BMJ Open

Comparison of emergency department trauma triage performance of clinicians and clinical prediction models: a cohort study in India

Ludvig Wärnberg Gerdin,[1] Monty Khajanchi,[2] Vineet Kumar,[3] Nobhojit Roy,[4,5] Makhan Lal Saha,[6] Kapil Dev Soni,[7] Anurag Mishra,[8] Jyoti Kamble,[9] Nitin Borle,[10] Chandrika Prasad Verma,[9] Martin Gerdin Wärnberg [5]

To cite: Wärnberg Gerdin L, Khajanchi M, Kumar V, *et al*. Comparison of emergency department trauma triage performance of clinicians and clinical prediction models: a cohort study in India. *BMJ Open* 2020;**10**:e032900. doi:10.1136/bmjopen-2019-032900

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-032900).

Received 11 July 2019
Revised 23 January 2020
Accepted 28 January 2020

For numbered affiliations see end of article.

Correspondence to
Dr Martin Gerdin Wärnberg; martin.gerdin@ki.se

ABSTRACT

Objective The aim of this study was to evaluate and compare the abilities of clinicians and clinical prediction models to accurately triage emergency department (ED) trauma patients. We compared the decisions made by clinicians with the Revised Trauma Score (RTS), the Glasgow Coma Scale, Age and Systolic Blood Pressure (GAP) score, the Kampala Trauma Score (KTS) and the Gerdin *et al* model.

Design Prospective cohort study.

Setting Three hospitals in urban India.

Participants In total, 7697 adult patients who presented to participating hospitals with a history of trauma were approached for enrolment. The final study sample included 5155 patients. The majority (4023, 78.0%) were male.

Main outcome measure The patient outcome was mortality within 30 days of arrival at the participating hospital. A grid search was used to identify model cut-off values. Clinicians and categorised models were evaluated and compared using the area under the receiver operating characteristics curve (AUROCC) and net reclassification improvement in non-survivors (NRI+) and survivors (NRI–) separately.

Results The differences in AUROCC between each categorised model and the clinicians were 0.016 (95% CI −0.014 to 0.045) for RTS, 0.019 (95% CI −0.007 to 0.058) for GAP, 0.054 (95% CI 0.033 to 0.077) for KTS and −0.007 (95% CI −0.035 to 0.03) for Gerdin *et al*. The NRI+ for each model were −0.235 (−0.37 to −0.116), 0.17 (−0.042 to 0.405), 0.55 (0.47 to 0.65) and 0.22 (0.11 to 0.717), respectively. The NRI– were 0.385 (0.348 to 0.4), −0.059 (−0.476 to −0.005), −0.162 (−0.18 to −0.146) and 0.039 (−0.229 to 0.06), respectively.

Conclusion The findings of this study suggest that there are no substantial differences in discrimination and net reclassification improvement between clinicians and all four clinical prediction models when using 30-day mortality as the outcome of ED trauma triage in adult patients.

Trial registration number ClinicalTrials.gov Registry (NCT02838459).

Strengths and limitations of this study

- ► To the best of our knowledge, this is the first study to prospectively compare the performance of clinicians' trauma triage decisions with the performance of clinical prediction models.
- ► The patient case mix resulting from the multicentre design is likely to improve the external validity of our findings.
- ► We did not assess model performance measures other than discrimination and net reclassification improvement.
- ► We used all cause 30-day mortality as the patient outcome. Future research should explore more short-term outcomes.
- ► Neither patients nor the public were involved in the study design, which may have resulted in a loss of transparency as well as a loss of valuable perspectives to the study design.

INTRODUCTION

Trauma, defined as an external injury and the body's response to that injury, is a major health threat worldwide. In the last 10 years, almost 50 million people died from trauma, and currently, approximately 4.5 million people die from trauma each year.[1] This situation calls for more research on effective trauma care.

Trauma care is highly time sensitive, and the early identification of potentially fatal injuries and conditions is crucial for survival.[2 3] Therefore, triage is a key component of trauma care and is here defined as the process of assigning patients to different levels of urgency for treatment and investigations.

One key challenge in trauma care, especially in many low-resource settings in which prehospital care may be limited or non-existent, is how to triage patients when they arrive at the emergency department (ED).

In settings without formalised triage systems, the level of urgency is often decided based on the clinical discretion of the clinician on duty.

There is an abundance of clinical prediction models in the literature aimed at facilitating the triage of trauma patients.[4] Such models could potentially aid ED trauma triage and allow clinicians to focus on treating the most severe patients first. No study has thus far prospectively compared the performance of clinicians' triage decisions with the performance of clinical prediction models.

Therefore, the aim of this study was to evaluate and compare the abilities of clinicians and clinical prediction models to accurately triage ED trauma patients.

METHODS

Study design

We conducted a prospective cohort study as part of the Trauma Triage Study in India (TTRIS). The TTRIS is a project of the Towards Improved Trauma Care Outcomes consortium.

Setting

The data analysed for this study came from patients enrolled between 28 July 2016 and 05 May 2018 at the following three hospitals: Khershedji Behramji Bhabha Hospital (KBBH) in Mumbai, Lok Nayak Hospital of Maulana Azad Medical College (MAMC) in Delhi and the Institute of Post-Graduate Medical Education and Research and Seth Sukhlal Karnani Memorial Hospital (SSKM) in Kolkata. The time frame was chosen to ensure that all included patients had completed 6 months of follow-up to minimise the loss to follow-up.

KBBH is a community secondary-level teaching hospital with 436 inpatient beds. There are surgery, orthopaedics and anaesthesia departments and both adult and paediatric intensive care units. It has a general ED where all patients are seen. Most patients present directly and are not transferred from another health centre. Plain X-rays and ultrasonography are available around the clock, but CT is only available in-house during the day. During evenings and nights, patients in need of a CT are referred elsewhere.

MAMC and SSKM are both university and tertiary referral hospitals. This means that all specialities and imaging facilities relevant to trauma care are available in-house around the clock. MAMC has approximately 2200 inpatient beds, and SSKM has approximately 1775 inpatient beds. Both MAMC and SSKM have general EDs. Because both MAMC and SSKM are tertiary referral hospitals, a large proportion of patients arriving at these EDs have been transferred from other health facilities, with almost no transfer protocols in place.

Prehospital care is rudimentary in all three cities, with no organised emergency medical services. Ambulances are predominately used for interhospital transfers, and most patients who arrive directly from the scene of the incident are brought by the police or in private vehicles.

At all centres, patients arriving at the ED are first seen by a casualty medical officer largely on a first come, first served basis. There is no formalised system for prioritising ED patients at KBBH or SSKM. In MAMC, there are different coloured zones but no formalised system for how to assign patients to the different zones.

Participants

Eligibility criteria

We included any person aged ≥18 years presenting to the EDs of the participating sites with a history of trauma. A history of trauma was defined here as having any of the external causes of morbidity and mortality listed in block V01-Y36, chapter XX of the International Classification of Disease version 10 online code book as the primary complaint, with some exclusions (see online supplementary material).

Source and methods of selection of participants and follow-up

The project officers worked morning, evening and night shifts, and data were collected from the first 10 consecutive patients during each shift. The rationale for including only the first 10 patients was that this was the number of participants that we considered feasible to follow-up. Follow-up was performed by the project officer at 24 hours, 30 days and 6 months after a participant arrived at a participating hospital. The follow-up was completed in person or by phone, depending on whether the patient was still hospitalised or if the patient had been discharged. The phone numbers of one or more contact persons, for example, relatives, were collected on enrolment and those people were contacted if the participant did not reply to the follow-up phone call. Only if neither the participant nor the contact person answered any of three phone calls was the outcome recorded as missing.

Variables

Outcome

The outcome variable was mortality within 30 days, henceforth referred to as 30-day mortality.

Clinician triage

For the purpose of this study, the clinicians who first assessed the patients on arrival to the ED were instructed by the project officers to categorise patients into four colour-coded triage level groups, henceforth referred to as only triage levels. The triage levels were green, yellow, orange and red. The risk of mortality was assumed to increase from green to red along the corresponding colour spectrum, with green and red indicating the least and most urgent patients, respectively. The clinicians were allowed to use all information available at the time when they assigned the triage level, which was as soon as they had first seen the patient. The triage levels were not used to guide further patient care, and no interventions were implemented as part of the study for patients assigned to the more urgent triage levels.

Table 1 Model predictors with cut-off values where relevant

Predictor	RTS	GAP	KTS	Gerdin *et al*
Age in years	–*	<60	<5, 5–55, >55	–
AVPU	–	–	1–4	–
GCS	3, 4–5, 6–8, 9–12, 13–15	3–15	–	3–15
HR	–	–	–	0–300
NSI	–	–	No, single, multiple	–
RR	0, 1–5, 6–9, 10–29,>29	–	<9, 10–29, >29	–
SBP	0, 1–49, 50–75, 76–89,>89	<60, 60–120, >120	0, 1–49, 50–75, 76–89, >89	0–300

*Indicates that a given predictor is not included in the model.
AVPU, alert, voice, pain or unresponsive; GAP, Glasgow Coma Scale, Age and Systolic Blood Pressure; GCS, Glasgow Coma Scale; HR, heart rate; KTS, Kampala Trauma Score; NSI, number of serious injuries; RR, respiratory rate; RTS, Revised Trauma Score; SBP, systolic blood pressure.

Predictors

The four prediction models we compared with the clinicians' triage decisions were the Revised Trauma Score (RTS),[5] the Glasgow Coma Scale, Age and Systolic Blood Pressure (GAP) score,[2] the Kampala Trauma Score (KTS)[6] and a prediction model previously published by us, here referred to as the Gerdin *et al* model.[7] The rationale for studying these specific models was that RTS is commonly considered the gold standard of physiological trauma severity scoring,[8] KTS has been shown to accurately predict inhospital mortality in both low-income, middle-income and high-income settings,[9 10] and both GAP and the Gerdin *et al* model have been shown to predict short-term trauma mortality.[8 11] The models considered age, systolic blood pressure, heart rate, Glasgow Coma Scale score, alert, voice, pain or unresponsive, respiratory rate and number of serious injuries (table 1). All vital signs were recorded by the project officers who were trained by the project managers and regularly overseen by local supervisors. The number of serious injuries variable was collected by the project officers by asking the same clinician who assigned the triage level.

Other variables

Data on sex and mechanism of injury were recorded in addition to the predictors specified above to characterise the study sample.

Quality assurance

There were several layers of quality control. First, data were entered using a dedicated electronic data collection instrument with extensive logical checks and prompts for unlikely but possible values. Second, the collected data were reviewed on a weekly basis and discussed during weekly online conferences with all project officers. Third, on-site quality control sessions were conducted every 3–4 months. During these sessions, a second project officer collected data alongside the project officer who worked at the site. The quality control data were then compared with the standard data.

Analyses and statistical methods

First, the complete cohort was temporally split into two samples so that the earlier observations were in the first sample, and the later observations were in the second sample. We refer henceforth to these samples as the grid search and comparison samples, respectively.

The grid search sample was used to identify what cut-off values to apply when using the clinical prediction models to assign triage levels to patients. This was done using a grid search that optimised the discrimination of the categorised model. This method of identifying cut-off values was not specified in the original registered protocol. We changed the method because a grid search would result in less arbitrary cut-off values than would using percentiles as originally planned. We used the area under the receiver operating characteristic curve (AUROCC) as the discrimination metric. The grid search was run separately for each model. The categorised models are denoted RTS_{CUT}, Gerdin *et al* $_{CUT}$, GAP_{CUT} and KTS_{CUT}.

Then, the performance of the prediction models was analysed by treating the output as continuous. Models with continuous output are denoted RTS_{CON}, Gerdin *et al* $_{CON}$, GAP_{CON} and KTS_{CON}. Here, the purpose was to conduct a $model_{CUT}$ to $model_{CON}$ comparison—henceforth referred to as a model–model comparison—to evaluate the performance loss caused by categorising model output. The metric used for comparison was the difference in the model AUROCCs, and the comparison was conducted with the comparison sample.

Finally, the performances of the categorised models and clinicians were compared, also with the comparison sample. We refer to this comparison as the model–clinician comparison. The following three metrics were compared: AUROCC and Net Reclassification Improvement (NRI) in events, that is, patients who died within 30 days, (NRI+) and non-events (NRI−), respectively.[12] NRI+ equalled the difference in the proportions of events moving upwards and downwards in triage levels. An upward movement was defined as a move from a lower to a higher triage level, for example, from green to yellow. A downward movement

was defined as a movement from a higher to a lower level. NRI– was calculated in the same way as NRI+ but for non-events. NRI+ and NRI– range from −1 to 1, with positive values indicating that the grouping chosen by a model was superior compared with that chosen by clinicians and negative values indicating the reverse.

We conducted all analyses in the R statistical environment.[13] We calculated 95% CIs using empirical bootstrapping.[14] Observations with missing data were excluded; hence, we report a complete case analysis.

Study size

We estimated the sample size of the comparison sample to include a total of 200 events and all non-events enrolled during the same time period. This sample size was calculated based on published simulation studies of the number of events needed to detect a difference in AUROCCs between two models of approximately <0.05, with 80% power and 5% significance level, when the prevalence of the outcome is 10%.[15] To include the first 200 non-events, we identified the date when the 200th non-surviving patient, counting only complete cases, arrived at a participating centre. We then included all patients, both survivors and non-survivors, who arrived before or on this date. Because we are not aware of a straightforward way to calculate the sample size required for the grid search, we included the same number of events in the grid search sample as in the comparison sample.

Patient and public statement

Neither patients nor the public were involved in the study design nor the planning of this study.

RESULTS

In total, 7697 patients were approached during the study period. The study flowchart is shown in figure 1, and the patient characteristics are shown in table 2. Among the included patients, 4755 (92.2%) survived and 400 (7.8%) did not survive. The majority were male (4023, 78.0%), and the main mechanism of injury was transportation accidents (2170, 42.1%).

Table 3 presents the model cut-off points identified with the grid search method. The model–model AUROCC differences were –0.002 (–0.008 to –0.001), –0.007 (–0.017 to –0.015), –0.003 (–0.005 to –0.001) and –0.013 (–0.025 to 0.006), respectively, for the RTS, GAP, KTS,and the Gerdin *et al* model. Both RTS and KTS suffered a loss of performance when their outputs were categorised.

The triage levels assigned by the clinicians and prediction models in the comparison sample are presented in table 4. Compared with the clinicians RTS_{CUT} and Gerdin *et al*$_{CUT}$ triaged a higher percentage of patients as green whereas GAP_{CUT} and KTS_{CUT} triaged a lower percentage of patients as green. The number and percentages of non-survivors in each triage category are presented in table 5. Compared with the clinicians only KTS_{CUT} triaged fewer non-survivors as green, whereas remaining models had

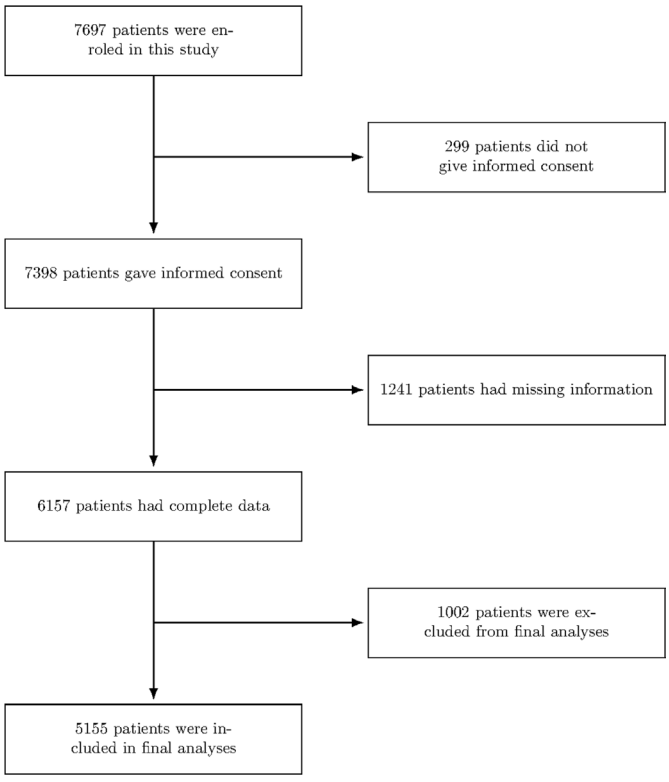

Figure 1 Study flowchart. 1002 patients were excluded from final analysis because they arrived at or after the date when data on the 400th non-surviving patient was collected.

substantially higher percentages of non-survivors in this group. In contrast, all models had higher percentages of non-survivors assigned to the red triage level compared with clinicians.

The AUROCC estimates and corresponding CIs, as well as the CIs for the model–clinician and model–model AUROCC differences, are reported in table 6. RTS_{CUT}, GAP_{CUT}, KTS_{CUT} and Gerdin *et al*$_{CUT}$ generated AUROCCs of 0.907 (0.88 to 0.936), 0.91 (0.884 to 0.951), 0.945 (0.931 to 0.963) and 0.884 (0.856 to 0.92), respectively. RTS_{CON}, GAP_{CON}, KTS_{CON} and Gerdin *et al*$_{CON}$ generated AUROCCs of 0.909 (0.884 to 0.939), 0.918 (0.892 to 0.952), 0.948 (0.933 to 0.966) and 0.897 (0.868 to 0.932), respectively. Clinicians generated an AUROCC of 0.891 (0.872 to 0.907). The ROC curves are shown in figure 2.

Reclassification estimates and the corresponding CIs for the categorised model scores are reported in table 7. RTS_{CUT}, GAP_{CUT}, KTS_{CUT} and Gerdin *et al*$_{CUT}$ generated NRI+ values of −0.235 (−0.37 to −0.116), 0.17 (−0.042 to 0.405), 0.55 (0.47 to 0.65) and 0.22 (0.11 to 0.717) and NRI– values of 0.385 (0.348 to 0.4), −0.059 (−0.476 to −0.005), −0.162 (−0.18 to −0.146) and 0.039 (−0.229 to 0.06) compared with clinicians, respectively.

DISCUSSION

We evaluated and compared the abilities of clinicians and four clinical prediction models to accurately triage ED trauma patients with regards to 30-day mortality. Our

Table 2 Sample characteristics

Characteristic	Level	Grid search	Comparison	Overall
n (%)		1437 (27.9)	3718 (72.1)	5155 (100.0)
Age in years (median (IQR))		33.0 (24.0–48.0)	32.0 (24.0–45.0)	32.0 (24.0–45.0)
Sex (%)	Female	308 (21.4)	824 (22.2)	1132 (22.0)
	Male	1129 (78.6)	2894 (77.8)	4023 (78.0)
Mechanism of injury (%)	Assault	213 (14.8)	564 (15.2)	777 (15.1)
	Burn	5 (0.3)	17 (0.5)	22 (0.4)
	Event of undetermined intent	2 (0.1)	2 (0.1)	4 (0.1)
	Fall	423 (29.4)	999 (26.9)	1422 (27.6)
	Intentional self-harm	4 (0.3)	13 (0.3)	17 (0.3)
	Other external cause of accidental injury	119 (8.3)	624 (16.8)	743 (14.4)
	Transportation accident	671 (46.7)	1499 (40.3)	2170 (42.1)
SBP (median (IQR))		121.0 (110.0–132.0)	123.0 (112.0–134.0)	122.0 (112.0–133.0)
HR (median (IQR))		87.0 (78.0–98.0)	84.0 (76.0–94.0)	84.0 (77.0–96.0)
RR (median (IQR))		21.0 (18.0–22.0)	22.0 (20.0–24.0)	22.0 (20.0–24.0)
AVPU (median (IQR))		4.0 (4.0–4.0)	4.0 (4.0–4.0)	4.0 (4.0–4.0)
All-cause 30-day mortality (%)	No	1237 (86.1)	3518 (94.6)	4755 (92.2)
	Yes	200 (13.9)	200 (5.4)	400 (7.8)
NSI (%)	No serious injury	591 (41.1)	1891 (50.9)	2482 (48.1)
	Single serious injury	713 (49.6)	1628 (43.8)	2341 (45.4)
	Multiple serious injuries	133 (9.3)	199 (5.4)	332 (6.4)
GCS (median (IQR))		15.0 (15.0–15.0)	15.0 (15.0–15.0)	15.0 (15.0–15.0)

AVPU, Alert, Voice, Pain, Unresponsive Scale; GCS, Glasgow Coma Scale; HR, heart rate; NSI, number of serious injuries; RR, respiratory rate in breaths per minute; SBP, systolic blood pressure in mm Hg.

findings indicate that clinicians do not discriminate or classify trauma patients in the ED substantially better than do the studied clinical prediction models. First, the discriminatory ability of clinicians was not superior to that of any of the study models. Second, the NRI+ of clinicians' triage was superior only to that of RTS. In non-events, the NRI of clinicians' triage was superior to that of GAP and KTS.

Table 3 Cut-off points identified with the grid search

	RTS	GAP	KTS	Gerdin *et al*
Green	>7.81	>23	>15	<0.02
Yellow	6.81–7.81	19–23	14–15	0.02–0.03
Orange	5.31–6.81	14–19	13–14	0.03–0.08
Red	<5.31	<14	<13	>0.08

GAP, Glasgow Coma Scale, Age and Systolic Blood Pressure; KTS, Kampala Trauma Score; RTS, Revised Trauma Score.

There was no single model that outperformed clinicians in all performance measures. Only KTS was superior to clinicians in terms of discrimination. In terms of reclassification, KTS and the Gerdin *et al* model reclassified events more accurately than did clinicians. This means that in general these two models assigned higher priority levels to patients who later died than did the clinicians. Only RTS reclassified non-events more accurately than did the clinicians. This means that only RTS assigned lower priority levels patients who survived than did the clinicians.

There were some noticeable differences between the clinicians and the clinical prediction models with regards to assignment of triage levels. Compared with RTS, GAP and Gerdin *et al*, KTS had lower percentages of non-survivors in all triage levels, meaning that in our sample this was the most sensitive model. Compared with clinicians, the other scores were more specific. RTS had the highest percentages of non-survivors in all but the green

Table 4 Priority levels assigned by models and clinicians in the comparison sample (%), n=3718

	Green	Yellow	Orange	Red
RTS_{CUT}	3318 (89.2)	240 (6.5)	102 (2.7)	58 (1.6)
GAP_{CUT}	1693 (45.5)	1713 (46.2)	209 (5.6)	103 (2.8)
KTS_{CUT}	1670 (44.9)	1327 (35.7)	424 (11.4)	297 (8.0)
Gerdin *et al*$_{CUT}$	2263 (60.9)	755 (20.3)	569 (15.3)	131 (3.5)
Clinicians	1967 (52.9)	1354 (36.4)	264 (7.1)	133 (3.6)

GAP, Glasgow Coma Scale, Age and Systolic Blood Pressure; KTS, Kampala Trauma Score; RTS, Revised Trauma Score.

triage level, and was also the model that triaged the largest number of patients as green.

To the best of our knowledge, this is the first study to prospectively compare the ED trauma triage performance of clinicians and clinical prediction models. We were therefore not able to identify any directly comparable studies. The study that most closely matched our research was recently published by Iversen *et al*.[16] They reported that what they referred to as 'eyeball triage', that is, clinicians' triage decisions in our study, was superior to formalised triage using the Danish Emergency Process Triage. They studied a general ED population and not only trauma patients, and the professionals performing the 'eyeball triage' in their study were medical students and phlebotomists.

Few studies have compared predictions of outcomes in trauma patients made by clinicians with those generated by prediction models. In 2015, Mahajan *et al* showed that compared with clinician decisions, a clinical prediction model had better sensitivity but worse specificity for the identification of children with intra-abdominal injuries after blunt torso trauma.[17] Pommerening *et al* have shown that clinicians are not adept at predicting the need for massive transfusion in trauma patients.[18] However, there is evidence that clinicians outperform prediction models in other areas of medicine; for example, Penaloza *et al* showed that clinicians were better than models at predicting pulmonary embolism.[19]

It is sometimes claimed that more experienced clinicians are better at triage than are less experienced clinicians. For example, in the American College of Chest Physicians (CHEST) Consensus Statement on Care of the Critically Ill and Injured During Pandemics and Disasters, it is recommended that a senior clinician perform the trauma triage.[20] We did not assess associations between individual clinician-related factors and the accuracy of triage decisions, but in a study conducted in 2013, Mohan *et al* reported that out of the factors they assessed, only high caseload was associated with the accuracy of triage decisions.[21]

Table 5 Number of non-survivors (%) in each triage category for models and clinicians in the comparison sample

	Green	Yellow	Orange	Red
RTS_{CUT}	30 (0.9)	54 (22.5)	63 (61.8)	53 (91.4)
GAP_{CUT}	12 (0.7)	24 (1.4)	78 (37.3)	86 (83.5)
KTS_{CUT}	1 (0.1)	17 (1.3)	14 (3.3)	168 (56.6)
Gerdin *et al*$_{CUT}$	20 (0.9)	15 (2.0)	66 (11.6)	99 (75.6)
Clinicians	2 (0.1)	62 (4.6)	78 (29.6)	58 (43.6)

GAP, Glasgow Coma Scale, Age and Systolic Blood Pressure; KTS, Kampala Trauma Score; RTS, Revised Trauma Score.

Methodological considerations

We focused on discrimination, measured using the AUROCC and net reclassification of events and non-events separately as performance measures. We chose to maximise AUROCC during the grid search to identify optimal cut-offs because it is one of the most widely used performance measures in prediction research, although we agree that its clinical usefulness can be questioned. Since we are dealing with a class-imbalanced dataset an alternative metric is the F1 score. However, the F1 score has no straightforward interpretation, whereas the AUROCC does. We report the NRI because we believe that its clinical interpretation is more straightforward and useful. In doing so, we acknowledge that the NRI is sensitive to the chosen cut-off values.

We did not include measures of calibration or net benefit.[22] In addition, we do not report specificity and sensitivity. To report the calibration for the prediction models, we would have needed to access the original model specifications, and these were not available for all four models. Furthermore, there is no straightforward method of estimating the calibration of the clinicians' triage decisions. The net benefit, sensitivity and specificity are relevant when a model is used to classify observations into two groups, but when observations are classified into more than two groups the interpretation of these measures per group is less useful.

We used different cut-off values compared with those used in the original studies. The reason was that no original study categorised patients into four risk groups, and we wanted to adjust the categorisation of the model scores to match existing triage systems, for example the South African Triage Scale (SATS),[23] to simplify the potential interpretation in terms of implications for clinical practice.

We handled missing data using listwise deletion. The rationale for using listwise deletion was that the level of missing predictor data was low, and that the study was

Table 6 AUROCCs (95% CI), model–model AUROCC differences and model-clinician AUROCC differences

	AUROCC (95% CI)	Model–model AUROCC difference (95% CI)*	Model-clinician AUROCC difference (95% CI)†
RTS_{CUT}	0.907 (0.880 to 0.936)	−0.002 (−0.008 to −0.001)	0.016 (−0.014 to 0.045)
GAP_{CUT}	0.910 (0.884 to 0.951)	−0.007 (−0.017 to 0.015)	0.019 (−0.007 to 0.058)
KTS_{CUT}	0.945 (0.931 to 0.963)	−0.003 (−0.005 to −0.001)	0.054 (0.033 to 0.077)
Gerdin *et al*$_{CUT}$	0.884 (0.856 to 0.920)	−0.013 (−0.025 to 0.006)	−0.007 (−0.035 to 0.030)
RTS_{CON}	0.909 (0.884 to 0.939)	0.002 (0.001 to 0.008)	0.018 (−0.009 to 0.051)
GAP_{CON}	0.918 (0.892 to 0.952)	0.007 (−0.015 to 0.017)	0.027 (0.000 to 0.061)
KTS_{CON}	0.948 (0.933 to 0.966)	0.003 (0.001 to 0.005)	0.057 (0.037 to 0.080)
Gerdin *et al*$_{CON}$	0.897 (0.868 to 0.932)	0.013 (−0.006 to 0.025)	0.005 (−0.024 to 0.042)
Clinicians	0.891 (0.872 to 0.907)	Not applicable	Not applicable

*The model–model comparison referred to is the AUROCC difference between, for example, RTS_{CUT} and RTS_{CON}.
†A positive difference indicates that the model discriminated better compared with the clinicians.
AUROCC, Area Under the Receiver Operating Characteristics Curve; GAP, Glasgow Coma Scale, Age and Systolic Blood Pressure; KTS, Kampala Trauma Score; RTS, Revised Trauma Score.

powered to accommodate the observed level of loss to follow-up. It is still possible that the amount of missing outcome data was not randomly distributed between survivors and non-survivors and that the mortality figure we report is therefore biased. We cannot know the extent or direction of this bias; however, why it is unclear if more advanced techniques to deal with missing data, such as multiple imputation, would generate less biased estimates. In addition, we argue that our use of all cause 30-day mortality as the outcome is an important improvement over the more commonly used measure of inhospital mortality.

One disadvantage of using 30-day mortality is that there might be patients whose deaths are not caused by the initial trauma and for whom earlier intervention would not have changed the outcome. There are other relevant outcomes that we could have used, such as all-cause 24 hours mortality, intensive care unit admission and major surgery, or we could have used a composite outcome. These outcomes are, however, also not without disadvantages but should be evaluated in future research.

The patient case mix resulting from the multicentre design is likely to limit the applicability of our findings to specific hospitals. To achieve hospital-specific results, larger studies that include only patients from a specific hospital will be needed. However, our case mix is likely to improve the external validity of our findings and their applicability to other heterogeneous patient cohorts. In terms of generalisability, we believe that our results are generalisable to other urban hospitals in India, as well as to urban areas in other low resource settings with similar systems for emergency and trauma care.

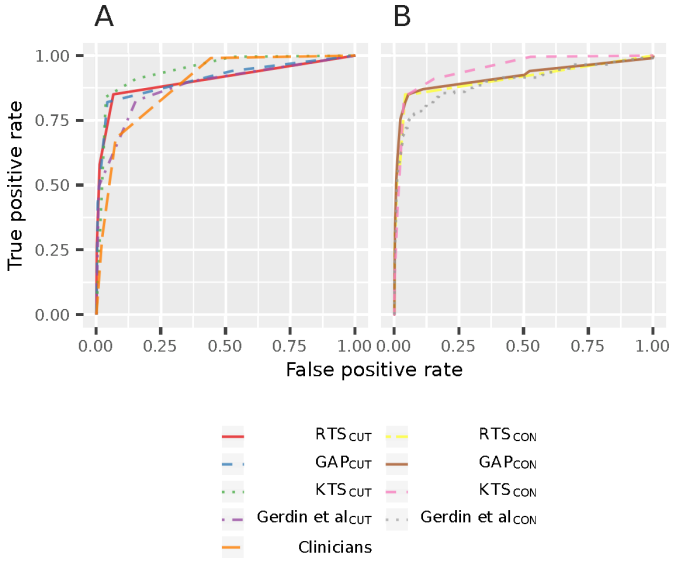

Figure 2 Receiver operating characteristic curves for categorised (A) and continuous models (B) in the comparison sample. GAP, Glasgow Coma Scale, Age and Systolic Blood Pressure; KTS, Kampala Trauma Score; RTS, Revised Trauma Score

Clinical implications and future research

Our study indicates that ED triage of adult trauma patients conducted by clinicians can be comparable to that conducted by clinical prediction models. This can be interpreted as favouring ED trauma triage by clinicians, as the findings are not compatible with an obvious performance benefit of using clinical prediction models. However, in our study the variables needed to assign the triage level using the clinical prediction models, except KTS, were collected by project officers without medical education. Our findings can therefore also be interpreted as favouring the use of clinical prediction models, as these can most likely be used by paramedical professionals and hence potentially reduce the work load of ED clinicians.

Whether this potential can be realised depends on how such models are implemented and integrated into existing systems. It can be well argued that in a single ED there should only be one system in place to triage patients regardless of presenting complaint. The use of

Table 7 NRI+ and NRI– (95% CI)*

	RTS_{CUT}	GAP_{CUT}	KTS_{CUT}	Gerdin *et al* $_{CUT}$
NRI+	−0.235 (−0.370 to −0.116)	0.170 (−0.042 to 0.405)	0.550 (0.470 to 0.650)	0.220 (0.110 to 0.717)
NRI–	0.385 (0.348 to 0.400)	−0.059 (−0.476 to −0.005)	−0.162 (−0.180 to −0.146)	0.039 (−0.229 to 0.060)

Positive values indicate that the grouping according to the model was superior to that of the clinicians, and negative values indicate the reverse.
GAP, Glasgow Coma Scale, Age and Systolic Blood Pressure; KTS, Kampala Trauma Score; NRI, Net Reclassification Improvement; NRI+, NRI in events; NRI–, NRI in non-events; RTS, Revised Trauma Score.

parallel systems for different presenting complaints could be detrimental in low-resource settings, where the time and resources available to teach and use multiple tools is particularly scarce. It is not obvious however that a clinical prediction model for trauma patients cannot be integrated into a more complete triage system, such as SATS.[23]

Such an integration could be worthwhile if it can be shown that clinical prediction models for trauma patient triage outperform how complete systems currently triage these patients. This is one important area for future research. In addition, more research is needed to evaluate if our findings are robust on the individual centre level; for that, a larger study is needed. More research is also needed to evaluate alternative outcomes, such as composite outcomes of death, intensive care unit admission and major surgery. Finally, studies comparing the triage performance of experienced clinicians and clinical prediction models should be conducted.

CONCLUSION

The findings of this study suggest that there are no substantial differences in discrimination and net reclassification improvement between clinicians and four clinical prediction models when using 30-day mortality as the outcome of ED trauma triage in adult patients. Whereas some clinical prediction models classified survivors more appropriately and others were superior in their handling of non-survivors, no model performed substantially better than clinicians in classifying both survivors and non-survivors.

Author affiliations
[1]Department of Industrial Economics and Management, KTH Royal Institute of Technology, Stockholm, Sweden
[2]Department of Surgery, Seth Gowardhandas Sunderdas Medical College and King Edward Memorial Hospital, Mumbai, India
[3]Department of Surgery, Lokmanya Tilak Municipal General Hospital, Mumbai, India
[4]Surgical Unit, WHO Collaborating Centre for Research on Surgical Care Delivery in LMICs, BARC Hospital (Government of India), Mumbai, India
[5]Department of Global Public Health, Karolinska Institutet, Stockholm, Sweden
[6]Department of Surgery, Institute of Post-Graduate Medical Education and Research and Seth Sukhlal Karnani Memorial Hospital, Kolkata, India
[7]Critical and Intensive Care, JPN Apex Trauma Centre, All India Institute of Medical Sciences, New Delhi, India
[8]Department of General Surgery, Maulana Azad Medical College, New Delhi, India
[9]Doctors for You, Mumbai, India
[10]Department of General Surgery, KB Bhabha Municipal General Hospital Mumbai, Mumbai, India

Twitter Nobhojit Roy @#nobsroy

Acknowledgements We would like to thank all participating centres, clinicians and patients. Further, we would like to thank the Thursday Truth Seekers, a multidisciplinary group of researchers based in Mumbai, India, for their advice and input. This paper was proofread and edited for spelling, grammar, clarity and style by editors at Springer Nature Author Services.

Contributors LWG drafted the manuscript, conducted the statistical analysis and interpreted the study results. MGW took the lead in study design and interpreted the results. LWG and MGW developed the code for the analysis. NR, MK, VK, MLS, NB, AM, KDS and JK contributed to the design of the study. JK and CPV conducted data collection. All authors revised the manuscript and approved the final version of the manuscript.

Funding This study was funded by the Swedish National Board of Health and Welfare, grant numbers 22464/2017 and 23745/2016.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The research was approved by the ethics review board at each participating hospital. The names of the boards and the approval numbers are as follows: the Ethics and Scientific Committee (KBBH, HO/4982/KBB), the Institutional Ethics Committee (MAMC, F.1/IEC/MAMC/53/2/2016/No97) and the IPGME&R Research Oversight Committee (SSKM, Inst/IEC/2016/328). The data management and analysis was also approved by the Stockholm Regional Ethics Committee (2018/2201-31/2).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Individual participant data will be available after deidentification with no end date (including data dictionaries). Data will be shared with researchers who provide a methodologically sound proposal to martin.gerdin@ki.se. Statistical code is available at https://github.com/itslwg/prediction_models_vs_clinicians.

ORCID iD
Martin Gerdin Wärnberg http://orcid.org/0000-0001-6069-4794

REFERENCES

1 Roth GA, Abate D, Abate KH, *et al*. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 2018;392:1736–88.
2 Kondo Y, Abe T, Kohshi K, *et al*. Revised trauma scoring system to predict in-hospital mortality in the emergency department: Glasgow coma scale, age, and systolic blood pressure score. *Crit Care* 2011;15:R191.
3 Fitzgerald M*et al*. Trauma resuscitation errors and computer-assisted decision support. *Arch Surg* 2011;146:218–25.
4 Rehn M, Perel P, Blackhall K, *et al*. Prognostic models for the early care of trauma patients: a systematic review. *Scand J Trauma Resusc Emerg Med* 2011;19:17.

5 Champion HR, Sacco WJ, Copes WS, *et al*. A revision of the trauma score. *J Trauma* 1989;29:623–9.
6 Kobusingye OC, Lett RR. Hospital-Based trauma registries in Uganda. *J Trauma* 2000;48:498–502.
7 Gerdin M, Roy N, Dharap S, *et al*. Early hospital mortality among adult trauma patients significantly declined between 1998-2011: three single-centre cohorts from Mumbai, India. *PLoS One* 2014;9:e90064.
8 Laytin AD, Kumar V, Juillard CJ, *et al*. Choice of injury scoring system in low- and middle-income countries: lessons from Mumbai. *Injury* 2015;46:2491–7.
9 Weeks SR, Stevens KA, Haider AH, *et al*. A modified Kampala trauma score (KTS) effectively predicts mortality in trauma patients. *Injury* 2016;47:125–9.
10 Gardner A, Forson PK, Oduro G, *et al*. Diagnostic accuracy of the Kampala trauma score using estimated abbreviated injury scale scores and physician opinion. *Injury* 2017;48:177–83.
11 Gerdin M, Roy N, Khajanchi M, *et al*. Predicting early mortality in adult trauma patients admitted to three public university hospitals in urban India: a prospective multicentre cohort study. *PLoS One* 2014;9:e105606–7.
12 Pencina MJ, D'Agostino RB, D'Agostino RB, *et al*. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–72.
13 R Core Team. *R: A language and environment for statistical computing [Internet*. Vienna, Austria: R Foundation for Statistical Computing, 2017. https://www.r-project.org/
14 Efron B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 1979;7:1–26.
15 Steyerberg EW. *Clinical Prediction Models - A Practical Approach to Development, Validation, and Updating*. New York: Springer-Verlag New York, 2009.
16 Iversen AKS, Kristensen M, Østervig RM, *et al*. A simple clinical assessment is superior to systematic triage in prediction of mortality in the emergency department. *Emerg Med J* 201910.1136/emermed-2016-206382. [Epub ahead of print: 16 Oct 2018].
17 Mahajan P, Kuppermann N, Tunik M, *et al*. Comparison of clinician suspicion versus a clinical prediction rule in identifying children at risk for intra-abdominal injuries after blunt torso trauma. *Acad Emerg Med* 2015;22:1034–41.
18 Pommerening MJ, Goodman MD, Holcomb JB, *et al*. Clinical gestalt and the prediction of massive transfusion after trauma. *Injury* 2015;46:807–13.
19 Penaloza A, Verschuren F, Meyer G, *et al*. Comparison of the unstructured clinician gestalt, the wells score, and the revised Geneva score to estimate pretest probability for suspected pulmonary embolism. *Ann Emerg Med* 2013;62:117–24.
20 Christian MD, Sprung CL, King MA, *et al*. Triage: care of the critically ill and injured during pandemics and disasters: chest consensus statement. *Chest* 2014;146:e61S–74.
21 Mohan D, Barnato AE, Rosengart MR, *et al*. Trauma triage in the emergency departments of nontrauma centers: an analysis of individual physician caseload on triage patterns. *J Trauma Acute Care Surg* 2013;74:1541–7.
22 Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35:1925–31.
23 South African Triage Group. *The South African Triage Scale Training Manual 2012 [Internet*. Westerns Cape Government, 2012: 1–34. https://emssa.org.za/sat/