EMETT: Ensemble Model for Emergency department Trauma Triage

Ludvig Wärnberg Gerdin Martin Gerdin Wärnberg

1/17/2022

Introduction

Trauma is a major threat to population health globally (Brohi & Schreiber, 2017; GBD 2016 Causes of Death Collaborators, 2017). Every year about 4.6 million people die because of trauma - more than the deaths from HIV/AIDS, malaria and tuberculosis combined. 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 (O'Reilly et al., 2013; Roy et al., 2017; Yeboah et al., 2014). Early identification and management of potentially life threatening injuries are crucial. Trauma triage - the process of prioritizing patients to match level of care with clinical acuity - is a key component of trauma care (Eastern Association for the Surgery of Trauma (EAST), 2010; National Institute for Health and Care Excellence (NICE), 2016).

In health systems with formalised criteria for emergency department trauma triage, all patients are assigned a priority coupled with a target time to treat. These priorities are may be coded with numbers (Agency for Healthcare Research and Quality, 2012) or colors (South African Triage Group, 2012), 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 triage trauma patients in the emergency department (Baker et al., 2013). Where prehospital care is lacking patients often arrive to the emergency department without warning (Choi et al., 2017). Identifying ways to quickly triage patients would therefore be valuable such settings.

The approach to triage trauma patients arriving to the emergency department 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 neural networks.

The uptake and use of such learners in trauma research has been slow (Liu & Salinas, 2017). One recent study used a random forest learner to assign priority to patients in a general emergency department population, and found a slight performance improvement using this learner compared with the standard criteria (Levin et al., 2018).

Given the paucity of research leveraging machine learning to triage trauma patients in the emergency department, we aimed to compare the performance of an ensemble machine learning model to that of clinician gestalt based on patients' presentation. Our hypothesis was that the performance of the this ensemble model would be non-inferior to that of clinician gestalt.

Materials and Methods

Study Design

We used data from the ongoing Trauma Triage Study (TTRIS) in India, a Towards Improved Trauma Care Outcomes (TITCO) study. This study is a prospective cohort study in three public hospitals in urban India.

Study Setting

Data analysed 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, orthopaedics, anaesthesia, 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 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 emergency departments. 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 predominately 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 emergency department 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 emergency department 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,

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 emergency department and the remaining two following up patients. The collected data were then transferred 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 emergency departements 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 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) code book 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 emergency department during each shift. The number of

patients to enrol was set to ten to make follow up feasible. Written informed consent from the patient or a patient representative was obtained either in the emergency department 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 enrolment 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 ensemble model variables The ensemble model were trained on two target variables. The first target 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 second target was a composite outcome of early mortality, ICU admission, major urgent surgery and severe injury. Early mortality was defined as death within 24 hours of arrival to the participating hospitals. ICU admission was defined as admission to the ICU within 48 hours of arrival. While most ICU admissions occur within hours of hospital arrival, we extended the time frame to compensate for bed availability and transfer delays. Major urgent surgery was defined as a major surgery performed within 24 hours of arrival to the participating hospitals. Surgical excerpts were reworked into standardized nomenclature using the Nordic Medico-Statistical Committee (NOMESCO) Classification of surgical Procedures, and the Systematized Nomenclature in Medicine Clinical Terms (SNOMED CT). In the lack of a general definition of major surgery, a team of experienced surgeons and researches decided on what surgeries to consider as major. Severe injury was defined as an Injury Severity Score (ISS) over 15. This cutoff-value for severe injury is traditionally sued in trauma research since it is said to indicate a 10 % mortality). **TAKEN FROM CELINAS PAPER**

The 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 (World Health Organization, 2018). 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 (glasgowcomascale.org, 2018). 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 (Rehn, Perel, Blackhall, & Lossius, 2011).

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 identified, 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. We used Python and R for all analysis (R Core Team, 2017; Van Rossum & Drake, 2009). In particular, we utilised packages (Pollard, Johnson, Raffa, & Mark, 2018) for generating sample characteristics table. 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 inter quartile 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 Ensemble Model The study sample was split into three parts, henceforth referred to as the training, validation, and test sets. We then developed our ensemble model in the training and validation setsusing the SuperLearner R package (Polley, LeDell, & Laan, 2016). SuperLearner is an ensemblemachine learning algorithm, meaning that it combines predictions several learners to come up with an "optimal" learner. Table@ref(tab:superlearner-library) shows our library of learners. Allwere implemented using the default hyperparameters. Short descriptionsof the individual learners are available as supportinginformation.

The ensemble model 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.

The ensemble model predictions was then used to assign levels of priority to patients. This was done by binning the ensemble model 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 cutoffs were identified in the validation set in order to prevent information leakage and limit bias. The performance of both the continuous ensemble model prediction and the ensemble model priority levels was then evaluated by estimating their AUROCC. We also visualised the performance by plotting ROC curves.

Algorithm	Package
Gradient Boosting Machine	LightGBM
Random Forest Classifier	scikit-learn
Multi-layer Perceptron	pytorch

Comparing the ensemble model and Clinicians We then used the ensemble model to predict the outcomes of the patients in the test set and used the cutoff values from the validation set to assign a level of priority to each patient in this set. The performance of the continuous ensemble prediction, the ensemble model priority levels, and the clinicians' priority levels, was then evaluated by estimating and comparing their AUROCC.

The levels of priority assigned by the ensemble model 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 ensemble model 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 ensemble model 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 ensemble model.

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 ensemble model.

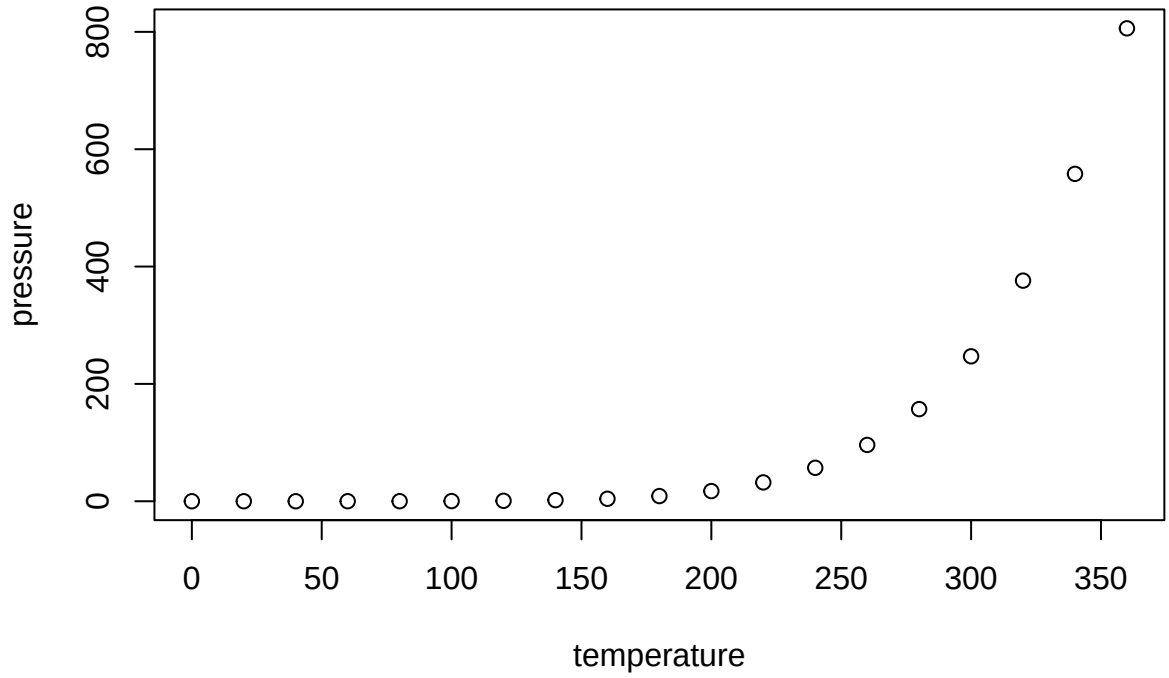
Results

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Agency for Healthcare Research and Quality. (2012). *Emergency Severity Index (ESI). A Triage Tool for Emergency Department Care* (No. Version 4; Vol. 117). U.S. Department of Health & Human Services. https://doi.org/10.1016/j.cmpb.2014.08.006

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

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

Choi, S. J., Oh, M. Y., Kim, N. R., Jung, Y. J., Ro, Y. S., & Shin, S. D. (2017). Comparison of trauma care systems in Asian countries: A systematic literature review. *Emergency Medicine Australasia*, *29*(June), 697–711. https://doi.org/10.1111/1742-6723.12840

Eastern Association for the Surgery of Trauma (EAST). (2010). *Practice Management Guidelines for the Appropriate Triage of the Victim of Trauma* (pp. 1–34). EAST.

GBD 2016 Causes of Death Collaborators. (2017). 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*, *390*(September 16), 1151–1210. https://doi.org/10.1016/S0140-6736(17)32152-9

glasgowcomascale.org. (2018). *GLASGOW COMA SCALE: Do it this way.* Retrieved from http://www.glasgowcomascale.org/downloads/GCS-Assessment-Aid-English.pdf?v=3

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

Liu, N. T., & Salinas, J. (2017). Machine Learning for Predicting Outcomes in Trauma. *Shock*, *48*(5), 504–510. https://doi.org/10.1097/SHK.0000000000000898

National Institute for Health and Care Excellence (NICE). (2016). *Major trauma: service delivery* (No. February). NICE.

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

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., & Mark, R. G. (2018). tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open*, *1*(1), 26–31. https://doi.org/10.1093/jamiaopen/ooy012

Polley, E., LeDell, E., & Laan, M. van der. (2016). *SuperLearner: Super Learner Prediction*. Retrieved from https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf$/backslash$nhttps://github.com/ecpolley/SuperLearner

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Rehn, M., Perel, P., Blackhall, K., & Lossius, H. M. (2011). Prognostic models for the early care of trauma patients: a systematic review. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *19*(1), 17. https://doi.org/10.1186/1757-7241-19-17

Roy, N., Veetil, D. K., Khajanchi, M. U., Kumar, V., Solomon, H., Kamble, J., ... Schreeb, J. V. (2017). Learning from 2523 trauma deaths in India- opportunities to prevent in-hospital deaths. *BMC Health Services Research*, *17*(142), 1–8. https://doi.org/10.1186/s12913-017-2085-7

South African Triage Group. (2012). *The South African Triage Scale Training Manual 2012* (pp. 1–34). Western Cape Government. Retrieved from Western Cape Government website: https://emssa.org.za/sats/

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

World Health Organization. (2018). *ICD-10 Interactive Self Learning Tool*. Retrieved from http://apps.who.int/classifications/apps/icd/icd10training/

Yeboah, D., Mock, C., Karikari, P., Agyei-Baffour, P., Donkor, P., & Ebel, B. (2014). 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*, *38*(7), 1707–1712. https://doi.org/10.1007/s00268-014-2452-z