A SuperLearner Ensemble Machine Learning Algorithm is Non-inferior to Clinicians in Prioritising Among Adult Trauma Patients in the Emergency Department: a Prospective Cohort Study in India

Ludvig Wärnberg Gerdin[1, ◕], Alan Hubbard[2], Anurag Mishra[3], Catherine Juillard[4], Debojit Basak[5], Deepa Kizhakke Veetil[6], Greeshma Abraham[5] Jyoti Kamble[5], Kapil Dev Soni[7], Makhan Lal Saha[8], Monty Khajanchi[9], Nitin Borle[10], Vineet Kumar[11], Sara Moore[2], Martin Gerdin Wärnberg[1*¤, ◕, ‡]

1 Global Health: Health Systems and Policy, Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden
2 Division of Epidemiology and Biostatistics, School of Public Health, University of California, Berkeley, California, USA.
3 Department of Surgery, Maulana Azad Medical College, New Delhi, Delhi, India
4 Department of Surgery, Center for Global Surgical Studies, University of California, San Francisco, California, USA.
5 Tata Institute of Social Sciences and Doctors for You, Mumbai, Maharashtra, India
6 Department of General Surgery, Manipal hospital, Human Care Medical Charitable Trust, New Delhi, Delhi, India
7 Critical & Intensive Care, JPN Apex Trauma Centre, All India Institute of Medical Sciences, New Delhi, Delhi, India
8 Department of Surgery, Institute of Post-Graduate Medical Education and Research and Seth Sukhlal Karnani Memorial Hospital, Kolkata, West Bengal, India
9 Department of Surgery, Seth Gowardhandas Sunderdas Medical College and King Edward Memorial Hospital, Mumbai, Maharashtra, India
10 Department of Surgery, Khershedji Behramji Bhabha hospital, Mumbai, Maharashtra, India
11 Department of Surgery, Lokmanya Tilak Municipal General Hospital, Mumbai, Maharashtra, India

◕These authors contributed equally to this work.
‡Senior authorship.
¤Current Address: Martin Gerdin Wärnberg, Department of Public Health Sciences, Karolinska Institutet, 171 77 Stockholm, Sweden
* martin.gerdin@ki.se

Abstract

Background

Trauma is a major threat to population health globally. Trauma care is highly time sensitive and delays in treatment are associated with preventable mortality across settings. A key component of trauma care is therefore the process of prioritizing patients to match level of care with clinical acuity. In many emergency departments in low resource setting hospitals trauma patients arrive with no or little prenotification. In such settings patients are often prioritised by clinicians based on patients' clinical gestalt. We aimed to compare the performance of an ensemble machine learning methodology called SuperLearner to that of clinician gestalt based on patients' presentation.

Methods and findings

Our hypothesis was that the performance of the SuperLearner would be non-inferior to that of clinician gestalt. We used data from an ongoing prospective cohort study in three public hospitals in urban India. All adult patients presenting to the emergency departments of these hospitals with history of trauma were approached for enrollment. The outcome was all cause mortality within 30 days of arrival to a participating centre. For the purpose of this study, clinicians were instructed to assign patients to one of four levels corresponding to clinical acuity. The levels were green (least urgent), yellow, orange, and red (most urgent). The SuperLearner included five machine learners and was developed in a training sample and then compared to clinicians in a test sample. Performance was compared in terms of reclassification and area under the receiver operating characteristics curve (AUROCC). We concluded that the SuperLearner was non-inferior to clinicians if the lower bound of the 95% confidence intervals (CI) of the net reclassification in events was not less than -0.05. From 28 July 2016 to 21 November 2017 we approached a total of 5667 patients for enrollment. Out of these, 4545 patients consented, had a priority level assigned by a clinician, and had complete outcome data and were therefore included in subsequent analysis. A total of 404 (9%) patients died within 30 days. We used a temporal split to divide the cohort into a training and test sample. The training sample included 3408 patients and the test sample 1137 patients. The AUROCCs of the priority levels assigned by the SuperLearner and clinicians were 0.9574 and 0.8727 respectively. The difference in AUROCC was -0.0846 (95% CI -0.1228 - -0.0451). The net reclassification in events was 0.0114 (95% CI -0.0185 - 0.0299) and in non-events 0.3500 (95% CI 0.2405 - 0.6895).

Conclusions

In terms of reclassification and discrimination an ensemble machine learning algorithm developed using the SuperLearner was non-inferior in prioritising among adult trauma patients in the ED compared to clinician gestalt based on patients' presentation.

Author Summary

Why was this study done?

Trauma kills almost five million people each year. A majority of these deaths occur in low resource settings. New methods are needed to prioritise among trauma patients in the emergency department and quickly identify patients in need of immediate care. Machine learning could potentially help to do so, but so far the use of machine learning in trauma research has been slow. We aimed to compare the performance of an ensemble machine learning methodology called SuperLearner to that of clinician gestalt based on patients' presentation.

What did the researchers do and find?

We analysed data from 4545 adult trauma patients who presented to emergency departments at three public hospitals in urban India. Out of these 404 (9%) patients died from any cause within 30 days of arrival to a participating hospital. We used the SuperLearner to combine multiple machine learning techniques to assign priority levels to included trauma patients based on demographic and clinical patient characteristics. We asked clinicians to also assign priority levels to the same patients and compared the performance of the SuperLearner and clinicians. We found that the SuperLearner was non-inferior to clinicians.

What do these findings mean?

Using an ensemble machine learning algorithm to prioritise among trauma patients in the ED may allow clinicians to focus on treating patients. This would free valuable resources that are particularly scarce the low resource settings where most trauma deaths happen.

Introduction

Trauma is a major threat to population health globally [1, 2]. Every year about 4.6 million people die because of trauma - a number that exceeds the total number of yearly deaths from HIV/AIDS, malaria and tuberculosis combined. The most common cause of trauma is road traffic injuries (RTIs); in 2016 an estimated 1.3 million people died from RTIs alone [2]. Global actors have targeted a 50% reduction of deaths from road trauma by 2020, but this sustainable development goal is far from being realized [3]. This situation calls for not only more interventions, but also strengthened research on effective trauma care delivery.

Trauma care is highly time sensitive and delays to treatment have been associated with increased mortality across settings [4–6]. Early identification and management of potentially life threatening injuries are crucial for survival. A key component of trauma care is therefore the process of prioritizing patients to match level of care with clinical acuity [7, 8]. The existing literature on how to prioritise trauma patients focuses largely on two issues. First, in the prehospital setting the main focus has been to identify patients who merit transfer to a trauma centre [9]. Second, in the hospital setting a substantial body of research has focused on the appropriate criteria for trauma team activation [10, 11].

Although both these issues are important, clinicians all over the world are on a daily basis faced with the more complex problem of how to decide in what order to assess and treat trauma patients that arrive to the emergency department (ED). In health systems with formalised criteria for prioritizing ED patients, all patients are assigned a priority coupled with a target time to treat. These priorities are may be coded with numbers [12] or colors [13], for example red, orange, yellow and green, with red being assigned to the most urgent patients and green to the least urgent.

In health systems without formalized criteria, for example in many low resource settings, clinician gestalt is used informally to prioritize among trauma patients arriving to the ED [14]. As there are commonly no formal prehospital care systems in such settings, trauma patients often arrive to the ED without warning and without any form of previous prioritisation to guide the appropriate level of in-hospital care [15]. Identifying ways to quickly prioritize the patients in need of more immediate care would therefore be very valuable in many low resource settings.

In contrast to trauma centre transfer or trauma team activation, the approach to prioritization among trauma patients arriving to the ED has received little attention from the research community. Framed as a classification problem this challenge can be addressed using a statistical learner. Logistic or proportional hazards models are common classification learners whereas more modern alternatives include random forests or convolutional neural networks. These learners all exist along the machine learning spectrum governed by their relative "human-to-machine decision-making-effort", with regression learners in the more-human-than-machine (MHTM) end and networks at the other, more machine than human (MMTH), end of the spectrum [16].

MMTH learners have been used to solve classification problems in other fields of medicine [17], but the uptake and use of such learners in trauma research has been slow [18]. Some studies have approached the trauma centre transfer and trauma team activation issues using MMTH learners, and the results are conflicting with regards to the superiority of such learners over MHTM learners or standard criteria [19–22]. One very recent study used a random forest learner to assign priority to patients in a general ED population, and found a slight performance improvement using this MMTH learner compared to the standard criteria [23].

Given the paucity of research leveraging machine learning to prioritise among trauma patients in the ED, we aimed to compare the performance of an ensemble machine learning methodology called SuperLearner to that of clinician gestalt based on patients' presentation. Our hypothesis was that the performance of the SuperLearner would be non-inferior to that of clinician gestalt.

Materials and Methods

Study Design

We used data from an ongoing prospective cohort at three public hospitals in urban India. Our analysis is an adjunct to a registered observational study to compare the performance of clinical prediction models with clinicians (ClinicalTrials.gov identifier NCT02838459).

Study Setting

Data anlysed for this study came from patients enrolled between 28 July 2016 and 21 November 2017 at the 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 decided to ensure that all included patients had completed six months follow up. KBBH is a community hospital with 436 inpatient beds. There are departments of surgery, orthopedics, anesthesia, and both adult and pediatric 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 computed tomography (CT) is only available in-house during day-time. 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, except emergency medicine, are available in-house around the clock. MAMC has approximately 2200 inpatient beds and SSKM has around 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 their EDs are 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 predominatly used for inter-hospital transfers and most patients who arrive directly from the scene of the incident are brought by the police or in private vehicles. Patients arriving to the ED are at all centres first seen by a casualty medical officer on a largely first come first served basis. There is no formalised system for prioritising ED patients at any of the centres. The research was approved by the ethical review board at each participating hospital. The names of the boards and the approval numbers were 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).

Data Collection

Data were collected by one dedicated project officer at each site. The project officers all had a masters degree in life sciences. They worked five shifts per week, and each shift was about eight hours long, so that mornings, evenings and nights were covered according to a rotating schedule. In each shift, project officers spent approximately six hours collecting data in the ED and the remaining two following up patients. The collected data were then transferred this data to a digital database. The rationale for this setup was to ensure collection of high-quality data from a representative sample of trauma patients arriving to the EDs at participating centres, while keeping to the projects budget constraints.

Participants

Eligibility criteria

Any person aged $\geq$ 18 years or older and who presented alive to the emergency department (ED) of participating sites with history of trauma was included. The age cutoff was chosen to align with Indian laws on research ethics and informed consent. We defined history of trauma 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 (ICD-10) codebook as primary complaint. Drownings, inhalation and ingestion of objects causing obstruction of respiratory tract, contact with venomous snakes and lizards, accidental poisoning by and exposure to drugs, and overexertion were excluded because they are not considered trauma at the participating centres.

Source and methods of selection of participants and follow up

The project officers enrolled the first ten consecutive patients who presented to the ED during each shift. The number of patients to enroll was set to ten to make follow up feasible. Written informed consent from the patient or a patient representative was obtained either in the ED or in the ward if the patient was admitted. A follow-up was completed by the project officer 30 days and 6 months after participant arrived at participating hospital. The follow-up was completed in person or on phone, depending on whether the patient was still hospitalised or if the patient had been discharged. Phone numbers of one or more contact persons (e.g. relatives), were collected on enrollment and contacted if the participant did not reply on follow up. Only if neither the participant nor the contact person answered any of three repeated phone calls was the outcome recorded as missing and the patient was considered lost to follow up.

Variables, Data Sources and Measurement

Patient characteristics and SuperLearner variables

The dependent variable, or label, used to train the SuperLearner was all-cause 30 day mortality, defined as death from any cause within 30 days of arrival to a participating centre. These data were extracted from patient records if the patient was still in hospital 30 days after arrival, or collected by calling the patient or the patient representative if the patient was not in hospital.

The independent variables, or features, included patient age in years, sex, mechanism of injury, type of injury, mode of transport, transfer status, time from injury to arrival in hours. The project officers collected data on these features by asking the patient, a patient representative, or by extracting the data from the patient's file. Sex was coded as male or female. Mechanism of injury was coded by the project officers using ICD-10 after completing the World Health Organization's (WHO) electronic ICD-10-training tool [24]. The levels of mechanism of injury was collapsed for analysis into transport accident (codes V00-V99), falls (W00-W19), burns (X00-X19), intentional self harm (X60-X84), assault (X85-X99 and Y00-Y09), and other mechanism (W20-99, X20-59 and Y10-36). Type of injury was coded as blunt, penetrating, or both blunt and penetrating. Mode of transport was coded as ambulance, police, private vehicle, or arrived walking. Transfer status was a binary feature indicating if the patient was transferred from another health facility or not.

The features also included vital signs measured on arrival to the ED at participating centres. The project officers recorded all vital signs using hand held equipment, i.e. these were not extracted from patient records, after receiving two days of training and yearly refreshers. Only if the hand held equipment failed to record a value did the project officers extract data from other attached monitoring equipment, if available. Systolic and diastolic blood pressure (SBP and DBP) were measured using an automatic blood pressure monitor. Heart rate (HR) and peripheral capillary oxygen saturation (SpO_2) were measured using a portable non-invasive fingertip pulse oximeter. Respiratory rate (RR) was measured manually by counting the number of breaths during one minute. Level of consciousness was measured using both the Glasgow coma scale (GCS) and the Alert, Voice, Pain, and Unresponsive scale (AVPU). In assigning GCS the project officers used the official Glasgow Coma Scale Assessment Aid [25]. AVPU simply indicates whether the patient is alert, responds to voice stimuli, painful stimuli, or does not respond at all. These represent standard variables commonly collected in many health systems. They are also included in several well known clinical prediction models designed to predict trauma mortality [26].

Clinicians' priority levels

For the purpose of this study, clinicians were instructed by the project officers to assign a priority to each patient. The priority levels were color coded. Red was assigned to the most serious patients that should be treated first. Green was assigned to the least serious patients that should be treated last. Orange and yellow were intermediate levels, where orange patients were less serious than red but more serious than yellow and green whereas yellow patients were less serious than red and orange patients but more serious than green patients. The clinicians were allowed to use all information available at the time when they assigned the priority level, which was as soon as they had first seen the patient. The priorities 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 priority levels.

Bias

Project officers underwent two days of training in study procedures and were then supervised locally. We conducted continuous data quality assurance by having weekly online data review meetings during which data discrepancies were identifed, discussed and resolved. We conducted quarterly on site quality control sessions during which data collection was conducted both by the centre's own project officer and a quality control officer. Data entry errors were prevented by having extensive logical checks in the digital data collection instrument.

Statistical Methods

All data was de-identified before it was analysed for this study. Details of the de-identification procedures are available as supporting information in S1 Text. We used R for all analyses [27]. We first made a non-random temporal split of the complete data set into a training and test set. The split was made so that 75% of the complete cohort was assigned to the training set and the remaining 25% to the test set, ensuring that the relative contribution of each centre was maintained in both sets. We then calculated descriptive statistics of all variables, using medians and interquartile ranges (IQR) for continuous variables and counts and percentages for qualitative variables. All quantitative features (age, SBP, DBP, HR, SpO_2, and RR) were treated as continuous and the levels of all qualitative variables (sex, mechanism of injury, type of injury, mode of transport, transfer status, and GCS components) were treated as bins (dummy variables).

Development of the SuperLearner

We then developed our SuperLearner in the training set using the SuperLearner R package [28]. SuperLearner is an ensemble machine learning algorithm, meaning that it uses a library of techniques or specific learners, in principle any technique or learner that the analyst wants, to come up with an “optimal learner”. Table 1 show our library of techniques that included three MHTM and two MMTH learners. All were implemented using the default hyperparameters. Short descriptions of the individual learners are available as supporting information in S2 Text. The SuperLearner was trained using ten fold cross validation. This procedure is implemented by default in the SuperLearner package and entails splitting the development data in ten mutually exclusive parts of approximately the same size. All learners included in the library are then fitted using the combined data of nine of these parts and evaluated in the tenth. This procedure is then repeated ten times, i.e. each part is used once as the evaluation data, and is intended to limit overfitting and reduce optimism.

Table 1. Techniques included in our SuperLearner library

Learner	R package	SuperLearner function
Breiman’s random forest algorithm	randomForest [29]	SL.randomForest
Extreme Gradient Boosting machine	XGboost [30]	SL.xgboost
Generalized Linear Model	glm (built-in)	SL.glm
Generalized Additive Model	gam [31]	SL.gam
Penalized regression model using elastic net	glmnet [32]	SL.glmnet

The SuperLearner was then used to assign levels of priority to the patients in the training set. This was done by binning the SuperLearner prediction into four bins using cutoffs identified using a grid search to optimize the area under the receiver operation characteristics curve (AUROCC) across all possible combinations of unique cutoffs, where each cutoff could take any value from 0.01 to 0.99 in 0.01 unit increments. These bins corresponded to the green, yellow, orange, and red priority levels assigned by the clinicians. The performance of both the continuous SuperLearner prediction and the SuperLearner priority levels in the training set was then evaluated using AUROCC. We then used the SuperLearner to predict the outcomes of the patients in the test set and used the cutoff values from the training set to assign a level of priority to each patient in this set.

Comparing the SuperLearner and Clinicians

The perfomance of the continuous SuperLearner prediction, the SuperLearner priority levels, and the clinicians’ priority levels, was then evaluated by estimating and comparing their AUROCC. The levels of priority assigned by the SuperLearner and clinicians respectively were then compared by estimating the net reclassification, in events (patient with the outcome, i.e. who died within 30-days from arrival) and non-events (patient without the outcome) respectively. The net reclassification in events was defined as the difference between the proportion of events assigned a higher priority by the SuperLearner than the clinicians and the proportion of events assigned a lower priority by the SuperLearner than the clinicians. Conversely, the net reclassification in non-events was defined as the difference between the proportion of non-events assigned to a lower priority by the SuperLearner than the clinicians and the proportion of non-events assigned a higher priority by the SuperLearner than the clinicians. We used an empirical bootstrap with 1000 draws of the same size as the original set to estimate 95% confidence interval (CI) around differences. We concluded that the SuperLearner was non-inferior to clinicians if the 95% CI of the net reclassification in events did not exceed a pre-specified level of -0.05, indicating that clinicians correctly classified 5 in 100 events more than the SuperLearner.

Handling of missing data

Observations with missing data on all cause 30-day mortality or priority level assigned by clinicians were excluded. Missing data in features was treated as informative. For each feature with missing data we created a non-missingness indicator, a variable that took the value of 0 if the feature value was missing and 1 otherwise. Missing feature values were then replaced with the median of observed data for quantitative features and the most common level for qualitative features. We included the non-missingness indicators as features in the SuperLearner.

Results

During the study period, we approached a total of 5724 patients for enrollment. A random sample of 57 observations were removed during data de-identification. Consent was declined by 215 patients. Out of the 5452 patients who provided informed consent, 1 had missing data on priority level assigned by clinicians, leaving 5451 patients. An additional 906 were excluded because of missing outcome data. Thus, the final study sample included 4545 patients.

Table 2 shows the characteristics of our study sample. A total of 46 (1%) patients had missing values in at least one feature. Among the included patients the median age was 32 (IQR 24-45) years. A majority, 3539 (78%) patients, were males. The most common mechanism of injury was transport accidents, accounting for 1925 (42%) patients. A total of 1973 (43%) patients were transported to participating centres in some sort of private vehicle, such as a car, taxi, or rickshaw. A majority of patients had normal vital signs on arrival to participating centres. Out of all patients, 404 (9%) died within 30 days of arrival. The number of patients in the training and test samples were 3408 and 1137 respectively.

The AUROCC of the continuous SuperLearner prediction in the training sample was 0.9829 (Fig. 1A). The cutpoints identified by the grid search were 0.05, 0.08, and 0.61. We used these cutpoints to bin the continuous SuperLearner prediction into the four priority levels green, yellow, orange, and red. The AUROCC of the SuperLearner priority levels in the training sample was 0.9785. Fig. 2A shows the precision-recall curves in the training sample.

We then applied the SuperLearner to the test sample. The AUROCC of the continuous SuperLearner prediction was 0.9828 Figure 1B. We used the same cutpoints as in the training sample to bin the continuous predictions into the four priority levels. The AUROCC of the SuperLearner priority levels in the test sample was 0.9574. Fig. 2B shows the precision-recall curves in the test sample.

Table 2. Characteristics of the samples analysed in this study

Characteristic	Level	Training	Test	Overall	Missing values, n (%)
n (%)		3408 (75.0)	1137 (25.0)	4545 (100.0)	46 (1)*
Age in years (median [IQR])		32.0 [24.0, 46.0]	31.0 [24.0, 45.0]	32.0 [24.0, 45.0]	0 (0)
Sex (%)	Female	763 (22.4)	243 (21.4)	1006 (22.1)	0 (0)
	Male	2645 (77.6)	894 (78.6)	3539 (77.9)	
Mechanism of injury (%)	Assault	515 (15.1)	168 (14.8)	683 (15.0)	0 (0)
	Burn	11 (0.3)	7 (0.6)	18 (0.4)	
	Event of undetermined intent	4 (0.1)	0 (0.0)	4 (0.1)	
	Fall	948 (27.8)	296 (26.0)	1244 (27.4)	
	Intentional self harm	12 (0.4)	4 (0.4)	16 (0.4)	
	Other external cause of accidental injury	489 (14.3)	166 (14.6)	655 (14.4)	
	Transport accident	1429 (41.9)	496 (43.6)	1925 (42.4)	
Type of injury (%)	Blunt	3372 (98.9)	1133 (99.6)	4505 (99.1)	1 (0)
	Penetrating	30 (0.9)	3 (0.3)	33 (0.7)	
	Blunt and penetrating	6 (0.2)	1 (0.1)	7 (0.2)	
Mode of transport (%)	Ambulance	1825 (53.6)	498 (43.8)	2323 (51.1)	4 (0.1)
	Police	91 (2.7)	20 (1.8)	111 (2.4)	
	Private vehicle	1395 (40.9)	578 (50.8)	1973 (43.4)	
	Arrived walking	97 (2.8)	41 (3.6)	138 (3.0)	
Transferred (%)	No	1534 (45.0)	538 (47.3)	2072 (45.6)	0 (0)
	Yes	1874 (55.0)	599 (52.7)	2473 (54.4)	
SBP (median [IQR])		121.0 [111.0, 132.0]	125.0 [112.0, 136.0]	122.0 [111.0, 133.0]	10 (0.2)
DBP (median [IQR])		80.0 [70.0, 87.0]	81.0 [73.0, 91.0]	80.0 [70.0, 89.0]	11 (0.2)
SpO_2 (median [IQR])		98.0 [97.0, 98.0]	98.0 [98.0, 98.0]	98.0 [97.0, 98.0]	4 (0.1)
HR (median [IQR])		86.0 [77.0, 97.0]	83.0 [77.0, 92.0]	85.0 [77.0, 96.0]	4 (0.1)
RR (median [IQR])		22.0 [19.0, 24.0]	22.0 [20.0, 24.0]	22.0 [20.0, 24.0]	3 (0.1)
EGCS (%)	1	178 (5.2)	41 (3.6)	219 (4.8)	0 (0)
	2	74 (2.2)	23 (2.0)	97 (2.1)	
	3	121 (3.6)	28 (2.5)	149 (3.3)	
	4	3007 (88.2)	1043 (91.7)	4050 (89.1)	
	Non testable	28 (0.8)	2 (0.2)	30 (0.7)	
VGCS (%)	1	196 (5.8)	35 (3.1)	231 (5.1)	0 (0)
	2	89 (2.6)	24 (2.1)	113 (2.5)	
	3	40 (1.2)	20 (1.8)	60 (1.3)	
	4	166 (4.9)	78 (6.9)	244 (5.4)	
	5	2911 (85.4)	980 (86.2)	3891 (85.6)	
	Non testable	6 (0.2)	0 (0.0)	6 (0.1)	
MGCS (%)	1	67 (2.0)	10 (0.9)	77 (1.7)	1 (0)
	2	37 (1.1)	10 (0.9)	47 (1.0)	
	3	35 (1.0)	8 (0.7)	43 (0.9)	
	4	39 (1.1)	8 (0.7)	47 (1.0)	
	5	186 (5.5)	62 (5.5)	248 (5.5)	
	6	3040 (89.2)	1039 (91.4)	4079 (89.7)	
	Non testable	4 (0.1)	0 (0.0)	4 (0.1)	
AVPU (%)	Unresponsive	68 (2.0)	9 (0.8)	77 (1.7)	1 (0)
	Pain responsive	212 (6.2)	78 (6.9)	290 (6.4)	
	Voice responsive	118 (3.5)	27 (2.4)	145 (3.2)	
	Alert	3010 (88.3)	1023 (90.0)	4033 (88.7)	
Delay (median [IQR])		329.5 [65.0, 1381.2]	480.0 [65.0, 1705.0]	360.0 [65.0, 1500.0]	30 (0.7)
All cause 30-day mortality (%)	No	3093 (90.8)	1048 (92.2)	4141 (91.1)	0 (0)
	Yes	315 (9.2)	89 (7.8)	404 (8.9)	

*The total number (%) of observations with missing data. Abbreviations and explanations: AVPU, Alert, voice, pain, unresponsive scale; DBP, Diastolic blood pressure in mmHg; Delay, Time between injury and arrival to participating centre in minutes; EGCS, Eye component of the Glasgow Coma Scale; HR, Heart rate; MGCS, Motor component of the Glasgow Coma Scale; RR, Respiratory rate in breaths per minute; SBP, Systolic blood pressure in mmHg; SpO_2, Peripheral capillary oxygen saturation; Transferred, Transferred from another health facility; VGCS, Verbal component of the Glasgow Coma Scale

In the test sample we compared the performance of the binned SuperLearner prediction with that of clinicians. The AUROCC of priority levels assigned by clinicians was 0.8727. The difference in AUROCC between the SuperLearner priority levels and clinicians was -0.0846 (95% CI -0.1228 - -0.0451). The net reclassification in events and non-events were 0.0114 (95% CI -0.0185 - 0.0299) and 0.3500 (95% CI 0.2405 - 0.6895) respectively. The overall reclassification is shown in Table 3.

Fig 1. Receiver operating characteristics curves in training (A) and test (B) samples

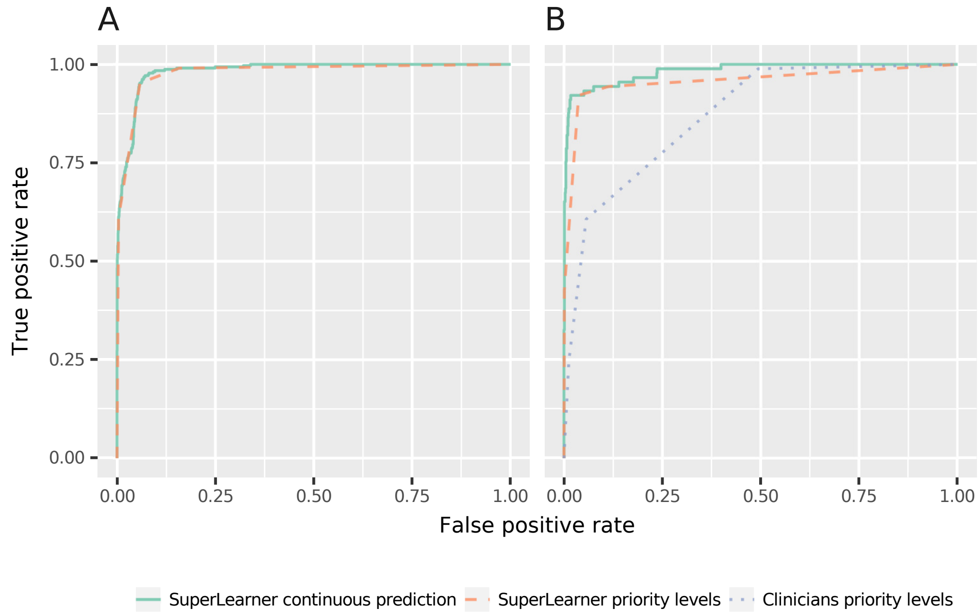


Fig 2. Precision-recall curves in training (A) and test (B) samples

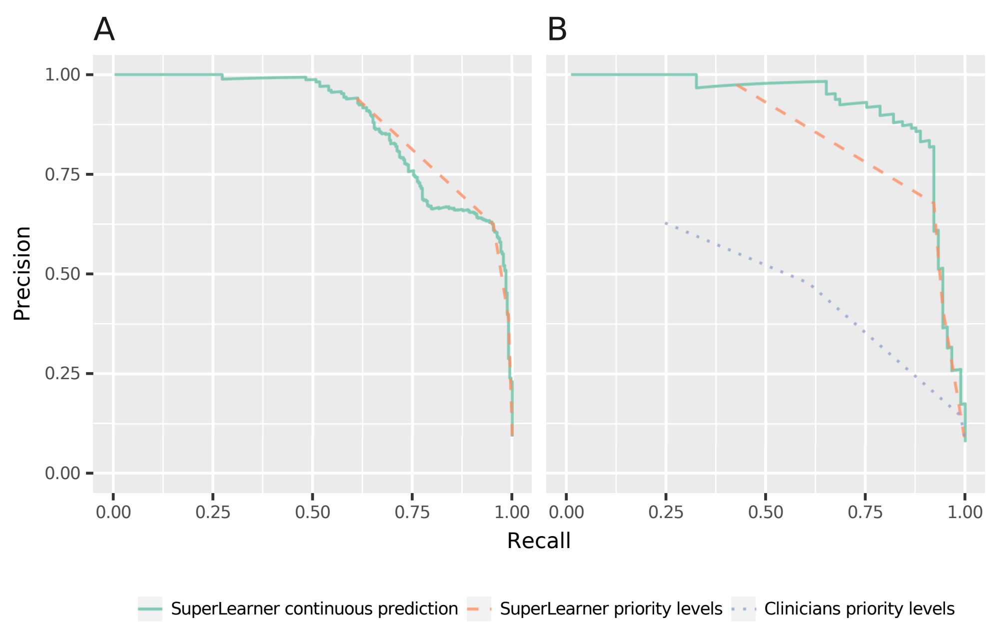


Fig. 3 shows the number of patients assigned to each priority level by the SuperLearner and clinicians. The number of patients assigned to the four priority levels differed substantially between the two. This difference was particularly marked in the green and yellow priority levels. The SuperLearner assigned the green priority level to 934 patients whereas clinicians assigned this level to 532. Among the patients that the SuperLearner prioritised as green 5 died. The corresponding figure for the clinicans was

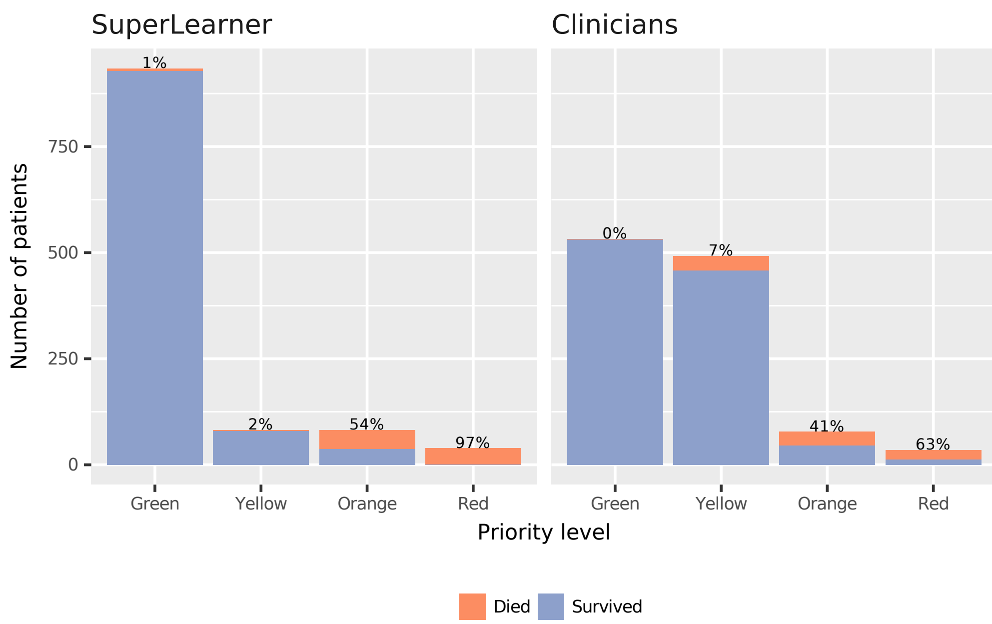
Table 3. Priority levels assigned by SuperLearner and clinicians in complete test sample (n = 1137)

	SuperLearner						
Clinicians	Green	Yellow	Orange	Red	Rec. %	Rec. up %	Rec. down %
Green	522	7	3	0	2	2	
Yellow	369	68	46	9	86	11	75
Orange	30	7	31	10	60	13	47
Red	13	0	2	20	43		43

Reclassification (Rec.) figures refer to % of patients reclassified by the SuperLearner compared to clinicians. Rec. up and Rec. down indicates % of patients reclassified to a higher or lower priority level respectively.

1. In contrast, the SuperLearner assigned the yellow priority level to 82 patients, out of which 2 died. Corresponding figures for the clinicians were 492 and 34.

Fig 3. Number of patients assigned to each priority level by the SuperLearner and clinicians in the test sample. Percentages are % with all cause mortality at each level.



Discussion

Our results indicate that using an ensemble machine learner developed with the SuperLearner to prioritise among adult trauma patients in the ED is non-inferior to that of clinician gestalt. In fact, our results suggest that the ensemble learner is superior to clinician gestalt in terms of discrimination. We have not been able to identify any previous study that has applied machine learning to prioritise among trauma patients in the ED. Hence, as far as we know this is the first study of its kind in this area and we hope that our results can work as benchmarks to which future work can be compared.

We found that the ensemble learner in general reclassified events to a higher priority level and non-events to a lower priority level, compared to clinicians. Specifically, the

SuperLearner reclassified a majority of patients from the yellow priority to the green priority level. This is analogous to reduced overtriage. Overtriage and undertriage are concepts used extensively in the trauma literature. Undertriage refers to for example patients with major trauma not being transferred to a trauma centre and overtriage to patients with minor trauma being transferred to a trauma centre. Our findings would indicate that most of the patients assigned to the yellow priority level by clinicians were overtriaged and strain the health system in face of limited resources.

Three studies have used MMTH learners to limit under and overtriage of trauma patients. Talbert et al. applied a tree based learner but found no improvement over standard criteria [19]. More recent research by Follin et al. demonstrated superior performance of the tree based learner compared to a model based on logistic regression [22]. Pearl et al. used neural networks but could not demonstrate a difference [20]. Only Follin report performance measures that can be compared to our results. Their learner achieved an AUROCC of 0.82, which is substantially lower than that of our ensemble learner.

In contrast, the literature is replete with studies using MHTM learners to reduce under and overtriage, or predict trauma mortality [10, 26, 33]. The performance of these learners vary substantially, but many studies report AUROCCs that approaches that of our ensemble learner. For example, Miller et al. and Kunitake et al. achieved AUROCCs of almost 0.97 and 0.94 with their models based on logistic regression [34]. Neither of these studies however approached the problem of prioritising among trauma patients in the ED, or suggested how the models could be used to assign patients to different priority levels.

Our study was limited by the relatively small sample size. For example, we did not have enough data to run centre wise analysis, which should be a focus of future studies. Instead we concentrated on data quality and had dedicated project officers record all data. This resulted in very low levels of missing feature data. In contrast, we did have a considerable amount of missing outcome data, with about 20% of patients being lost to follow up. We handled this missingness using list wise deletion, aware of the potential bias introduced by this approach. One alternative would have been to use multiple imputation to replace missing values, however we had no way of determining the mechanism underlying the missing outcomes why results based on multiple imputed data might be biased as well. Further, we did not consider it computionally feasible to combine multiple imputation and bootstraping for uncertainty estimation. We do however consider it a strength of our study that the outcome included out of hospital deaths, when comparably recent research does not [23, 35].

We used point measurements to train the ensemble learner, meaning that we failed to account for potential changes in patients' clinical condition between the time when feature and outcome data were collected. The clinicians were however also limited to the data available when they decided on a priority level, although this could have included laboratory or imaging findings from a transferring health facility. Future research may improve the predictions by both the ensemble machine learner and clinicians by including data from multiple time points.

As opposed to the clinicians the ensemble learner was limited by the features that we defined. For example, in our setting with no or very limited electronic record keeping it would have been challenging to incorporate for example imaging data. In settings with more extensive electronic records this should be more feasible. Further, the ensemble learner was limited by the techniques included in its library. We included a mix of MHTM and MMTH learners, for example logistic regression and random forest. The performance of our ensemble learner was already very good, but extending the list of features and techniques available to the learner would likely improve it further. Also, we used the default hyperparameter settings for each technique. Future research may improve the learner's performance by modifying the included learners' hyperparameters.

Several steps remain before a system to prioritse among adult trauma patients in the ED based on our algorithm can and should be implemented. These steps involve refining the algorithm, comparing it with other commonly used methods to prioritise patients in the ED, incorporating it into usable software that may be used in parallel even in settings with no electronic health records, and designing an implementation study to assess both its effectiveness and safety. There are many ways in which the algorithm could be refined but we regard defining a sequence in which the variables should be measured as the most important. We think that this sequence should be based on a combination of individual variable importance and how feasible the variables are to record. We assume that once this sequence is defined the patients with the most severe trauma could be identified very quickly using only a small subset of the variables. Further, our ensemble learner did assign more events to the green priority level than the clinicians. This should be explored in depth in future studies.

Conclusion

An ensemble machine learner developed with the SuperLearner to prioritise among adult trauma patients in the ED is non-inferior to that of clinician gestalt. The SuperLearner may be especially useful to reduce the number of patients that would be unnecessarily prioritised to a high priority level.

Supporting Information

S1 Text Details of de-identification procedures.

S2 Text Short descriptions of included learners.

Acknowledgments

We would like to thank the Towards Improved Trauma Care Outcomes and the Trauma Triage Study in India teams.

References

1. Brohi K, Schreiber M. The new survivors and a new era for trauma research. PLoS Medicine. 2017;14(7):3–5. doi:10.1371/journal.pmed.1002354.

2. GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980 – 2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(September 16):1151–210. doi:10.1016/S0140-6736(17)32152-9.

3. United Nations Division for Sustainable Development. Sustainable development goal 3. Ensure healthy lives and promote well-being for all at all ages; 2018. Available from: `https://sustainabledevelopment.un.org/sdg3`.

4. Yeboah D, Mock C, Karikari P, Agyei-Baffour P, Donkor P, Ebel B. Minimizing preventable trauma deaths in a limited-resource setting: A test-case of a multidisciplinary panel review approach at the Komfo Anokye Teaching Hospital in Ghana. World Journal of Surgery. 2014;38(7):1707–1712. doi:10.1007/s00268-014-2452-z.

5. O'Reilly D, Mahendran K, West A, Shirley P, Walsh M, Tai N. Opportunities for improvement in the management of patients who die from haemorrhage after trauma. British Journal of Surgery. 2013;100:749–755. doi:10.1002/bjs.9096.

6. Roy N, Veetil DK, Khajanchi MU, Kumar V, Solomon H, Kamble J, et al. Learning from 2523 trauma deaths in India- opportunities to prevent in-hospital deaths. BMC Health Services Research. 2017;17(142):1–8. doi:10.1186/s12913-017-2085-7.

7. Eastern Association for the Surgery of Trauma (EAST). Practice Management Guidelines for the Appropriate Triage of the Victim of Trauma. EAST; 2010.

8. National Institute for Health and Care Excellence (NICE). Major trauma: service delivery. NICE; 2016. February.

9. Voskens FJ, van Rein EAJ, van der Sluijs R, Houwert RM, Lichtveld RA, Verleisdonk EJ, et al. Accuracy of Prehospital Triage in Selecting Severely Injured Trauma Patients. JAMA Surgery. 2018;153(4):322–327. doi:10.1001/jamasurg.2017.4472.

10. van Rein EAJ, van der Sluijs R, Houwert RM, Gunning AC, Lichtveld RA, Leenen LPH, et al. Effectiveness of prehospital trauma triage systems in selecting severely injured patients: Is comparative analysis possible? American Journal of Emergency Medicine. 2018;doi:10.1016/j.ajem.2018.01.055.

11. Tignanelli CJ, Vander Kolk WE, Mikhail JN, Delano MJ, Hemmila MR. Noncompliance with American College of Surgeons Committee on Trauma recommended criteria for full trauma team activation is associated with undertriage deaths. Journal of Trauma and Acute Care Surgery. 2018;84(2):287–294. doi:10.1097/TA.0000000000001745.

12. Agency for Healthcare Research and Quality. Emergency Severity Index (ESI). A Triage Tool for Emergency Department Care. U.S. Department of Health & Human Services; 2012. Version 4. Available from: `http://dx.doi.org/10.1016/j.cmpb.2014.08.006`.

13. South African Triage Group. The South African Triage Scale Training Manual 2012. Western Cape Government; 2012. Available from: `https://emssa.org.za/sats/`.

14. Baker T, Lugazia E, Eriksen J, Mwafongo V, Irestedt L, Konrad D. Emergency and critical care services in Tanzania: a survey of ten hospitals. BMC Health Services Research. 2013;13(1):140. doi:10.1186/1472-6963-13-140.

15. Choi SJ, Oh MY, Kim NR, Jung YJ, Ro YS, Shin SD. Comparison of trauma care systems in Asian countries: A systematic literature review. Emergency Medicine Australasia. 2017;29(June):697–711. doi:10.1111/1742-6723.12840.

16. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. Jama. 2018;(March 12):E1–E2. doi:10.1001/jama.2017.18391.

17. Nevin L. Human Intelligence & Artificial Intelligence in Medicine: A day with the Stanford Presence Center; 2018. Available from: `http://blogs.plos.org/speakingofmedicine/2018/04/24/human-intelligence-artificial-intelligence-in-medicine-a-day-with-the-sta`

18. Liu NT, Salinas J. Machine Learning for Predicting Outcomes in Trauma. Shock. 2017;48(5):504–510. doi:10.1097/SHK.0000000000000898.

19. Talbert S, Talbert DA. A comparison of a decision tree induction algorithm with the ACS guidelines for trauma triage. AMIA Annual Symposium proceedings. 2007; p. 1127.

20. Pearl A, Bar-Or R, Bar-Or D. An artificial neural network derived trauma outcome prediction score as an aid to triage for non-clinicians. Studies in Health Technology & Informatics. 2008;136:253–258.

21. Scerbo M, Radhakrishnan H, Cotton B, Dua A, Del Junco D, Wade C, et al. Prehospital triage of trauma patients using the Random Forest computer algorithm. Journal of Surgical Research. 2014;187(2):371–376. doi:10.1016/j.jss.2013.06.037.

22. Follin A, Jacqmin S, Chhor V, Bellenfant F, Robin S, Guinvarc'h A, et al. Tree-based algorithm for prehospital triage of polytrauma patients. Injury. 2016;47(7):1555–1561. doi:10.1016/j.injury.2016.04.024.

23. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. Annals of Emergency Medicine. 2018;71(5):565–574.e2. doi:10.1016/j.annemergmed.2017.08.005.

24. World Health Organization. ICD-10 Interactive Self Learning Tool; 2018. Available from: http://apps.who.int/classifications/apps/icd/icd10training/.

25. glasgowcomascale org. GLASGOW COMA SCALE: Do it this way; 2018. Available from: http://www.glasgowcomascale.org/downloads/GCS-Assessment-Aid-English.pdf?v=3.

26. Rehn M, Perel P, Blackhall K, Lossius HM. Prognostic models for the early care of trauma patients: a systematic review. Scandinavian journal of trauma, resuscitation and emergency medicine. 2011;19(1):17. doi:10.1186/1757-7241-19-17.

27. R Core Team. R: A language and environment for statistical computing; 2017. Available from: https://www.r-project.org/.

28. Polley E, LeDell E, van der Laan M. SuperLearner: Super Learner Prediction; 2016. Available from: https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf$\delimiter"026E30F$nhttps://github.com/ecpolley/SuperLearner.

29. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

30. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al.. xgboost: Extreme Gradient Boosting; 2018. Available from: https://CRAN.R-project.org/package=xgboost.

31. Hastie T. gam: Generalized Additive Models; 2018. Available from: https://CRAN.R-project.org/package=gam.

32. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software. 2011;39(5):1–13.

33. de Munter L, Polinder S, Lansink KWW, Cnossen MC, Steyerberg EW, de Jongh MAC. Mortality prediction models in the general trauma population: A systematic review. Injury. 2017;48(2):221–229. doi:10.1016/j.injury.2016.12.009.

34. Miller RT, Nazir N, McDonald T, Cannon CM, Pearson WS, Dulski T, et al. The modified rapid emergency medicine score: A novel trauma triage tool to predict in-hospital mortality. Injury. 2017;67(0):71–75. doi:10.1016/j.injury.2017.04.048.

35. Kunitake RC, Kornblith LZ, Cohen MJ, Callcut RA. Trauma Early Mortality Prediction Tool (TEMPT) for assessing 28-day mortality. Trauma Surg Acute Care Open. 2018;3:1–6. doi:10.1136/tsaco-2017-000131.