 / 17	37.9 (22.7-56.0)	65.4 (46.2-80.6)	55.0 (34.2-74.2)	48.6 (33.0-64.4)	1.10	0.95	0.033
≥ 14	11 / 6 / 18 / 20	37.9 (22.7-56.0)	76.9 (57.9-89.0)	64.7 (41.3-82.7)	52.6 (37.3-67.5)	1.64	0.81	0.149
≥ 15	11 / 5 / 18 / 21	37.9 (22.7-56.0)	80.8 (62.1-91.5)	68.8 (44.4-85.8)	53.8 (38.6-68.4)	1.97	0.77	0.187
≥ 16	8 / 5 / 21 / 21	27.6 (14.7-45.7)	80.8 (62.1-91.5)	61.5 (35.5-82.3)	50.0 (35.5-64.5)	1.43	0.90	0.084
≥ 17	8 / 3 / 21 / 23	27.6 (14.7-45.7)	88.5 (71.0-96.0)	72.7 (43.4-90.3)	52.3 (37.9-66.2)	2.39	0.82	0.160
≥ 18	6 / 3 / 23 / 23	20.7 (9.8-38.4)	88.5 (71.0-96.0)	66.7 (35.4-87.9)	50.0 (36.1-63.9)	1.79	0.90	0.092
≥ 20	1 / 3 / 28 / 23	3.4 (0.6-17.2)	88.5 (71.0-96.0)	25.0 (4.6-69.9)	45.1 (32.3-58.6)	0.30	1.09	-0.081
≥ 21	1 / 0 / 28 / 26	3.4 (0.6-17.2)	100.0 (87.1-100.0)	100.0 (20.7-100.0)	48.1 (35.4-61.1)	∞	0.97	0.034

Notes: TP, true positive; FP, false positive; FN, false negative; TN, true negative; PPV, positive predictive value; NPV, negative predictive value; LR+, positive likelihood ratio; LR-, negative likelihood ratio; ∞ denotes a likelihood ratio that is not finite because the corresponding false-positive or false-negative count is zero. Every observed cut-off is listed so that the reported optimum is not a selectively reported threshold. The Youden maximum is 0.317 at Boey ≥ 2 and 0.305 at Jabalpur ≥ 10; each maximum was attained at a single cut-off.

3.5 Calibration of the published Boey risk table

Discrimination describes ranking; it says nothing about whether the absolute risks a score quotes are correct[27]. Table 5 lists the published Boey mortality table against what was observed and Figure 4A plots the same comparison. Applying the published risks of 0%, 10%, 45% and 100% to the observed strata predicts 17.05 deaths against 29 observed, a ratio of observed to expected of 1.70. The discrepancy is concentrated almost entirely in one stratum: among the 24 patients scoring 1, the table predicts 2.4 deaths and 12 occurred (exact binomial $p<0.001$). Level 2 was well predicted ($p=0.330$). Levels 0 and 3 admit no informative test, because a published risk of exactly 0% or 100% carries no sampling variability; what can be said is that the observations at both ends agree with the table, none of the 7 patients scoring 0 having died and all 7 scoring 3 having done so.

Two summary measures of calibration are reported, and both need a caveat that is easy to omit. Only two of the four Boey strata carry a published risk strictly between 0 and 1, so both the calibration slope and calibration-in-the-large are estimated from those two strata alone, on 41 of the 55 patients. With two design points and two free parameters the slope of 0.179 is a saturated fit that passes exactly through both observed proportions; it is a description of how much flatter the observed gradient is than the published one, not a test with residual degrees of freedom. Calibration-in-the-large with the slope fixed at unity was 1.581 on the same 41 patients, corresponding to a 4.86-fold increase in the odds of death over the published risk; within that subset the ratio of observed to expected deaths is 2.19 rather than the whole-cohort 1.70.

The global test does not depend on those two strata. Convolving the four stratum-specific binomial distributions

gives the exact probability of observing 29 or more deaths under the published risks as 9.19e-06; in 400,000 simulated cohorts the largest death count reached was 29, attained 4 times. The Brier score was 0.2597 against a no-information value of 0.2493, so the scaled Brier score was -0.042: in this cohort, quoting the published Boey risk to a family was slightly worse than quoting the overall unit mortality of 52.7%.

A Hosmer-Lemeshow-type statistic on the two informative levels was 43.979 on 2 degrees of freedom (p<0.001). This is reported because it was requested, not because it is valid: the test was designed for a model fitted in the data being tested, its strata here are the score's own fixed levels rather than data-derived deciles, and its reference distribution does not hold for an externally specified model[27]. The exact binomial tests within strata and the exact global tail above are the defensible substitutes and reach the same conclusion.

No published risk table exists for the Jabalpur bands, so only internal calibration could be assessed, and it is plotted by risk quartile in Figure 4B. Refitted within this cohort the Jabalpur score carried an odds ratio of 1.074 per point (95% CI 0.971-1.188), an interval that includes unity. Its mean cross-validated calibration slope was -0.048, but that mean is not a stable quantity: across the fifty repetitions it had a standard deviation of 0.241 and ranged from -0.510 to 0.586, so it should be read as showing that the score carries no reliably transportable signal in this sample rather than as an estimate of a slope. The Boey score carried an odds ratio of 3.972 per point (95% CI 1.687-9.353), a cross-validated slope of 0.824 (SD 0.056) and a cross-validated area of 0.659 against 0.538 for the Jabalpur score.

Table 5. External calibration of the published Boey mortality table in this cohort.

Stratum	n	Published risk (%)	Expected deaths	Observed deaths	Observed mortality, % (95% CI)	Exact binomial p
Boey 0	7	0	0.00	0	0.0 (0.0-35.4)	1.000
Boey 1	24	10	2.40	12	50.0 (31.4-68.6)	<0.001
Boey 2	17	45	7.65	10	58.8 (36.0-78.4)	0.330
Boey 3	7	100	7.00	7	100.0 (64.6-100.0)	1.000
Total	**55**		**17.05**	**29**	**52.7**	**O:E 1.70**

Notes: Exact binomial tests compare the observed number of deaths in each stratum with the number the published risk predicts; a stratum whose published risk is 0% or 100% admits no informative test and is shown as 1.000 where the observation agrees with it. Calibration slope 0.179; calibration-in-the-large with the slope fixed at unity 1.581 (odds x 4.86); Brier score 0.2597 against a no-information value of 0.2493, scaled Brier -0.042. Convolving the four stratum-specific binomial distributions gives the exact probability of 29 deaths or more under the published risks as 9.19e-06. The calibration slope and calibration-in-the-large are estimated from levels 1 and 2 only (41 patients), the two strata whose published risk lies strictly between 0 and 1; with two design points the slope is a saturated fit and carries no residual degrees of freedom.

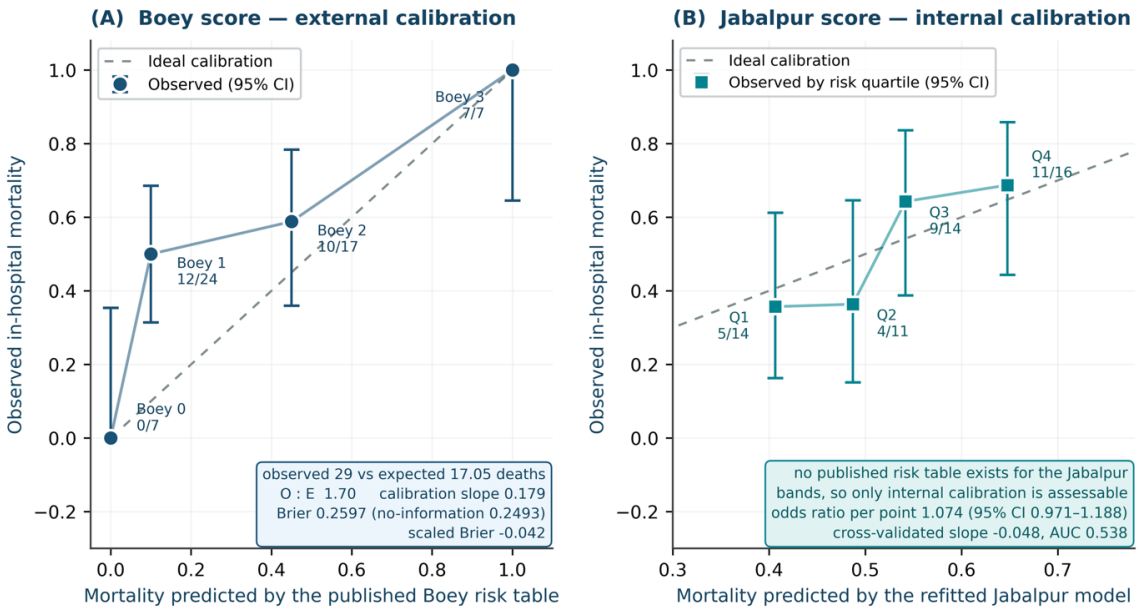

Figure 4. Calibration of the two index tests. (A) The Boey score against its own published mortality table, an external calibration. (B) The Jabalpur score against risks refitted within this cohort by quartile, an internal calibration, since no mortality table has been published for its bands.

3.6 Anatomical distribution of the perforation

The site of the breach was recorded in every patient and is displayed in Figure 5A. Forty-three of the 55 perforations (78.2%) lay in the prepyloric segment of the anterior gastric wall, 10 (18.2%) were recorded as postpyloric, one was at the pylorus itself and one in the body of the stomach. Mortality by site was 21 of 43 (48.8%) for prepyloric, 7 of 10 (70.0%) for postpyloric, 1 of 1 at the pylorus and 0 of 1 in the body, and the differences were not significant (chi-square 3.470, df 3, p=0.325; prepyloric site versus every other site, odds ratio for death 0.48, Fisher p=0.339). With 43 of the 55 perforations at a single site the study has very little power to detect such a difference, and the anatomical result should be read as descriptive. The route by which a breach at that site becomes a systemic illness, and the point at which each score variable enters that route, is traced in Figure 5B.

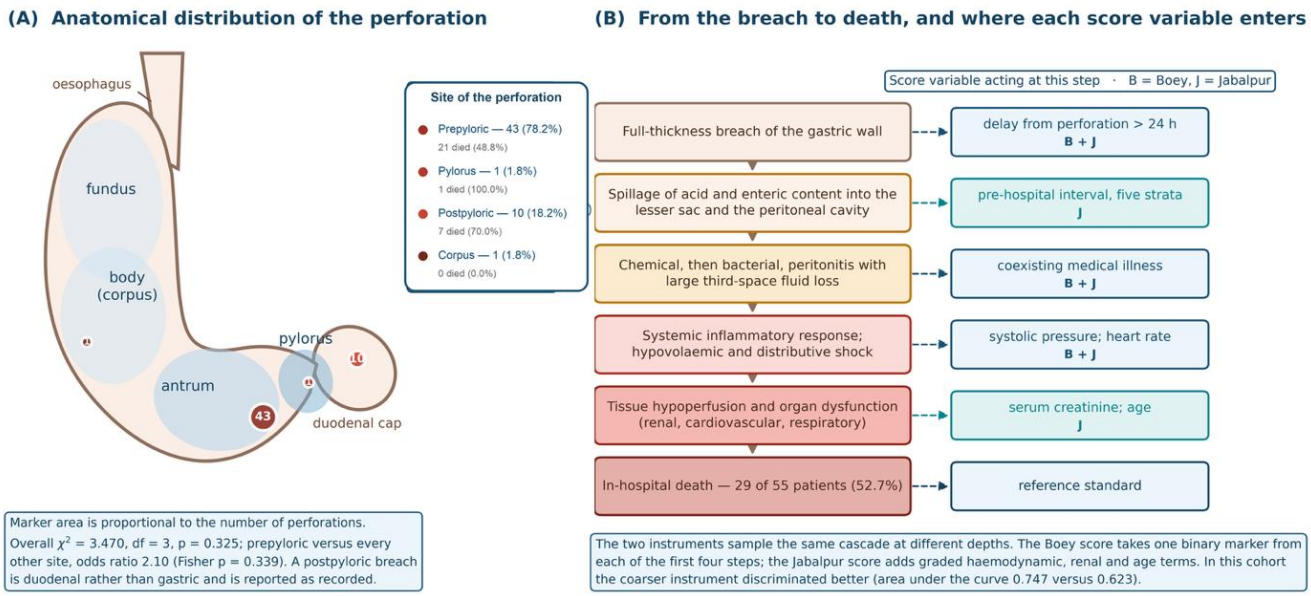

Figure 5. (A) Anatomical distribution of the perforation in the 55 patients, with the mortality observed at each site. (B) The cascade from a full-thickness breach of the gastric wall to death, with the variable each index test contributes at every step.

3.7 Clinical usefulness

Decision-curve analysis of the two prescribed decision rules is shown in Figure 6A-B. Neither rule is useful across the whole range of threshold probabilities. Treating every patient as high risk carries the greatest net benefit below a threshold probability of about 0.39, which is unsurprising in a cohort whose mortality is 52.7%: when more than half of the patients die, a policy of maximal support for everyone is hard to beat. The rule Boey ≥2 exceeds both treat-all and treat-none over threshold probabilities of 0.39 to 0.70, and the rule Jabalpur ≥10 over 0.37 to 0.66. Where the two rules are compared with each other, Boey ≥2 carries the higher net benefit above a threshold of 0.51 and the Jabalpur rule is marginally higher below it.

The practical reading is narrower than a decision curve of the envelope over all cut-offs would suggest, and deliberately so: an envelope chosen in the same data is optimistic[27]. What the fixed rules support is that in the region where a surgeon is genuinely undecided – roughly a two-in-five to two-in-three probability of death, the region in which the question is whether to commit a scarce intensive care bed – either score adds something over treating everyone alike, and above about one in two the Boey rule adds more[27]. Below that region neither score should displace maximal support for all.

3.8 The Boey score rebuilt from its components: bounds

The Boey score analysed above is the score as it was computed and analysed in the source output. It is also possible to ask what the score would look like if it were rebuilt from the three component variables as they were separately tabulated: preoperative shock in 12 patients (2 survivors, 10 deaths), comorbidity in 25 (10, 15) and a delay beyond twenty-four hours in 49 (22, 27). Those margins do not fix which patient carried which combination, so the distribution of the score is only partially identified, and they are also not unique: the source records the delay in two different ways, as a binary variable at twenty-four hours (6 patients within the window) and as the five-stratum pre-hospital interval used by the Jabalpur score (3 patients within it). Both readings are carried through.

Under the binary delay variable, 495 joint distributions are admissible and the area under the curve is bounded between 0.615 and 0.756; under the five-stratum interval, 198 are admissible and the bounds are 0.610 to 0.732. Taking the union, the component-based area lies between 0.610 and 0.756 (Figure 6B). This is a wider and less comfortable result than the analysed value alone, and it matters: the lower end of the region, 0.610, falls **below** the Jabalpur point estimate of 0.623. The ordering of the two scores is therefore not guaranteed by the component reading, even though it holds at the analysed value and across most of the admissible set. That is stated here rather than buried, because an earlier draft of this analysis asserted that the direction was robust to this uncertainty, and it is not.

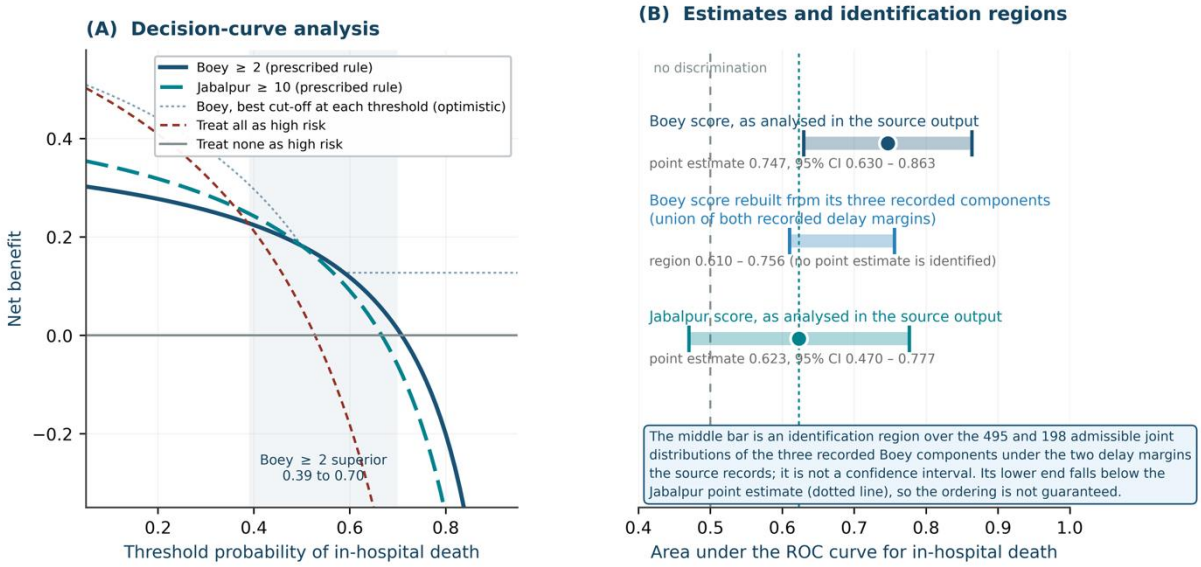

Figure 6. (A) Decision-curve analysis of both index tests against the treat-all and treat-none strategies. (B) Point estimates with 95% confidence intervals for the two areas under the curve, together with the sharp identification region for the Boey score rebuilt from its three separately tabulated components.

4. Discussion

4.1 Principal findings

In 55 consecutive adults operated for non-traumatic gastric perforation at the tertiary surgical referral centre for South Sumatera, the three-variable Boey score separated survivors from decedents and the six-variable Jabalpur score did not. The areas under the curve were 0.747 and 0.623 respectively; only the first is distinguishable from chance. Mortality across the four Boey strata was 0%, 50.0%, 58.8% and 100%, a monotone gradient with a trend statistic of $p<0.001$, and the two extreme strata separated the cohort completely. The full head-to-head comparison is set out in Table 3.

Two qualifications belong in the same breath as that finding, and neither is a formality. The difference between the two areas, 0.123, is exactly identified, but its standard error is not, so the formal comparison of the two scores is unresolved: the DeLong p lies somewhere between 0.002 and 0.350 depending on how the two scores are paired within patients. And the Jabalpur null is inconclusive rather than negative: at the area observed the study had 34.1% power. What this cohort supports is that the Boey score works here and that the Jabalpur score has not been shown to; it does not support a formal claim of superiority.

4.2 Discrimination in the context of published validations

An area of 0.747 for the Boey score sits comfortably within the published range, and the operating characteristics behind that estimate are set out at every cut-off in Table 4, with the curves themselves drawn in Figure 2. The score has been evaluated against mortality after perforated peptic ulcer in an Indonesian series from Dr. M. Djamil Padang[32], prospectively in Nepal[16], in two Indian series[17,18] and in two Egyptian ones[19,20], and against the Mannheim Peritonitis Index in India[33]; mortality rises across the four strata in each. A 2026 systematic review and meta-analysis of the Boey score against the PULP score in perforated peptic ulcer[34] and a head-to-head comparison of the ASA, Boey and PULP scores published the same year[35] both retain the Boey score as a reference instrument, and an earlier comparison against the POMPP score reached the same conclusion[36]. The agreement is reassuring, because it suggests that a score built in Hong Kong in the 1980s still ranks Indonesian patients correctly four decades later, in a case mix with a far higher baseline mortality.

The comparison closest to ours in setting is the 2025 study of 31 patients at this hospital[29]. At a Boey cut-off of 2 it reported a sensitivity of 75.0% and a specificity of 53.3%, where the same threshold in the cohort reported here gives 58.6% and 73.1% (Table 4). The two series were assembled at the same centre over adjacent periods, so the difference between them is a difference in case mix and in threshold behaviour within one referral stream rather than evidence about how the score travels between centres.

The Jabalpur result is harder to reconcile with its literature, and the first thing to say about it is that our study was not well placed to detect a moderate effect: at the area observed the power was only 34.1%, and the smallest area this cohort could detect at 80% power was 0.720. The finding is an absence of evidence at a resolution the study can support, not evidence of absence. With that said, its developers reported a sensitivity of 87% and a specificity of 85%[21], and later series have reported sensitivities from 69% to 95%[22,23,25,26], whereas our sensitivity of 69.0% at the same threshold falls at the bottom of that range and our specificity of 61.5% below it.

Two explanations for the gap are worth separating from the question of power. The first is case mix, and it is weaker than it first appears. The derivation paper is titled for peptic ulcer perforation[21], so the index cannot simply be dismissed as an instrument built for a different disease; what has changed is the company it keeps, since most of its subsequent use has been in undifferentiated perforation peritonitis, in which ileal and appendicular perforations are prominent[10,22,24,37]. The same is true of the score with which it is most often compared: the Mannheim Peritonitis Index and its recent combinations and

rivals have almost all been evaluated in undifferentiated perforation peritonitis or in digestive tract perforation at large[38,39,40,41]. Our cohort is restricted to gastric perforation and excludes malignancy, which is a narrower and physiologically more homogeneous population, and homogeneity in the predictors is exactly what flattens a discrimination statistic[28].

There is also a structural reason why more variables need not help. Every additional term in a score is an additional opportunity for a value to be missing, mis-banded or recorded at a different moment of the resuscitation. The Jabalpur score bands systolic pressure and heart rate non-monotonically – a systolic pressure of 130-159 mmHg scores 3 while 70-109 mmHg scores 0 – so a single reading taken after fluid resuscitation rather than on arrival can move a patient two points in either direction. The Boey score's three dichotomies are far more robust to when the observation was made. In small samples a coarse, stable instrument frequently outperforms a fine, unstable one[27,28], and that is what was observed here.

4.3 Discrimination is not calibration

The Boey score ranked patients well and quantified their risk badly, and the two findings must be kept apart[27]. Applying the published mortality table to this cohort predicts 17.05 deaths where 29 occurred. Under those published risks the exact probability of observing 29 deaths or more, obtained by convolving the four stratum-specific binomial distributions, is 9.19e-06. The scaled Brier score of -0.042 makes the practical consequence plain: a surgeon who told the family of a patient with a Boey score of 1 that the risk of death was 10% would have been less accurate than one who simply quoted the unit's overall mortality of 52.7%.

The miscalibration is not uniform, which is what makes it interesting. Scores of 0, 2 and 3 were predicted well; the whole of the discrepancy lies at a score of 1, where 12 of 24 patients died against a published expectation of 2.4. A single Boey point in Hong Kong in 1987 evidently did not mean what a single Boey point means in Palembang in 2024. The most likely reason is visible in Table 1: the commonest way to score a single point in this cohort was a delay beyond twenty-four hours, present in 49 of 55 patients, and the pre-hospital interval here extends far beyond anything contemplated in the original series – 13 patients reached hospital more than 120 hours after onset, of whom 12 died. A dichotomy at twenty-four hours cannot distinguish a patient who arrived at hour twenty-six from one who arrived on day six. The stratum-by-stratum discrepancy is set out in Table 5.

The implication for practice is specific. The ordering the Boey score produces can be used; the percentages it quotes cannot, at least not without local recalibration. We deliberately do not propose a multiplicative recalibration factor derived from the observed ratio of observed to expected deaths. Such a rule would be fitted and evaluated in the same 55 patients, it would be dominated by the single stratum in which the score fails, and applied to the top stratum it would predict a mortality above 100%[27,28]. What is needed is a locally refitted table from a larger multicentre Indonesian series, not an arithmetic correction to this one.

4.4 The anatomical dimension

Perforations in this cohort were overwhelmingly prepyloric: 43 of 55 (78.2%), with 10 recorded as postpyloric, one at the pylorus itself and one in the body of the stomach (Figure 5A). This distribution is the expected one. The prepyloric segment of the anterior gastric wall is where acid-peptic injury concentrates, the mucosa there is thinner than in the fundus or the body, and the region lies at the watershed of the right and left gastroepiploic territories[1,2,42]. The single corpus perforation in the series is consistent with how uncommon that site is outside malignancy, and malignant perforation was an exclusion criterion.

Site was not associated with mortality (chi-square 3.470, df 3, p=0.325; prepyloric site versus all other sites, odds ratio for death 0.48, Fisher p=0.339), and with 43 of the perforations at one site the study has almost no power to detect such an association. The anatomical contribution to this paper is therefore explanatory rather than predictive, and it is worth stating what it explains. A prepyloric anterior breach discharges directly into the supracolic compartment of the greater sac and tracks down the right paracolic gutter, so contamination is generalised from the outset and its extent is governed by elapsed time rather than by site. That is precisely why the delay variable dominates both scores, and why a score that measures elapsed time crudely still works while one that measures haemodynamics finely does not: at this site, in this population, time is the dominant determinant of the peritoneal insult. Early source control remains the single intervention with the largest effect on that insult[14], which is why the delay terms of both scores measure something the surgeon can still act on. The mapping of each score variable onto the successive steps of that cascade is drawn in Figure 5B.

4.5 Measurement of the index tests

Three measurement issues qualify the interpretation. The first is that scoring was retrospective. Both index tests were computed after discharge from a record in which the outcome was already present, so the blinding that STARD asks for could not be achieved[30]. The mitigation is that every component is an objective measurement recorded contemporaneously for clinical purposes – a blood pressure, a pulse, a creatinine, a documented time of onset – rather than a judgement made at the time of scoring. It is not a complete mitigation: the banding of a borderline value is a judgement, and a reviewer who knows the outcome may band it differently.

The second is that the Jabalpur rubric as reproduced in the source protocol contains bands that do not tile the real line: heart rates between 130 and 139 per minute, systolic pressures between 160 and 179 mmHg, creatinine values below 0.6 mg/dL and an age of exactly 75 years all fall between defined categories. When the score is rebuilt from the recorded components under the most natural reading of those gaps the totals differ from the analysed totals by about 11% among survivors and 3% among decedents. Some of the Jabalpur score's poor showing may therefore be attributable to ascertainment rather than to the instrument, and a prospective application with a fully specified rubric would be a fairer test of it.

The third is that the record itself contains at least one impossible value: a heart rate of 0 per minute in a patient who

underwent laparotomy. That value is treated as missing here, but its presence is a reminder that the Jabalpur score's four physiological domains are only as good as the observations entered into the chart, and that a score with more domains inherits more of that noise than a score with three binary ones. The corresponding data-quality audit was not performed in the earlier analysis and should be routine.

4.6 Implications for practice

For a surgeon receiving a patient with a gastric perforation at Dr. Mohammad Hoesin Central General Hospital, Palembang or a comparable Indonesian centre, three recommendations follow, each stated with the precision the data allow.

First, calculate the Boey score at admission. It takes under a minute, requires no laboratory result, and it is the only one of the two instruments that separated survivors from decedents here. That is not the same as saying it is proven superior to the Jabalpur score, and the paper does not say so.

Second, treat the extremes of the scale as informative and the middle as uninformative. All 7 patients with a score of 3 died, and none of the 7 with a score of 0 did. Those are the strongest signals in the data and they are also the least precise: 7 of 7 gives a 95% interval of 64.6% to 100%, and 0 of 7 an interval of 0% to 35.4%. A score of 3 is a reason to secure the highest available level of perioperative support – early source control, protocolised resuscitation and an intensive care bed booked before induction[42,43] – and to speak to the family before operation; it is not a reason to withhold treatment, and the ceiling of 35.4% at the other end of the scale means a score of 0 should reassure rather than reclassify. The intermediate scores of 1 and 2 carried mortality of 50.0% and 58.8% and are not separated by either score; for those patients the score should not be the deciding input.

Third, do not quote the published Boey percentages to families. Quote the local stratum-specific figures in Table 2 with their confidence intervals, and recognise that at a score of 1 the local risk is five times the published one. This is a general point about importing a risk table across four decades and two continents, and it applies to any of the scores in this literature[27,28]. The decision-curve analysis sharpens the same point: below a threshold probability of about 0.39 neither score improves on treating every patient as high risk, which in a cohort with 52.7% mortality is the honest default.

Underlying all three is the finding in Table 1 that most strongly resists any score. Twelve of the thirteen patients who reached hospital more than 120 hours after the onset of symptoms died. No preoperative score can recover a delay of that magnitude, and the largest available gain in survival at this centre lies in the referral pathway rather than in the choice of scoring instrument[2,9].

These three recommendations are consistent with what the wider literature identifies as the dominant determinants of outcome after peptic ulcer perforation. Preoperative shock, delay between onset and operation, and advanced age recur as independent predictors of early death across series from several health systems[1,36,42,44] – and all three are variables the Boey score already carries. A score that adds terms without adding any of these is unlikely to add much.

4.7 Strengths

The cohort is consecutive and complete: total sampling within a defined accrual window, no missing outcome and no loss to follow-up, and a single unambiguous reference standard requiring no adjudication. Both index tests were available for every patient, so the two scores are compared on identical patients rather than on overlapping subsets. The analysis is reported to STARD 2015 with a complete participant flow diagram, and operating characteristics are given at every observed cut-off rather than only at the one that performed best, which removes the optimism of a selectively reported threshold. Discrimination is tested by permutation as well as by the DeLong statistic, because the DeLong variance is anti-conservative at this sample size, and the optimal cut-off is reported with the plateau convention. Every quantity in the paper is re-derived by an executable verification harness that recomputes it from the reconstructed data rather than matching it against a stored value, and both the analysis code and that harness are supplied.

4.8 Limitations

The study is retrospective and confined to a single tertiary referral centre, and its case mix – a mortality of 52.7% and a pre-hospital interval extending beyond five days – is not that of a well-resourced elective-referral system. The findings should be generalised to comparable settings and not beyond them. The sample of 55 patients with 29 deaths is small for a comparison of two areas under the curve. It supported 92.0% power at the area observed for the Boey score but only 34.1% at the area observed for the Jabalpur score, so the Jabalpur result is inconclusive; and the confidence interval around the difference remains wide under every pairing. Fifty-five operative cases over thirty-three months is also a low count for a provincial referral centre, and the authors cannot exclude that records were missed at the point of case ascertainment.

4.10 Future work

The obvious next study is a prospective multicentre Indonesian cohort applying both scores with a fully specified rubric at a fixed time point, powered for a difference in areas under the curve of about 0.10 – roughly 46 patients under the most favourable pairing and appreciably more under realistic ones[28]. Such a study should refit the Boey mortality table locally rather than importing it, should record the perforation site prospectively against a fixed anatomical nomenclature, should include malignant perforation so that the age and comorbidity terms of the Jabalpur score are given a fair test, and should extend follow-up to thirty days. It should also record complication grades, so that the morbidity endpoint this analysis could not address becomes available[42].

A second line of work is worth naming because it lies outside what any purely clinical score can reach. Recent series have added information to these models from sources the Boey and Jabalpur scores do not use: admission biomarkers[46], nutritional indices derived from routine blood counts[47], and radiologically measured sarcopenia in older patients[48]. Not one of these was recorded in the source records used here, and they cannot be recovered retrospectively; a prospective cohort at this centre could collect all three at negligible marginal cost, and would be better placed than the present study to say whether anything is gained by extending a bedside score at all.

5. Conclusion

In 55 consecutive adults undergoing emergency laparotomy for non-traumatic gastric perforation at an Indonesian tertiary referral centre, the three-variable Boey score separated survivors from decedents (area under the curve 0.747, 95% CI 0.630-0.863) and the six-variable Jabalpur score did not (0.623, 0.470-0.777), although the latter result was underpowered. The difference of 0.123 favours the Boey score but is not formally resolvable from aggregate data, with a DeLong p bounded within 0.002 to 0.350. Mortality rose monotonically across the Boey strata and the extreme strata separated the cohort completely, so the Boey score is the more defensible bedside instrument in this setting, with a score of 3 read as a reason to secure maximal support and a score of 0 as reassurance rather than reclassification. Its published mortality table, however, underestimated death by a factor of 1.70 and was less accurate than quoting the unit's overall mortality, so local stratum-specific risks should be used when counselling families. A prospective multicentre Indonesian validation with a locally refitted risk table is required.

6. Declarations

Ethics Approval and Consent to Participate

The Research Ethics Committee of RSUP Dr Mohammad Hoesin Palembang, Universitas Sriwijaya, granted an ethical exemption for this study. The ethics reference is DP.04.03/D.XVII.9.4/ETIK/184/2026. The committee waived individual written informed consent because this retrospective study used de-identified records. Institutional access was authorised, and the study followed the Declaration of Helsinki and applicable WHO/CIOMS ethical guidance.

Consent for Publication

Not applicable. The manuscript contains no information or images that identify an individual participant.

Availability of Data and Materials

The data and analysis materials are available from the corresponding author upon reasonable request, subject to institutional and ethics requirements. They are not publicly available because the source records are confidential.

Artificial Intelligence Use Statement

The authors declare that no generative artificial intelligence tools were used in the design, analysis, or writing of this manuscript. The authors reviewed the final manuscript and accept full responsibility for its content.

Author Contribution Statement

SBHM contributed to conceptualization, investigation, data curation, formal analysis, visualization, and preparation of the original draft. EEM contributed to methodology, clinical supervision, validation, interpretation, and critical revision. EB contributed to anatomical interpretation, supervision, validation, and critical revision. All authors reviewed and approved the final manuscript and accept accountability for the work.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest / Competing Interests

The authors declare that they have no competing interests.

Acknowledgements

None.

7. References

1. Shreya A, Sahla S, Gurushankari B, et al. Spectrum of perforated peptic ulcer disease in a tertiary care hospital in South India: predictors of morbidity and mortality. ANZ Journal of Surgery. 2024;94(3):366-370. doi: 10.1111/ans.18831.
2. Bejiga G, Negasa T, Abebe A. Treatment outcome of perforated peptic ulcer disease among surgically treated patients: A cross-sectional study in Adama hospital medical college, Adama, Ethiopia. International Journal of Surgery Open. 2022;48:100564. doi: 10.1016/j.ijso.2022.100564.
3. Pouroushaninia N, Bozorgmehr R, Alibeik N, et al. Surgical Mortality and Morbidity Predictors in Patients With Perforated Peptic Ulcer: A Cohort Study. Journal of Surgical Research. 2025;312:75-82. doi: 10.1016/j.jss.2025.04.038.
4. Sultany A, Rao A, Gondal A, et al. Outcomes and Predictors of Mortality in Perforated Versus Non-Perforated Peptic Ulcer Disease: A U.S. Nationwide Propensity-Matched Analysis, 2016–2021. JCM. 2026;15(11):4358. doi: 10.3390/jcm15114358.
5. Xie X, Ren K, Zhou Z, et al. The global, regional and national burden of peptic ulcer disease from 1990 to 2019: a population-based study. BMC Gastroenterol. 2022;22(1):58. doi: 10.1186/s12876-022-02130-2.
6. Hao W, Zheng C, Wang Z, et al. Global burden and risk factors of peptic ulcer disease between 1990 and 2021: An analysis from the global burden of disease study 2021. PLoS One. 2025;20(7):e0325821. doi: 10.1371/journal.pone.0325821.
7. Li Y, Choi H, Leung K, et al. Global prevalence of Helicobacter pylori infection between 1980 and 2022: a systematic review and meta-analysis. The Lancet Gastroenterology & Hepatology. 2023;8(6):553-564. doi: 10.1016/s2468-1253(23)00070-5.
8. Li H, Shi Q, Chen C, et al. Smoking-attributable peptic ulcer disease mortality worldwide: trends from 1990 to 2021 and projections to 2046 based on the global burden of disease study. Front. Public Health. 2024;12:1465452. doi: 10.3389/fpubh.2024.1465452.
9. An SJ, Davis D, Kayange L, et al. Predictors of mortality for perforated peptic ulcer disease in Malawi. The American Journal of Surgery. 2023;225(6):1081-1085. doi: 10.1016/j.amjsurg.2022.11.029.
10. Shakya VC, Pangeni A, Karki S, et al. Evaluation of Mannheim's Peritonitis Index in Prediction of Mortality in Patients with Non-traumatic Hollow Viscus Perforation Peritonitis. J Nepal Health Res Counc. 2021;19(1):179-184. doi: 10.33314/jnhrc.v19i1.3258.
11. Tartaglia D, Strambi S, Coccolini F, et al. Laparoscopic versus open repair of perforated peptic ulcers: analysis of outcomes and identification of predictive factors of conversion. Updates Surg. 2023;75(3):649-657. doi: 10.1007/s13304-022-01391-6.
12. Kim CW, Kim JW, Yoon SN, et al. Laparoscopic repair of perforated peptic ulcer: a multicenter, propensity score matching analysis. BMC Surg. 2022;22(1):230. doi: 10.1186/s12893-022-01681-1.
13. Chan KS, Ng STC, Tan CHB, et al. A systematic review and meta-analysis comparing postoperative outcomes of laparoscopic versus open omental patch repair of perforated peptic ulcer. J Trauma Acute Care Surg. 2023;94(1):e1-e13. doi: 10.1097/ta.0000000000003799.
14. Reitz KM, Kennedy J, Li SR, et al. Association Between Time to Source Control in Sepsis and 90-Day Mortality. JAMA Surg. 2022;157(9):817. doi: 10.1001/jamasurg.2022.2761.
15. BOEY J, CHOI SKY, ALAGARATNAM TT, et al. Risk stratification in perforated duodenal ulcers: a prospective validation of predictive factors. Annals of Surgery. 1987;205(1):22-32. doi: 10.1097/00000658-198701000-00005.
16. Shrestha S, Shrestha M, Rai P, et al. Predictive validity of the Boey scoring system for postoperative outcomes in perforated peptic ulcer disease: a prospective study from a tertiary center of Nepal. Int Surg J. 2025;12(10):1630-1636. doi: 10.18203/2349-2902.isj20253008.
17. Tudu P, Roy P, Naskar PS. Role of BOEY score in association with age in predicting mortality and morbidity in peptic perforation. Int Surg J. 2022;9(4):841. doi: 10.18203/2349-2902.isj20220944.

18. Jain DU, Chauhan A, Gupta J, et al. Evaluation of Boey scoring in predicting morbidity and mortality in peptic ulcer perforation peritonitis. Int. J. Surg. Sci.. 2021;5(3):41-43. doi: 10.33545/surgery.2021.v5.i3a.735.
19. al-salahi HAA, Deabes SM, Khairy M. Evaluation of Boey's score in predicting morbidity and mortality in peptic ulcer perforation complicated by peritonitis undergoing surgical repair in compare with other scoring system. Al-Azhar International Medical Journal. 2025;2025(10):189-193. doi: 10.21608/aimj.2025.400924.2620.
20. Ghobashy AM, Shafik IA, Milad NM, et al. Comparative performance of Boey, peptic ulcer perforation, and American Society of Anesthesiologists scores in predicting outcomes in patients with perforated peptic ulcer.. The Egyptian Journal of Surgery. 2024;43(4):1268-1277. doi: 10.21608/ejsur.2024.282169.1048.
21. Mishra A, Sharma D, Raina VK. A simplified prognostic scoring system for peptic ulcer perforation in developing countries. Indian J Gastroenterol. 2003;22(2):49-53.
22. Gupta PK, Shrestha PS, Shakya BM, et al. Comparative Performance of POSSUM and Jabalpur Peritonitis Index for Predicting Outcomes after Emergency Laparotomy for Perforation Peritonitis: An Observational Study. Bali J Anesthesiol. 2025;9(4):236-242. doi: 10.4103/bjoa.bjoa_215_25.
23. Ehiagwina LA, Tagar E, Kpolugbo J, et al. Outcome Prediction Following Peritonitis Secondary to Perforated Peptic Ulcer: A Comparison of the Jabalpur Prognostic Scoring System with Mannheim Peritonitis Index in Irrua Specialist Teaching Hospital, Nigeria. J West Afr Coll Surg. 2026;:. doi: 10.4103/jwas.jwas_12_25.
24. Pathak AA, Agrawal V, Sharma N, et al. Prediction of mortality in secondary peritonitis: a prospective study comparing p-POSSUM, Mannheim Peritonitis Index, and Jabalpur Peritonitis Index. Perioper Med. 2023;12(1):65. doi: 10.1186/s13741-023-00355-7.
25. Koranne A, Byakodi KG, Teggimani V, et al. A Comparative Study between Peptic Ulcer Perforation Score, Mannheim Peritonitis Index, ASA Score, and Jabalpur Score in Predicting the Mortality in Perforated Peptic Ulcers. Surg J (N Y). 2022;08(03):e162-e168. doi: 10.1055/s-0042-1743526.
26. Dharmendra BL, Bagalkot VS. Role of Jabalpur Score in Predicting Mortality Among Patients with Peritonitis Secondary to Peptic Ulcer Perforation. Tropical Gastroenterology. 2021;42(3):122-125. doi: 10.7869/tg.638.
27. Cox EGM, Wiersema R, Eck RJ, et al. External Validation of Mortality Prediction Models for Critical Illness Reveals Preserved Discrimination but Poor Calibration. Critical Care Medicine. 2023;51(1):80-90. doi: 10.1097/ccm.0000000000005712.
28. Yamamoto R, Hirakawa S, Tachimori H, et al. Simple severity scale for perforated peptic ulcer with generalized peritonitis: a derivation and internal validation study. International Journal of Surgery. 2024;110(11):7134-7141.
doi: 10.1097/js9.0000000000002037.
29. Bobi Wijaya, Alsen Arlan, Theodorus. Predicting Mortality in Gastric Perforation: A Comparative Analysis of Boey Score and Mannheim Peritonitis Index Accuracy in an Indonesian Tertiary Hospital. sjs. 2025;8(1):840-852. doi: 10.37275/sjs.v8i1.124.
30. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;:h5527. doi: 10.1136/bmj.h5527.
31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837. doi: 10.2307/2531595.
32. Rivai IM, Suchitra A, Janer A. Evaluation of clinical factors and three scoring systems for predicting mortality in perforated peptic ulcer patients, a retrospective study. Annals of Medicine & Surgery. 2021;69:. doi: 10.1016/j.amsu.2021.102735.
33. H. R. H, V. P. M. A comparative study between Mannheims peritonitis index and Boeys score in predicting the morbidity and mortality in perforated pertic ulcer patients in a tertiary health care center in Bangalore. Int Surg J. 2022;9(3):644. doi: 10.18203/2349-2902.isj20220636.
34. Christanto AGR, Adrianto AA. Comparative accuracy of Boey score and PULP score for predicting prognosis in patients with perforated peptic ulcer: A systematic review and meta-analysis. Journal of Visceral Surgery. 2026;163(3):164-173. doi: 10.1016/j.jviscsurg.2026.02.010.
35. Lee JW, Kim EY. ASA, Boey, and PULP scores in perforated peptic ulcer: which best predicts clinical outcomes?. Foregut Surg. 2026;6(2):62. doi: 10.51666/fs.2026.6.e8.
36. K. R. BP, Patil S, Mohan M. Comparative study between POMPP score versus Boey score to predict morbidity and mortality in peptic perforation peritonitis. Int Surg J. 2021;8(2):543. doi: 10.18203/2349-2902.isj20210054.
37. Gupta S, Zingade A, Baviskar M, et al. Efficacy of the Mannheim Peritonitis Index (MPI) in Predicting Postoperative Outcomes in Patients With Perforation Peritonitis. Cureus. 2025;:. doi: 10.7759/cureus.83193.
38. Sangwan P, Agrawal H, Agarwal N, et al. Comparative Evaluation of the Mannheim Peritonitis Index and POSSUM Scoring Systems for Predicting Mortality and Morbidity in Perforation Peritonitis: A Prospective Observational Study. Indian J Surg. 2026;:. doi: 10.1007/s12262-026-04630-x.
39. Jameel H, Raza MH, Anees A, et al. Comparative evaluation of Mannheim Peritonitis Index and Mansoura Scoring System for risk stratification in perforation peritonitis: a prospective study from a resource-limited setting of North India. BMC Surg. 2026;:. doi: 10.1186/s12893-026-03918-9.
40. Ma X, Wang H, Liu W, et al. The Value of Mannheim Peritonitis Index Combined With ASA Classification in Predicting Postoperative Mortality of Patients With Digestive Tract Perforation. ANZ Journal of Surgery. 2026;96(4):1005-1012. doi: 10.1111/ans.70570.
41. Hota P, Vats S, Nagda P. The Role of the Mannheim Peritonitis Index in Predicting Mortality and Morbidity in Perforation Peritonitis Patients: A Prospective Cohort Study in a Tertiary Care Hospital in Southern Rajasthan. SN Compr. Clin. Med.. 2026;8(1):180. doi: 10.1007/s42399-026-02419-3.
42. Treuheit J, Krautz C, Weber GF, et al. Risk Factors for Postoperative Morbidity, Suture Insufficiency, Re-Surgery and Mortality in Patients with Gastroduodenal Perforation. JCM. 2023;12(19):6300. doi: 10.3390/jcm12196300.
43. Wilson I, Rahman S, Pucher P, et al. Laparoscopy in high-risk emergency general surgery reduces intensive care stay, length of stay and mortality. Langenbecks Arch Surg. 2023;408(1):62. doi: 10.1007/s00423-022-02744-w.
44. Yalcin M. Early post-operative morbidity and mortality predictors in peptic ulcer perforation. Ulus Travma Acil Cerrahi Derg. 2022;:. doi: 10.14744/tjtes.2022.85686.
45. Anwar S, Anees A, Afroz N, et al. H. pylori and peptic ulcer perforation: prevalence of infection and role in surgical outcome. Int Surg J. 2021;8(4):1243. doi: 10.18203/2349-2902.isj20211305.
46. Canlıkarakaya F, Ocaklı S, Doğan İ, et al. The role of biomarkers in predicting mortality in peptic ulcer perforation. BMC Surg. 2026;26(1):154. doi: 10.1186/s12893-026-03537-4.
47. Jia G, Li K. Prognostic Nutritional Index is A Useful Predictive Marker for Morbidity and Mortality in Surgery for Perforated Peptic Ulcer. JIR. 2025;18:11021-11028. doi: 10.2147/jir.s519037.
48. Wang YH, Tee YS, Wu YT, et al. Sarcopenia provides extra value outside the PULP score for predicting mortality in older patients with perforated peptic ulcers. BMC Geriatr. 2023;23(1):269. doi: 10.1186/s12877-023-03946-7.