




OPEN ACCESS

Triage in major incidents: development and external validation of novel machine learning-derived primary and secondary triage tools

Yuanwei Xu,^{1,2} Nabeela Malik ,^{3,4,5,6} Saisakul Chernbumroong,^{1,3} James Vassallo,^{7,8} Damian Keene,^{4,6} Mark Foster,^{3,4,6} Janet Lord,^{3,5} Antonio Belli,^{3,4} Timothy Hodgetts,⁹ Douglas Bowley,^{4,6} George Gkoutos^{1,2,3,10,11}

Handling editor Shammi L Ramlakhan

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/emermed-2022-212440>).

For numbered affiliations see end of article.

Correspondence to

Miss Nabeela Malik, NIHR Surgical Reconstruction Microbiology Research Centre, Birmingham B15 2TH, UK; nabeelamalik@nhs.net

YX and NM are joint first authors.

Received 7 March 2022
Accepted 12 August 2023
Published Online First
26 September 2023



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ.

To cite: Xu Y, Malik N, Chernbumroong S, et al. *Emerg Med J* 2024;**41**:176–183.

ABSTRACT

Background Major incidents (MIs) are an important cause of death and disability. Triage tools are crucial to identifying priority 1 (P1) patients—those needing time-critical, life-saving interventions. Existing expert opinion-derived tools have limited evidence supporting their use. This study employs machine learning (ML) to develop and validate models for novel primary and secondary triage tools.

Methods Adults (16+ years) from the UK Trauma Audit and Research Network (TARN) registry (January 2008–December 2017) served as surrogates for MI victims, with P1 patients identified using predefined criteria. The TARN database was split chronologically into model training and testing (70:30) datasets. Input variables included physiological parameters, age, mechanism and anatomical location of injury. Random forest, extreme gradient boosted tree, logistic regression and decision tree models were trained to predict P1 status, and compared with existing tools (Battlefield Casualty Drills (BCD) Triage Sieve, CareFlight, Modified Physiological Triage Tool, MPTT-24, MSTART, National Ambulance Resilience Unit Triage Sieve and RAMP). Primary and secondary candidate models were selected; the latter was externally validated on patients from the UK military's Joint Theatre Trauma Registry (JTTR).

Results Models were internally tested in 57 979 TARN patients. The best existing tool was the BCD Triage Sieve (sensitivity 68.2%, area under the receiver operating curve (AUC) 0.688). Inability to breathe spontaneously, presence of chest injury and mental status were most predictive of P1 status. A decision tree model including these three variables exhibited the best test characteristics (sensitivity 73.0%, AUC 0.782), forming the candidate primary tool. The proposed secondary tool (sensitivity 77.9%, AUC 0.817), applicable via a portable device, includes a fourth variable (injury mechanism). This performed favourably on external validation (sensitivity of 97.6%, AUC 0.778) in 5956 JTTR patients.

Conclusion Novel triage tools developed using ML outperform existing tools in a nationally representative trauma population. The proposed primary tool requires external validation prior to consideration for practical use. The secondary tool demonstrates good external validity and may be used to support decision-making by healthcare workers responding to MIs.

INTRODUCTION

In the immediate aftermath of a major incident (MI), patient needs exceed the resources available to treat

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ During major incidents (MIs) (eg, terrorist attacks), triage tools have a crucial role in maximising overall survival by identifying priority 1 (P1) patients.
- ⇒ Existing tools, derived using expert opinion, have limited evidence to support their use.

WHAT THIS STUDY ADDS

- ⇒ In this study, novel machine learning-based primary and secondary triage tools surpassed the current UK National Ambulance Resilience Unit Triage Sieve and other existing tools in identifying P1 patients within a nationally representative trauma population.
- ⇒ The secondary tool demonstrated favourable external validity. However, the primary tool could not be externally validated due to missing GCS component data.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ The proposed secondary tool, applicable using a portable device, may be used to support decision-making among healthcare workers responding to MIs.

them^{1–5}: triage tools seek to categorise patients, to guide the order of treatment, transport from the scene and the choice of medical facility for definitive care.^{5,6} A vital function of triage tools is to identify patients requiring time-critical, life-saving interventions (priority 1 or P1 patients). Failure to identify these patients (undertriage) is associated with absolute harm arising from delays in care or selection of an inappropriate medical facility.^{6,7} However, overtriage may risk overwhelming healthcare facilities with patients not requiring time-critical treatment.²

Primary triage, conducted at the scene of an MI, uses paper-based flow diagrams that are quick and simple to apply under challenging conditions.⁸ Existing primary triage tools have largely been developed using expert opinion, often with limited evidence to support their use.⁶ These include the National Ambulance Resilience Unit (NARU) Triage Sieve (current UK tool for adults), the Australian CareFlight and the US Simple Triage And Rapid Treatment (START).^{6,9,10} These tools



use ambulatory status to designate priority 3 (minor) category, followed by physiological assessments to distinguish P1 from P2 (less critical) patients. A recent study demonstrated that the UK military's Battlefield Casualty Drills (BCD) Triage Sieve attained greatest sensitivity among 10 international primary triage tools in detecting P1 status among adults; however, this was associated with an overtriage rate of 72%.¹⁰

Primary triage is often, but not always, followed by a further targeted prehospital clinical assessment of patients known as secondary triage. This is usually undertaken in a place of relative safety (eg, Casualty Clearing Station or hospital reception area)¹⁻⁸; thus, the additional use of medical equipment and/or portable devices is more plausible. Two existing secondary MI triage tools are the UK's Major Incident Medical Management and Support Triage Sort which has suboptimal sensitivity (15.7%) in predicting the need for life-saving intervention,¹¹ and the US points-based Sacco Triage Method (developed to predict mortality) which is time-consuming and complex to apply.⁹

Anatomical assessment of injuries has yet to feature in any MI triage tool, yet this is commonly used in the field triage of singly injured patients.¹² Advanced age is associated with worse outcomes following injury; however, existing tools do not incorporate this in patient assessment.¹³ There is scope to develop evidence-based primary and secondary MI triage tools which offer greater sensitivity while decreasing overtriage compared with the BCD Triage Sieve, yet preserve applicability. Tree-based machine learning models have demonstrated utility in clinical risk stratification, with the ability to capture non-linear interactions between input variables.¹⁴⁻¹⁵ This study aimed to develop machine learning models that can be adapted into primary and secondary MI triage tools and to externally validate these models using an independent population of injured patients.

METHODS

Database for model training and internal testing

Model development and validation were conducted according to Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis guidelines.¹⁶ Adult (16+ years) patients from the Trauma Audit and Research Network

(TARN) registry presenting between 1 January 2008 and 31 December 2017 were included.¹⁷ The TARN registry prospectively captures prehospital and hospital data from 169 hospitals in England and Wales for patients who meet the following inclusion criteria: length of stay >72 hours or admission to intensive care and/or death in hospital.¹⁷ TARN does not include prehospital deaths. Patients for whom any input variables required for modelling were missing were excluded. Using hospital arrival dates recorded by TARN, the database was split temporally (70:30) to generate model training and internal testing datasets, respectively.

Primary outcome of interest

The primary outcome of interest was P1 status, defined as the need for time-critical life or limb-saving surgery and/or advanced resuscitative measures.¹⁸ Each patient was retrospectively designated a triage category (priority 1, priority 2, priority 4/expectant or dead) (see flow diagram in online supplemental figure 1) using validated, consensus-derived definitions (table 1).^{10,18} Prior to the modelling phase, patients were designated either P1 or non-P1. The small numbers of P4 and dead patients (who share physiological similarities to P1 patients) were excluded from the modelling as these may impede model performance.

Input variables selected for modelling

Input variables differ in their complexity and time taken for measurement. Variables that can be readily assessed by first responders in the MI setting were included in the modelling process (summarised in online supplemental table 1). This included all physiological parameters used by existing MI triage tools (first-recorded prehospital HR, RR and systolic BP) with the exception of capillary refill time, which has been found to be a poor reflection of circulatory status and is difficult to measure reliably in challenging settings and in non-white patients.^{6,10,19} In addition to the ability to follow commands (GCS Motor) used by the CareFlight triage tool, all subcomponents of the GCS were included.⁶ However, total GCS score, although known to be an important predictor of outcomes in injured patients,

Table 1 Triage category definitions

Dead	<ul style="list-style-type: none"> ▶ Cardiac and/or respiratory arrest at initial prehospital evaluation that is not responsive to needle decompression or airway positioning (or the delivery of two rescue breaths in children less than 12 years old) ▶ Lack of palpable pulse and need for CPR (ie, cardiac arrest) within the first 15 min of EMS arrival on scene
Priority 4 (expectant)	<ul style="list-style-type: none"> ▶ In patients aged 0–49 years: third-degree (full thickness) burns to >90% of the body ▶ In patients aged 50 years and over: third-degree (full thickness) burns to >80% of the body ▶ Penetrating trauma to the head that crosses the midline with agonal respirations and/or no motor response, decorticate posturing or decerebrate posturing (ie, GCS Motor ≤3) ▶ Blunt trauma to the head with agonal respirations and/or no motor response, decorticate posturing or decerebrate posturing (ie, GCS Motor ≤3) ▶ Uncontrolled haemorrhage that resulted in cardiac arrest (defined as a lack of palpable pulse and EMS initiation of CPR) prior to EMS transport
Priority 1	<ul style="list-style-type: none"> ▶ Neurological, vascular or haemorrhage-controlling surgery to the head, neck or torso performed within 4 hours of arrival to hospital ▶ Limb-conserving surgery performed within 4 hours of arrival at hospital on a limb that was found to be pulseless distal to the injury prior to surgery ▶ Escharotomy performed on a patient with burns within 2 hours of arrival at a hospital ▶ Chest tube placed within 2 hours of arrival at hospital ▶ An advanced airway intervention (eg, intubation, LMA, surgical airway) performed in the prehospital setting or within 4 hours of arrival at hospital ▶ IV vasopressors administered within 2 hours of arrival at hospital ▶ Arrived in the ED with uncontrolled haemorrhage ▶ Patient who required EMS initiation of CPR (ie, had a cardiac arrest) during transport, in the ED or within 4 hours of arrival at a hospital
Priority 2	<ul style="list-style-type: none"> ▶ All patients who do not meet the criteria for the other categories are considered priority 2
Priority 3	<ul style="list-style-type: none"> ▶ Discharged from ED with no X-rays or an extremity X-ray that was negative or showed an uncomplicated fracture (ie, a closed extremity fracture without significant displacement or neurovascular compromise); no laboratory testing; received only simple wound repair (single-layer suturing only); and received no medications intravenously (does not include fluids), or inhaled (does not include oxygen) from EMS or in the hospital
<p>These definitions were derived by expert consensus and have been validated in a UK trauma population. Priority 4 (expectant) denotes injuries which are incompatible with life. CPR, cardiopulmonary resuscitation; EMS, emergency medical services; IV, intravenous; LMA, laryngeal mask airway.</p>	

was not included.^{7 12} Total GCS is time-consuming to calculate, with evidence suggesting that scores by paramedics frequently differ from those assigned by emergency physicians; hence, measurement under MI conditions may lack accuracy.^{5 19 20} The ability to breathe spontaneously is an important determinant of outcome and is assessed early within several existing triage tools.^{6 10} TARN does not explicitly record whether patients are spontaneously breathing at the scene of injury, nor does it record the indication for airway interventions.¹⁷ We assumed that all patients who received an advanced airway intervention at the scene (defined as intubation and ventilation and/or surgical airway and/or the need for airway support) were unable to breathe spontaneously.^{10 21}

The presence of injury in anatomical regions including the head, face, chest and limb(s) was included as input variables for modelling using retrospectively calculated Abbreviated Injury Severity (AIS) scores (TARN records AIS based on hospital rather than prehospital data). A binary input (AIS=0, AIS >0) was used rather than a graded assessment of severity. Due to the known difficulties in identifying intra-abdominal injuries based on clinical assessment alone, and the requirement to undertake detailed clinical assessment in order to reliably identify spinal injuries, the presence of spinal and abdominal injuries was not included as input variables.^{22 23} Patient age was dichotomised into age ≥ 65 years (yes or no), which may be reliably identified by first responders.¹² Broad injury mechanism (blunt or penetrating) was included.

Input variables described thus far were deemed appropriate for inclusion in both primary and secondary triage tools. Although not conducive to primary triage due to the need for calculation, shock index (HR/systolic BP), which may correlate better with outcome than HR or systolic BP alone, was included in the modelling process as a potential component of a secondary triage tool.²⁴

Model training and internal testing

Four machine learning methods were applied to the model training dataset to distinguish P1 from non-P1 patients. Decision tree (RPART) methodology was included because models can be visualised as bifurcating trees, closely resembling the format of existing primary triage tools. Two other tree-based models with demonstrated value in clinical risk stratification, random forest (RF) and eXtreme Gradient Boosting (XGB), were trained.^{25 26} Further methodological details are presented as online supplemental material. Finally, we included an L1-regularised logistic regression model. We anticipated that non-P1 patients would substantially outnumber P1 patients; hence, we adopted an undersampling strategy to balance the data by leaving out random samples of non-P1 patients.¹⁴ For each of these models, fivefold cross-validation was applied.²⁶

To generate models that were no more complex to apply than existing primary triage tools, modelling included all possible combinations of 3–7 of the available 13 input variables. Model building and selection strategy are summarised in online supplemental figure 2. Models trained using all 13 input variables, although too complex for practical application as triage tools, were also considered as comparators (online supplemental table 2). Additionally, we compared the triage assignments (namely, P1 status) of 10 existing international primary triage tools to the testing dataset (online supplemental table 3).¹⁰

Previous studies demonstrate that elders (aged 65+ years) are over-represented in the TARN population while constituting 18.3% of the UK population¹⁰; hence, during testing, we split

the TARN testing set by age (ages 16–64 years and 65+ years) to further evaluate model performance.

Determining feature importance

We assessed the relative importance of individual features (input variables) in model predictions using the TreeSHAP method, a model-agnostic, individualised feature attribution method for explaining predictions.²⁷ The resulting Shapley value for a particular feature measures the expected change in model prediction when that feature is present relative to the average model prediction. Additionally, feature importance was estimated by the contribution of each feature to the overall XGB model-predictive performance.²⁷

Selection of models as candidates for primary and secondary triage tools

We sought to identify models that achieved the best possible performance (maximal sensitivity in identifying P1 patients, but also favourable overtriage rate and area under the receiver operating curve (AUC)) across all ages as well as age subgroups, using the minimal number of input variables, to maintain practical applicability. We predetermined that selected models must outperform the best performing existing triage tool, as identified by our study.

In keeping with existing practice, the primary tool candidate was intended to be a paper-based, simple algorithm. The model selected as a secondary tool was adapted into a web-based prototype using the R shiny application.

External validation of models using the Joint Theatre Trauma Registry database

The UK military's Joint Theatre Trauma Registry (JTTR) (February 2002–December 2016) was used to externally validate the selected models. JTTR includes consecutive patients who triggered trauma team activation at a deployed medical treatment facility, largely comprising combat casualties during military operations in Iraq and Afghanistan.

Children (<16 years), patients with erroneous data (eg, age over 110 years) and those with injuries recorded as both blunt and penetrating were excluded from the validation (see online supplemental figure 1). As we expected a paucity of prehospital data in this population,²⁸ patients' first recorded hospital physiology was used. Patients with missing data for the input variables were not excluded. Subcomponents of GCS are not routinely recorded within JTTR; these were derived for patients with GCS 15 and unavailable for those with GCS <15. Furthermore, we evaluated candidate models on a subset of JTTR patients with sufficient data to apply the best performing existing tool (subsequently found to be the BCD Triage Sieve), thereby facilitating direct comparison. Triage category definitions were applied as described earlier (table 1): since JTTR does not record the time of interventions, those performed at deployed medical treatment facilities were presumed to have occurred within 4 hours.²⁸

Statistical analyses

Patient characteristics across the model training, internal testing and external validation datasets were compared using the χ^2 test (Injury Severity Score (ISS) and age compared using Mann-Whitney U test); $p < 0.05$ was considered statistically significant. Model performance is reported as sensitivity, specificity, undertriage (1-sensitivity) and overtriage (1-positive predictive value). The 95% CIs for the AUC were calculated using deLong's method (pROC R package, V.1.17.0.1).²⁹ The 95% CIs for

Table 2 Patient and injury characteristics for the model training, testing and external validation cohorts

	Model training dataset (70% TARN: 1 Jan 2008–14 Jul 2016)	Model testing dataset (30% TARN: 15 Jul 2016–31 Dec 2017)	External validation dataset (JTTR: 1 Feb 2002–31 Dec 2016)
Gender			
Male	72 817 (53.8%)	29 532 (50.9%)	5830 (97.9%)
Female	62 465 (46.2%)	28 447 (49.1%)	106 (1.8%)
Missing data	0 (0.0%)	0 (0.0%)	20 (0.3%)
Injury Severity Score			
Median (IQR)	9 (9–16)	9 (9–17)	8 (2–17)
Missing data	0 (0.0%)	0 (0.0%)	13 (0.2%)
Age			
Median (IQR)	64.3 (45.6–82.3)	70.9 (51.6–84.5)	24 (21–28)
16–64 years	69 237 (51.2%)	24 769 (42.7%)	5256 (88.2%)
65+ years	66 045 (48.8%)	33 210 (57.3%)	25 (0.4%)
Missing data	0 (0.0%)	0 (0.0%)	675 (11.3%)
Discharge status			
Alive	127 624 (94.3%)	54 383 (93.8%)	5681 (95.4%)
Dead	7657 (5.7%)	3596 (6.2%)	275 (4.6%)
Missing data	1 (0.0%)	0 (0.0%)	0 (0.0%)
Injury mode			
Blunt	131 208 (97.0%)	56 473 (97.4%)	1092 (18.3%)
Penetrating	4074 (3.0%)	1506 (2.6%)	4864 (81.7%)
Missing data	0 (0.0%)	0 (0.0%)	0 (0.0%)
Injury mechanism			
Fall less than 2 m	76 169 (56.3%)	36 380 (62.7%)	78 (1.3%)
Vehicle incident	30 195 (22.3%)	10 744 (18.5%)	389 (6.5%)
Fall more than 2 m	17 838 (13.2%)	6725 (11.6%)	37 (0.6%)
Blow(s)	4871 (3.6%)	1868 (3.2%)	0 (0.0%)
Stabbing	2871 (2.1%)	1192 (2.1%)	29 (0.5%)
Crush	1065 (0.8%)	268 (0.5%)	76 (1.3%)
Shooting	328 (0.2%)	91 (0.2%)	2316 (38.9%)
Burn	91 (0.07%)	27 (0.05%)	3 (0.1%)
Blast	88 (0.07%)	50 (0.09%)	2926 (49.1%)
Other	1766 (1.3%)	634 (1.1%)	86 (1.4%)
Missing data	0 (0.0%)	0 (0.0%)	16 (0.3%)

JTTR, Joint Theatre Trauma Registry; TARN, Trauma Audit and Research Network.

models' sensitivity at given specificity points were calculated using 500-stratified bootstrap replicates.²⁹

Patient and public involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

RESULTS

Training dataset and primary outcome of interest

A total of 200 728 patients were captured by TARN over the 10-year period. After exclusions, the sample consisted of 193 261 patients, of which 21 878 patients (11.3%) fulfilled P1 criteria.

The model training dataset comprised 135 282 patients, with a median age of 64.3 years, in-hospital mortality of 5.7% and predominantly blunt injuries (97%), most commonly low falls (56.3%) (table 2). Patients within the internal test dataset (n=57 979) were older (median age 70.9 years vs 64.3 years, respectively, $p<0.001$) and more often injured by a low-level fall (62.7% vs 56.3%, $p<0.001$) compared with patients within the model training dataset.

Model training and internal testing

In the test set, the BCD Triage Sieve demonstrated the greatest sensitivity at 68% with overtriage at 80.8% (table 3). Existing tools performed less well in the elders' subgroup compared with

younger (16–64 years) adults, with sensitivity 5.8–14.6% lower and overtriage rates 11.5–33.2% higher among elders (online supplemental table 3).

Four hundred fifty-six models were developed, which, when applied to the internal test dataset, demonstrated greater sensitivity and AUC than all existing tools. Model selection was initially narrowed down to five decision tree models as candidates for primary triage tools and 29 XGB models as candidates for secondary triage tools (see online supplemental figure 2). A comprehensive list, including performance by age subgroups within the internal (TARN) testing and external validation (JTTR) datasets (described later), is detailed in online supplemental table 4A–C. Receiver operating curves demonstrating the performance of the novel primary and secondary tool candidate models when applied to the internal testing dataset are shown in figure 1.

Feature importance

The top 10 features (figure 2A), and their relative contribution in predicting P1 status (figure 2B) are presented. By far, the most important variable was breathing status (mean Shapley value 1.2), followed by presence of a chest injury and GCS Verbal score. Age >65 years was negatively predictive of P1 status. Any abnormal GCS Verbal or GCS Motor score contributed substantially in predicting P1 status (see figure 2B). The XGB method of

Table 3 Performance characteristics of existing triage tools and novel machine learning models among adult patients (16+ years) in the testing (TARN) dataset

	Sensitivity	Specificity	Undertriage	Overtriage	AUC
Existing tools					
BCD Triage Sieve	68.2 (66.9, 69.4)	69.5 (69.1, 69.9)	31.8 (30.6, 33.1)	80.8 (80.2, 81.3)	0.688 (0.682, 0.695)
CareFlight	39.9 (38.6, 41.2)	94.5 (94.3, 94.7)	60.1 (58.8, 61.4)	56.4 (55.0, 57.8)	0.672 (0.666, 0.679)
MPTT-24	48.4 (47.1, 49.7)	66.4 (66.0, 66.8)	51.6 (50.3, 52.9)	86.7 (86.2, 87.2)	0.574 (0.567, 0.581)
MSTART	54.9 (53.6, 56.2)	88.4 (88.1, 88.7)	45.1 (43.8, 46.4)	66.5 (65.5, 67.5)	0.717 (0.710, 0.723)
NARU Triage Sieve	43.0 (41.7, 44.3)	88.3 (88.1, 88.6)	57.0 (55.7, 58.3)	71.8 (70.9, 72.8)	0.657 (0.650, 0.663)
RAMP	37.1 (35.9, 38.4)	94.6 (94.5, 94.8)	62.9 (61.6, 64.1)	57.5 (56.1, 58.9)	0.659 (0.653, 0.665)
Models selected as candidates for novel primary and secondary triage tools					
Primary triage tool candidate (decision tree)	73.0 (71.8, 74.2)	73.9 (73.5, 74.3)	27.0 (25.8, 28.2)	77.0 (76.4, 77.7)	0.782 (0.775, 0.789)
Secondary triage tool candidate (XGB)	77.9 (76.8, 79.0)	73.1 (72.7, 73.5)	22.1 (21.0, 23.2)	76.4 (75.8, 77.0)	0.817 (0.810, 0.824)

Values shown are percentages (except for AUC), accompanied by 95% CIs.
 *The best performing model using each method is shown. Both machine learning models and the triage tools were evaluated using the same TARN population (internal testing dataset).
 AUC, area under the receiver operating curve; BCD, Battlefield Casualty Drills (UK Military); MPTT-24, Modified Physiological Triage Tool 24 (2017); NARU, National Ambulance Resilience Unit (current UK civilian triage tool); TARN, Trauma Audit and Research Network; XGB, eXtreme Gradient Boosting.

determining feature importance yielded similar rankings (online supplemental figure 3).

Primary and secondary triage tool candidate models

The decision tree model selected for clinical adaptation into a primary triage tool (figure 3) used three qualitative binary (yes/no) assessments (breathing status at scene, ability to obey

commands, that is, GCS Motor score=6, and presence of a chest injury) to categorise patients as P1 or non-P1. This achieved 73.0% sensitivity, overtriage rate of 77.0% and AUC of 0.782 when applied to the internal testing dataset (see table 3).

The XGB model selected as a secondary triage tool (figure 4) combines four input variables: GCS Motor score, breathing status at scene, presence of chest injury and classification of injury as blunt or penetrating. This model achieved 77.9% sensitivity, overtriage of 76.4% and AUC of 0.817 when applied to the internal testing dataset (figure 1 and table 3). This has been adapted into an online interactive tool (accessible via link: <https://ywxtrriageapp.shinyapps.io/mltriage/>).

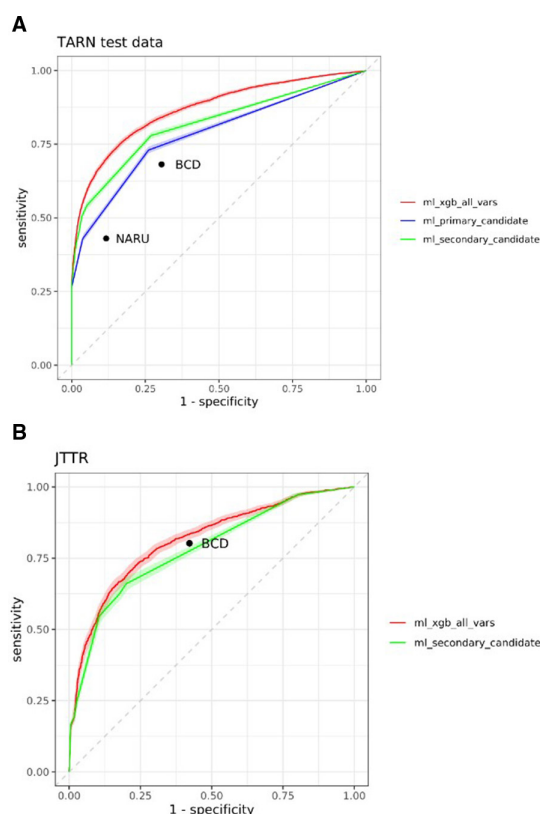


Figure 1 Performance of tool candidate models in the internal and external validation datasets compared with the Battlefield Casualty Drills (BCD) Triage Sieve (best performing existing tool) and the current UK tool, the National Ambulance Resilience Unit (NARU) Triage Sieve. Additionally, the performance of an XGB model using all 13 input variables is shown for comparison (see online supplemental material for more details). JTTR, Joint Theatre Trauma Registry; TARN, Trauma Audit and Research Network; XGB, eXtreme Gradient Boosting.

External validation of the secondary triage model (JTTR)

A total of 5956 JTTR patients met inclusion criteria (online supplemental figure 1). Median age was 24 years (IQR 21–28) and most were male (97.9%). Compared with patients in the TARN model training set, JTTR patients had lower mortality (4.6% vs 5.7%, $p < 0.001$) and lower injury severity (median ISS 8 (IQR 2–17) vs median ISS 9 (IQR 9–16), $p = 0 < 0.001$). A greater proportion of JTTR patients suffered penetrating trauma (81.7% vs 3.0%, $p = 0 < 0.001$), with high prevalence of blast injury (49.1% vs 0.07%, $p = 0 < 0.001$) and shooting (38.9% vs 0.2%, $p = 0 < 0.001$) (see table 2). A total of 2046 (34.3%) JTTR patients had missing GCS Motor scores.

Given the high proportion of JTTR patients missing GCS Motor scores, as well as inability for decision trees to perform predictions when data are missing (unlike XGB and RF), application of the primary tool candidate model to JTTR patients would not reliably measure the model's external validity. Hence, this was not performed.

Performance of the models shortlisted as candidates for a secondary triage tool for JTTR patients is shown in online supplemental table 4B and model calibration is presented as online supplemental figure 4. The model selected as a secondary tool (XGB model, ID 37) achieved sensitivity of 97.6%, overtriage of 57.5% and AUC of 0.778 (figure 1). Secondary candidate models were evaluated on a subset of JTTR patients containing sufficient data to apply the BCD Triage Sieve ($n = 5455$), thereby facilitating direct comparison (online supplemental table 5): the secondary tool candidate attained comparatively higher sensitivity (97.3% vs 80.2%), but had a higher overtriage rate (58.5% vs 47.4%).

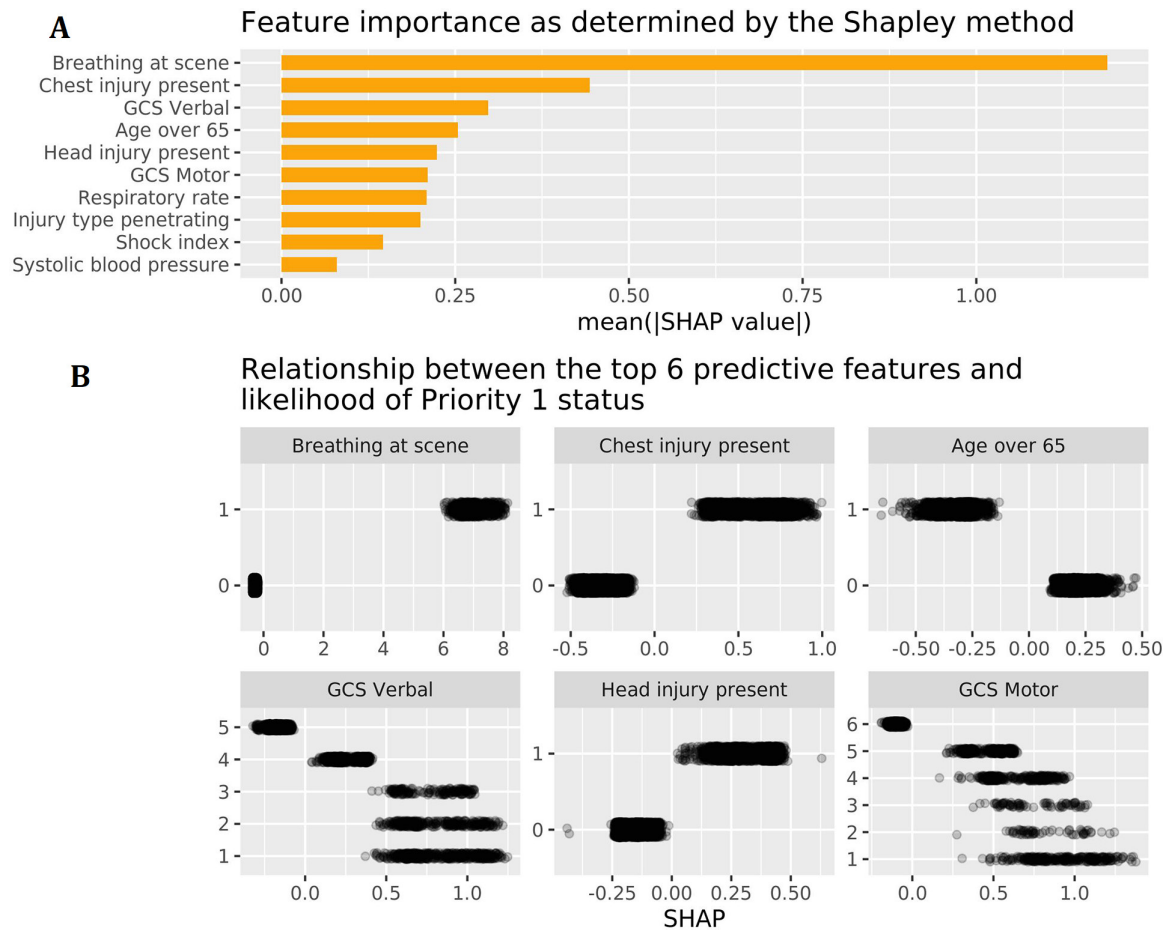


Figure 2 (A) Mean absolute Shapley value for the top 10 predictors. This is followed by the (B) Shapley values for the top six most important features (Shapley values are shown on the x axis, feature values are shown on the y axis). Large, positive Shapley values represent a greater contribution to the likelihood of P1 status. Negative Shapley values represent contributions to non-P1 status. Age over 65 years was found to be negatively predictive of P1 status. GCS Motor, motor subcomponent of the GCS; P1, priority 1.

DISCUSSION

We have developed MI triage tools based on machine learning models that outperform 10 existing international triage tools in predicting the need for time-critical interventions (P1 status) among adults. The best existing primary triage tool, the BCD Triage Sieve, demonstrated sensitivity of 68.2% and overtriage of 80.8% (AUC 0.688), while the selected machine learning primary triage tool achieved a sensitivity of 73% and overtriage of 77% (AUC 0.782). The model selected as a secondary

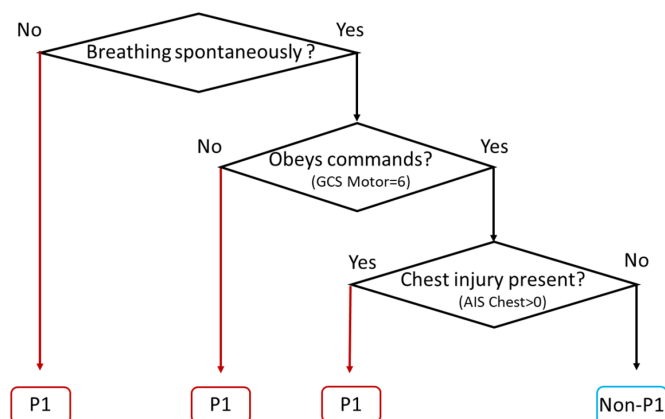


Figure 3 AIS, Abbreviated Injury Severity; P1, priority 1.

MI triage tool achieved sensitivity of 77.9% and an overtriage rate of 76.4% (AUC 0.817). When externally validated, the secondary tool demonstrated excellent performance with sensitivity of 97.6% and overtriage of 57.5% (AUC 0.778). External validation of the primary tool was precluded by a lack of GCS subcomponent data within the UK combat casualty registry. A novel aspect of this exercise was including anatomical assessment of injuries as part of an MI triage tool and presence of a chest injury was found to be one of the most important variables. Our models serve as evidence-based alternatives to existing tools.

The models proposed are based entirely on qualitative assessments. Eliminating arithmetic calculations (RR and HR) from triage under challenging circumstances has been advocated by expert consensus.¹⁹ The proposed four-variable secondary tool may also reduce triage time relative to the seven-step NARU and BCD Triage Sieve tools. In addition, decision support using portable device applications has established utility in the MI setting, exemplified by CitizenAID, which enables mutual aid by members of the general public.³⁰ Triage using a portable device could help to minimise interuser variability and human error.

Breathing status was the most important predictor of P1 status; this constitutes the opening step in several existing tools.⁶ Our study concurs with the findings of Wallis and Carley, who determined that the GCS Motor component was strongly predictive of P1 status.³¹ The finding that age >65 years is negatively associated with P1 status may be confounded by the predominantly

ML for Triage

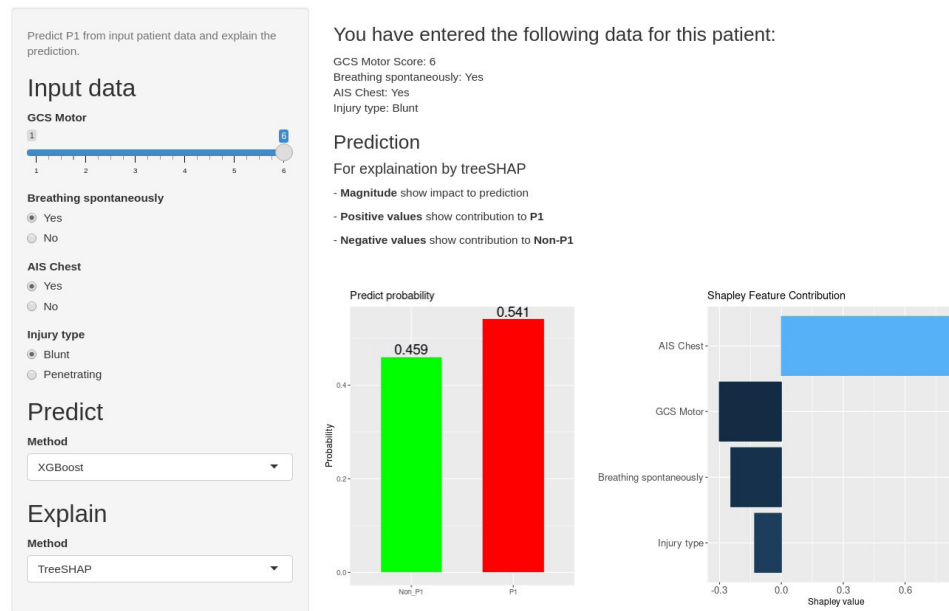


Figure 4 An interactive online application is demonstrated at <https://ywxtriageapp.shinyapps.io/mltriage/>. AIS, Abbreviated Injury Severity; GCS Motor, motor subcomponent of the GCS; ML, machine learning; P1, priority 1.

low-risk injury mechanism (low-level falls) in elders in our training dataset: hence, these patients are a poor surrogate for elders injured in an MI. Further work is required to develop effective trauma triage tools for elders, who differ in their physiology, and in whom presence of comorbidities and/or frailty is an important determinant of outcome.¹³ Penetrating mechanism was also an important predictor of P1 status: MIs involving penetrating trauma have historically yielded larger proportions of P1 patients.⁵

A key strength of this study is use of a large sample of injured patients using prospective data collected by trained TARN coordinators.¹⁷ The primary outcome measure chosen for this study is the only validated outcome measure for MI triage tool performance.¹⁰ A further strength is that the proposed secondary triage tool has undergone blinded, external validation using the UK military's JTTR database. This provides estimates of the model's predictive capability overall, but importantly, also among patients with blast and penetrating mechanisms (under-represented in the TARN dataset) typical of terrorist attacks, the prevalent type of UK MI in recent years.¹ Selection of an XGB model as a secondary tool, which can make predictions in the context of some missing data, has avoided the possible bias which can result from multiple imputation. Importantly, based on the TARN patients included in our study, both novel tools would generate proportions of P1 casualties that fall within UK national mass casualty planning assumptions.³² Notably, no UK or international guidance exists to define acceptable rates of undertriage and overtriage in the major incident setting.

Limitations of this study include use of retrospectively calculated AIS scores (incorporating CT and operative findings) during modelling in place of documented prehospital clinical assessment. While paramedics routinely conduct anatomical assessments during triage in singly injured patients using existing field triage tools and clinical assessment has proven effective in ruling out clinically significant chest injuries, some overtriage can be expected.^{12,33} Clinicians have performed improvised anatomical-based secondary triage following two mass shooting incidents, with a subsequent low rate of undertriage.⁵ Another limitation is

the use of singly injured patients within a civilian trauma registry as surrogates for those injured in an MI; outcomes in the MI setting may be worse. Our models focus on predicting P1 status only: however, these patients are at greatest risk of preventable death. In current UK practice, a small proportion of P1 patients may be subsequently assigned P4/expectant status by a senior clinician at scene; this contrasts with practice elsewhere, where triage tools fulfil this role (eg, Australian CareFlight and US START tools).^{6,32} Exclusion of P4 patients (<1% of the sample size) from the modelling process is unlikely to have impacted significantly on study findings. Application of models to the first recorded hospital physiology in JTTR may be biased by prehospital interventions; however, collection of prehospital physiological data during combat is particularly challenging.²⁸ The results of external validation in a military trauma population may have limited generalisability to the civilian setting. Further validation of our models in a true MI dataset or a prospective UK civilian database, including blast/penetrating trauma and burns, would provide further assurance of the models' performance. A further limitation is that we were unable to externally validate our proposed primary tool due to the paucity of prehospital vital signs (GCS) documented in the JTTR dataset.

In conclusion, using machine learning, we developed primary and secondary triage tools which differ from prior tools by incorporating anatomical assessment and have superior sensitivity and more favourable overtriage rates. Although the primary tool requires external validation among patients with injuries similar to those sustained in MI, the proposed secondary triage tool, which was externally validated, may be suitable for use in civilian hospital reception areas and in the military evacuation chain during MIs prior to or in conjunction with senior clinician triage using a portable device.

Author affiliations

¹Centre for Computational Biology, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK

²Health Data Science Centre, University of Birmingham, Birmingham B15 2TT, UK

³NIHR Surgical Reconstruction Microbiology Research Centre, Edgbaston, UK

⁴University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

⁵Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK
⁶Academic Department of Military Surgery & Trauma, Royal Centre for Defence Medicine, Mindelsohn Way, Edgbaston, Birmingham B152WB, UK
⁷Emergency Department, Derriford Hospital, Plymouth, UK
⁸Academic Department of Military Emergency Medicine, Royal Centre for Defence Medicine, Mindelsohn Way, Edgbaston, Birmingham B15 2WB, UK
⁹UK Strategic Command, Northwood Headquarters, Northwood, UK
¹⁰Institute of Translational Medicine, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
¹¹MRC Health Data Research UK (HDR UK), Birmingham, UK

Contributors NM and YX contributed equally to this study. NM and GG designed the study. YX, NM and SC accessed the TARN database, verified the underlying data and conducted analysis. JV and YX accessed JTTR data, JV verified the underlying data and YX conducted analysis. All authors contributed to data interpretation. NM (clinician) and YX (machine learning expert) wrote the initial draft of the manuscript. All authors contributed to critical revisions of subsequent manuscript drafts and approved the final version.

Funding This study is funded by the National Institute for Health Research (NIHR) Surgical Reconstruction and Microbiology Research Centre. GG acknowledges support from the NIHR Birmingham ECMC, Nanocommons H2020-EU (731032), MAESTRIA (grant agreement ID 965286) and the MRC Health Data Research UK (HDRUK/CFC/01), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities.

Disclaimer The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR, the Medical Research Council, the Department of Health and Social Care, or the Ministry of Defence.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval The UK Health Research Authority Patient Information Advisory Group (Section 20) has granted ethical approval and waived the requirement for individual patient consent for research using anonymised TARN data. The Ministry of Defence (through its Medical Directorate) granted approval for the use of anonymised JTTR data.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. All data relevant to the study are included in the article or uploaded as supplemental information. De-identified patient data used for this study are proprietary to the Trauma Audit and Research Network (TARN), University of Manchester, and may be requested directly from TARN.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iD

Nabeela Malik <http://orcid.org/0000-0003-0791-4158>

REFERENCES

- Dark P, Smith M, Ziman H, *et al*. Healthcare system impacts of the 2017 Manchester arena bombing: evidence from a national trauma registry patient case series and hospital performance data. *Emerg Med J* 2021;38:746–55.
- Frykberg ER, Tepas JJ. Terrorist bombings. Lessons learned from Belfast to Beirut. *Ann Surg* 1988;208:569–76.
- Hirsch M, Carli P, Nizard R, *et al*. The medical response to multisite terrorist attacks in Paris. *Lancet* 2015;386:2535–8.
- Moran CG, Webb C, Brohi K, *et al*. Lessons in planning from mass casualty events in UK. *BMJ* 2017;359:j4765.
- Turner CDA, Lockey DJ, Rehn M. Pre-hospital management of mass casualty civilian shootings: a systematic literature review. *Crit Care* 2016;20:362.
- Jenkins JL, McCarthy ML, Sauer LM, *et al*. Mass-casualty triage: time for an evidence-based approach. *Prehosp Disaster Med* 2008;23:3–8.
- Lerner EB, Schwartz RB, Coule PL, *et al*. Mass casualty triage: an evaluation of the data and development of a proposed national guideline. *Disaster Med Public Health Prep* 2008;2 Suppl 1:S25–34.
- Group ALS. *Major incident medical management and support*. 3rd Edition ed. Wiley-Blackwell, 2011.
- Jain TN, Ragazzoni L, Stryhn H, *et al*. Comparison of the Sacco triage method versus START triage using a virtual reality scenario in advance care paramedic students. *CJEM* 2016;18:288–92.
- Malik NS, Chernbumroong S, Xu Y, *et al*. The BCD triage sieve outperforms all existing major incident triage tools: comparative analysis using the UK national trauma registry population. *EclinicalMedicine* 2021;36:100888.
- Vassallo J, Smith J. Major incident triage and the evaluation of the triage sort as a secondary triage method. *Emerg Med J* 2019;36:281–6.
- Trauma ACoSo. *Resources for optimal care of the injured patient*. 2014.
- Sammy I, Lecky F, Sutton A, *et al*. Factors affecting mortality in older trauma patients—a systematic review and meta-analysis. *Injury* 2016;47:1170–83.
- Klug M, Barash Y, Bechler S, *et al*. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. *J Gen Intern Med* 2020;35:220–7.
- Ramlakhan S, Saatchi R, Sabir L, *et al*. Understanding and interpreting artificial intelligence, machine learning and deep learning in emergency medicine. *Emerg Med J* 2022;39:380–5.
- Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:735–6.
- The University of Manchester. The trauma audit and research network. n.d. Available: www.tarn.ac.uk
- Lerner EB, McKee CH, Cady CE, *et al*. A consensus-based gold standard for the evaluation of mass casualty triage systems. *Prehosp Emerg Care* 2015;19:267–71.
- Model uniform core criteria for mass casualty triage. *Disaster Med Public Health Prep* 2011;5:125–8.
- Bazarian JJ, Eirich MA, Salhanick SD. The relationship between pre-hospital and emergency department Glasgow coma scale scores. *Brain Inj* 2003;17:553–60.
- Cobas MA, De la Peña MA, Manning R, *et al*. Prehospital intubations and mortality: a level 1 trauma center perspective. *Anesth Analg* 2009;109:489–93.
- Jost E, Roberts DJ, Penney T, *et al*. Accuracy of clinical, laboratory, and computed tomography findings for identifying hollow viscus injury in blunt trauma patients with unexplained intraperitoneal free fluid without solid organ injury. *Am J Surg* 2017;213:874–80.
- Domeier RM, Evans RW, Swor RA, *et al*. The reliability of prehospital clinical evaluation for potential spinal injury is not affected by the mechanism of injury. *Prehosp Emerg Care* 1999;3:332–7.
- Vandromme MJ, Griffin RL, Kerby JD, *et al*. Identifying risk for massive transfusion in the relatively normotensive patient: utility of the prehospital shock index. *J Trauma* 2011;70:384–90.
- Raita Y, Goto T, Faridi MK, *et al*. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019;23:64.
- Soltan AAS, Kouchaki S, Zhu T, *et al*. Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. *Lancet Digit Health* 2021;3:e78–87.
- Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature Attribution for tree ensembles. *arXiv Preprint arXiv* 2018:180203888.
- Hettiarachy S, Tai N, Mahoney P, *et al*. UK's NHS trauma systems: lessons from military experience. *The Lancet* 2010;376:149–51.
- Robin X, Turck N, Hainard A, *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77.
- citizenAID A UK charity empowering the public to save lives. Available: <https://www.citizenaid.org/> [Accessed 21 Jan 2022].
- Wallis LA, Carley S. Comparison of paediatric major incident primary triage tools. *Emerg Med J* 2006;23:475–8.
- NHS England Emergency Preparedness RaR. *Concept of operations for managing mass casualties*. Emergency Preparedness RaR, 2017.
- Rodriguez RM, Anglin D, Langdorf MI, *et al*. NEXUS chest: validation of a decision instrument for selective chest imaging in blunt trauma. *JAMA Surg* 2013;148:940–6.

Triage in major incidents: development and external validation of novel machine-learning derived primary and secondary triage tools

Supplementary material

Additional details of machine learning modelling

An overview of study methodology and data processing is presented in Supplementary Figure 1, with a more detailed model development and selection strategy outlined in Supplementary Figure 2.

For the decision tree (also known as Recursive Partitioning And Regression Tree, RPART) method, a limit of a maximum tree depth of 3 was imposed for ease of interpretation. To guard against overfitting, we chose to tune the cost complexity and the tree depth parameter of the decision tree model. Effectively, the tree depth (distance from the root to a terminal node) represents the number of measurables needed in order to determine a triage category. However, we note that unlike triage tools conceived by human experts, it is possible to have the same variable used more than once to split the nodes, if the reuse of variables reduces classification error. A deep tree with many splits tends to overfit the data, and makes it difficult to adapt the model to a tool that can be implemented in practice.

Both random forest (RF) and gradient boosted tree (XGB) are popular machine learning algorithms with strong predictive power. RF is based on averaging an ensemble of trees and the idea of bagging, which lowers the prediction variance. Furthermore, instead of growing each tree using all variables, it randomly chooses a subset of variables at each split of the node in the tree, thereby forcing it to learn through all subsets of available variables. For XGB, the prediction target is estimated by sum-of-trees, and the model is built by successively fitting each tree to the residue of previously fitted trees with no single tree dominating the prediction, while regularizing the fit through multiplication by a scaling factor known as learning rate. In short, XGB estimates the target function by a sum of trees each of which explains a small and different portion of the target and no single tree dominates the prediction.

For the L1-regularized logistic regression model, the penalty parameter, specifying the amount of regularization, was tuned. We add a regularization term in logistic regression so that the solution is well-defined even if the data are perfectly linearly separable.

Initially, models were trained using all 13 input variables (summarised in Supplementary Table 1): the resulting models would be too complex for practical application as tools, but nonetheless act as a useful comparator for model performance (see Supplementary Figure 2 detailing the model building and selection strategy). The optimal hyperparameters that yield the best AUC were selected. For decision tree and logistic regression, a grid search was used; whereas for RF and XGB, random sampling of points in the parameter space was used to try to cover the space as uniformly as possible. For each model, having selected the hyperparameters, a final model was trained on the whole training set (70% of TARN data) and then evaluated on the remaining 30% hold-out data. Models developed using all 13 input variables yielded similar AUC values (range 0.862-0.868, see Supplementary Table 2), except for the decision tree model (AUC 0.782), which also exhibited lower specificity and higher over-triage than the other ML models. All models employing 13 variables attained sensitivity above 72%, exceeding that of the BCD Triage Sieve. Performance characteristics of models employing all 13 input variables were further evaluated by age subgroup (16-64 years and 65+ years (Supplementary Table 2)). We note that for ML models evaluated on the 65+ group, while there is slight decrease in AUC compared to the 16-64 group, sensitivity is much worse, except for the decision tree model which has the best sensitivity (66.3%) among all models and triage tools. However, the price of this relatively high sensitivity of decision tree is a high over-triage rate (87.2%).

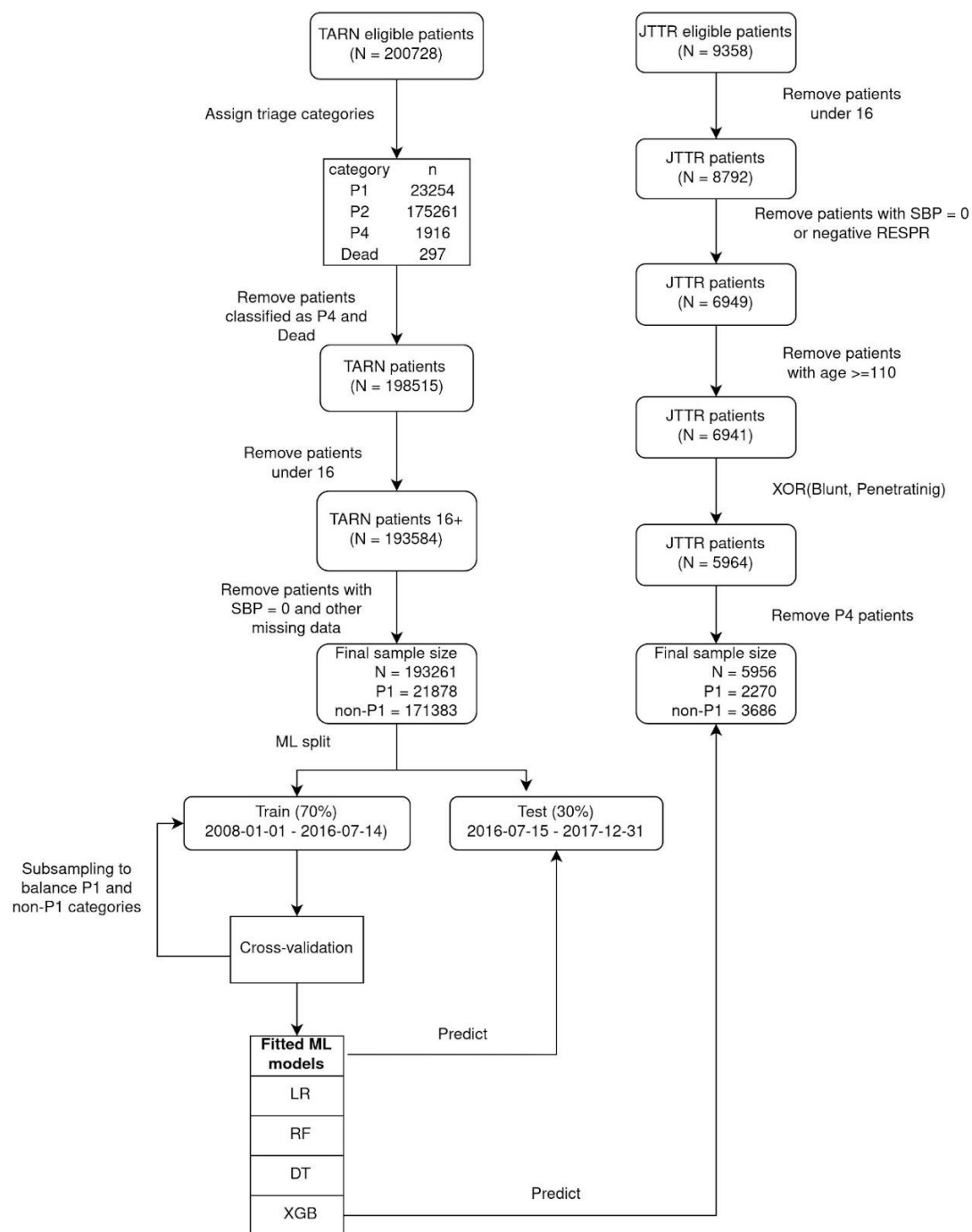
Existing triage tools were applied to the internal validation dataset to act as comparators to the models proposed as novel triage tools. To overcome the over-representation of elders (65+ years) within the TARN database (see Manuscript, Table 1), who also differ in their physiology to younger adults, tool performance was additionally tested in subgroups by age (16-64 years and 65+ years), as shown in Supplementary Table 3. Existing tools demonstrated lower sensitivity and higher over-triage rates amongst elders compared to younger adults (16-64 years).

We sought to combine the individual models in a weighted fashion by training a super model [1], in which weights are assigned to models based on their predictive power and the final predictions are driven by models with high weights. For the super model, a binomial likelihood maximization using the BFGS quasi-Newton optimization method was used, the model was fitted using the “SuperLearner” R package [2]. The weights are

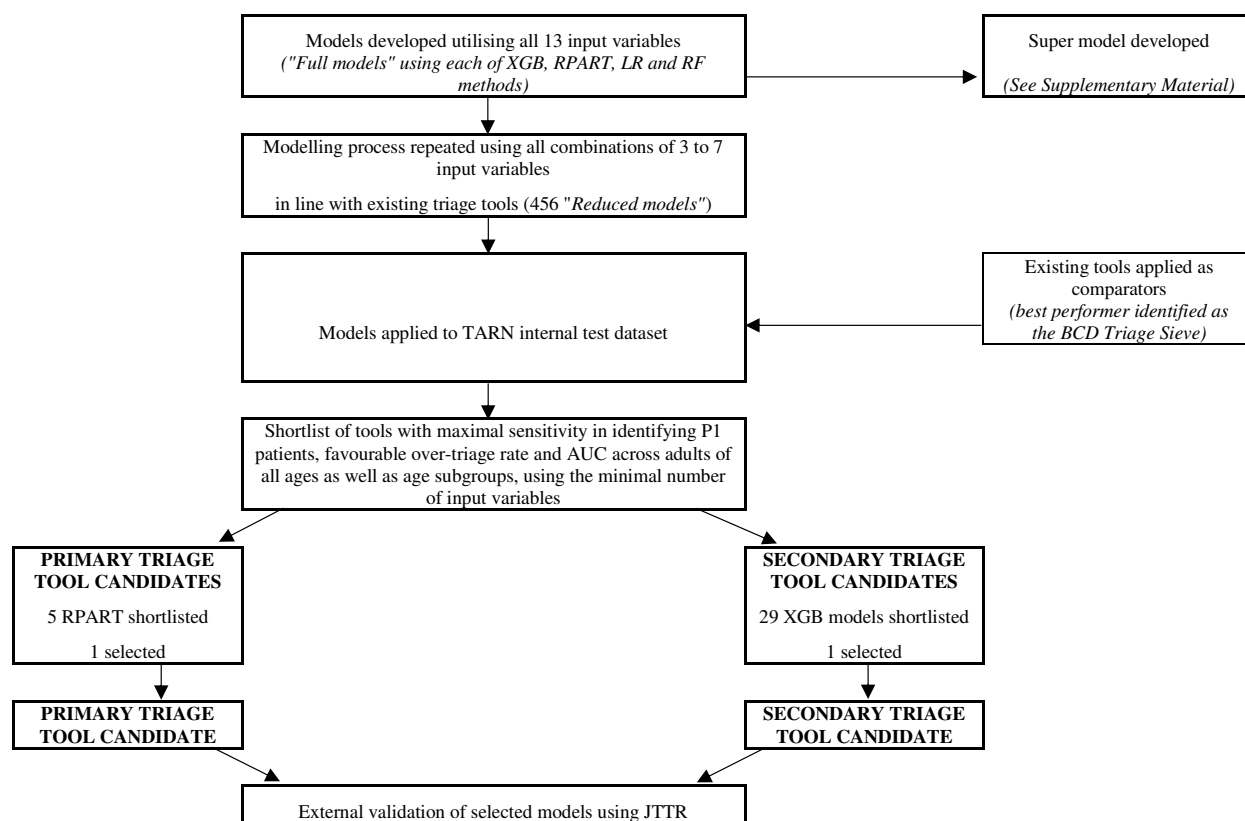
normalized and sum to one. The super model assigned coefficients (weights) to each individual model, along with the minimized risks. We note that the decision tree model was in fact excluded from the super model, since it has a weight of zero (risk 0.525). The XGB model has highest weight of 0.717 (risk 0.442). Random forest had the second highest weight (0.241, risk 0.454) whilst logistic regression had a low contribution to the overall super model (coefficient 0.041, risk 0.451). The AUC for the super model is 0.868.

The importance of individual features (input parameters) was also estimated using the XGB method (see Supplementary Figure 3). This method yielded similar rankings to those generated by the TREEShap method: breathing status contributed 36% of the total gain, followed by presence of chest injury (13%) and GCS verbal score (11%).

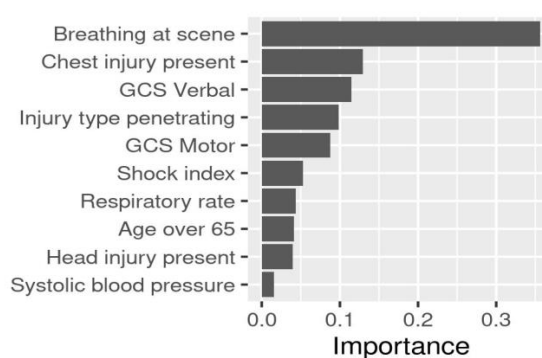
Secondary candidate models were subsequently evaluated on a smaller subset of JTTR patients (n=5455) for which there is complete data available to test the performance of the BCD Triage Sieve, thereby facilitating direct comparison (Supplementary Table 6). The secondary tool candidate (XGB 37) attained comparatively high sensitivity (97.3% vs 80.2%), although this was associated with an 11.1% increase in over-triage (58.5% vs 47.4%).

Supplementary Figure 1: Overview of study methodology and data processing

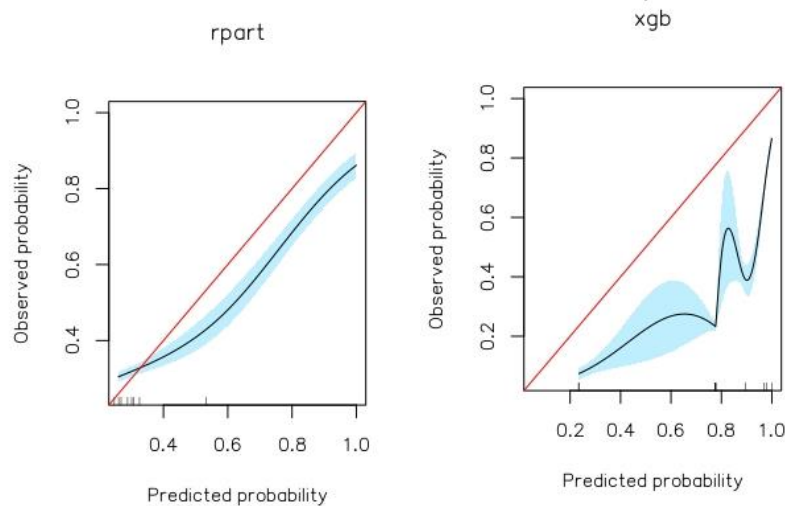
Ledger: Abbreviations: TARN= Trauma Audit and Research Network; JTTR= Joint Theatre Trauma Registry; SBP=Systolic Blood Pressure; XOR=Exclusive/or; LR=Logistic Regression; RF=Random Forest; DT=Decision Tree; XGB= Extreme Gradient Boost; RESPR=Respiratory rate

Supplementary Figure 2: Model building and selection strategy

Ledger: XGB=eXtreme Gradient Boosting, RPART=Recursive Partitioning And Regression Trees (i.e. Decision Tree), LR=Logistic regression and RF=random forest, TARN=Trauma Audit and Research Network Registry, JTTR=Joint Theatre Trauma Registry, BCD Triage Sieve=Battlefield Casualty Drills Triage Sieve.

Supplementary Figure 3: Feature importance plot for the XGB model

Ledger: Importance of top 10 predictors for the XGB model as measured by the fractional contribution of each feature to the model based on the total gain of each feature's splits. High values represent more predictive features. Respiratory rate is measured in breaths per minute, Systolic blood pressure is measured in mmHg. Presence of chest and head injuries are denoted by a positive Abbreviated Injury Severity score.

Supplementary Figure 4: Calibration plot for models selected as candidate primary and secondary triage tools

Ledger: calibration plot for the candidate primary (left) and secondary (right) ML models, evaluated using JTTR data.

The calibration curve was estimated by natural splines using the R package *gbm* [3]. 95% confidence intervals covering 2 standard errors are demonstrated (blue).

For perfect calibration, the calibration curve would align with the 45-degree line (red). It can be seen that the secondary tool (XGB model) over-predicted risk, since the predicted P1 probabilities were greater than the observed probabilities across all patients. This is expected as the secondary tool candidate (XGB model) had high sensitivity but low specificity. In contrast, the calibration curve of the primary tool candidate (decision tree or *rpart* model) was smoother and the over-prediction was less extreme than XGB, reflecting the fact that the decision tree model had lower sensitivity and higher specificity than XGB.

Supplementary Table 1: Clinical parameters included as input variables for modelling

	Input variables
Physiological parameters*	Heart rate (beats per minute), Respiratory rate (breaths per minute) Systolic blood pressure (mmHg) Ability to breathe spontaneously** GCS Verbal component GCS Motor component GCS Eyes component Shock index***
Anatomical parameters	Presence (AIS>0) or absence (AIS=0) of injury in the following anatomical regions: Head Face Thorax Limb
Age	Age 65 and over (Binary – Yes or No)
Injury Mechanism*	Blunt or penetrating injury

Ledger: GCS=Glasgow Coma Score, AIS=Abbreviated Injury Score. *First recorded pre-hospital physiological parameters and injury mechanism were utilised. **All patients who underwent an advanced airway intervention in the pre-hospital environment were assumed to be unable to breathe. ***Shock index=heart rate/systolic blood pressure.

Supplementary Table 2: Performance of machine learning models utilising all 13 input variables in predicting P1 status amongst patients in the internal (TARN) testing dataset

	Method	Sensitivity	Specificity	Under-triage	Over-triage	AUC
All adults (16+ years)	ml_rpart	73.0 [71.8, 74.2]	73.9 [73.5, 74.3]	27.0 [25.8, 28.2]	77.0 [76.4, 77.7]	0.782 [0.775, 0.789]
	ml_rf	72.6 [71.4, 73.8]	86.0 [85.7, 86.3]	27.4 [26.2, 28.6]	64.5 [63.6, 65.4]	0.867 [0.861, 0.873]
	ml_xgb	72.7 [71.5, 73.9]	85.9 [85.6, 86.2]	27.3 [26.1, 28.5]	64.6 [63.7, 65.5]	0.868 [0.862, 0.874]
	ml_lr	72.2 [71.0, 73.4]	85.2 [84.9, 85.5]	27.8 [26.6, 29.0]	65.8 [64.9, 66.6]	0.862 [0.857, 0.868]
16-64 years subgroup	ml_rpart	76.0 [74.6, 77.3]	71.8 [71.2, 72.4]	24.0 [22.7, 25.4]	66.8 [65.8, 67.8]	0.794 [0.786, 0.803]
	ml_rf	81.9 [80.7, 83.1]	76.1 [75.5, 76.7]	18.1 [16.9, 19.3]	61.2 [60.2, 62.3]	0.877 [0.871, 0.884]
	ml_xgb	82.3 [81.0, 83.5]	75.7 [75.2, 76.3]	17.7 [16.5, 19.0]	61.5 [60.4, 62.6]	0.879 [0.872, 0.885]
	ml_lr	82.5 [81.2, 83.7]	74.7 [74.1, 75.3]	17.5 [16.3, 18.8]	62.5 [61.4, 63.5]	0.873 [0.866, 0.879]
65+ years subgroup	ml_rpart	66.3 [64.0, 68.6]	75.3 [74.8, 75.8]	33.7 [31.4, 36.0]	87.2 [86.5, 87.9]	0.746 [0.733, 0.759]
	ml_rf	51.7 [49.3, 54.1]	92.5 [92.2, 92.8]	48.3 [45.9, 50.7]	72.7 [71.1, 74.2]	0.806 [0.793, 0.818]
	ml_xgb	51.3 [48.9, 53.6]	92.6 [92.3, 92.9]	48.7 [46.4, 51.1]	72.5 [70.9, 74.0]	0.807 [0.795, 0.820]
	ml_lr	49.2 [46.8, 51.5]	92.2 [91.9, 92.5]	50.8 [48.5, 53.2]	74.3 [72.8, 75.8]	0.800 [0.787, 0.812]

Ledger: Results shown are percentages (except for AUC). The best performing model amongst all adults for each method is shown, including performance by age subgroup. Abbreviations: ml=machine learning, rpart=decision tree, rf=random forest, xgb= extreme gradient boosting, lr=logistic regression.

Supplementary Table 3: Performance characteristics of existing triage tools when applied to the internal validation dataset

Tool	Sensitivity	Specificity	Under-triage	Over-triage	AUC
All adults (16+ years)					
BCD Triage Sieve	68.2 [66.9, 69.4]	69.5 [69.1, 69.9]	31.8 [30.6, 33.1]	80.8 [80.2, 81.3]	0.688 [0.682, 0.695]
CareFlight	39.9 [38.6, 41.2]	94.5 [94.3, 94.7]	60.1 [58.8, 61.4]	56.4 [55.0, 57.8]	0.672 [0.666, 0.679]
JumpSTART	42.5 [41.2, 43.8]	92.1 [91.8, 92.3]	57.5 [56.2, 58.8]	63.7 [62.5, 64.9]	0.673 [0.666, 0.679]
MIMMS Triage Sieve	40.5 [39.2, 41.8]	92.0 [91.8, 92.3]	59.5 [58.2, 60.8]	64.9 [63.7, 66.1]	0.663 [0.656, 0.669]
MPTT	50.5 [49.2, 51.8]	62.4 [62.0, 62.8]	49.5 [48.2, 50.8]	87.5 [87.1, 87.9]	0.565 [0.558, 0.571]
MPTT-24	48.4 [47.1, 49.7]	66.4 [66.0, 66.8]	51.6 [50.3, 52.9]	86.7 [86.2, 87.2]	0.574 [0.567, 0.581]
MSTART	54.9 [53.6, 56.2]	88.4 [88.1, 88.7]	45.1 [43.8, 46.4]	66.5 [65.5, 67.5]	0.717 [0.710, 0.723]
NARU Triage Sieve	43.0 [41.7, 44.3]	88.3 [88.1, 88.6]	57.0 [55.7, 58.3]	71.8 [70.9, 72.8]	0.657 [0.650, 0.663]
RAMP	37.1 [35.9, 38.4]	94.6 [94.5, 94.8]	62.9 [61.6, 64.1]	57.5 [56.1, 58.9]	0.659 [0.653, 0.665]
START	51.8 [50.5, 53.2]	90.0 [89.7, 90.2]	48.2 [46.8, 49.5]	64.5 [63.5, 65.6]	0.709 [0.702, 0.716]
16-64 years subgroup					
BCD Triage Sieve	72.7 [71.2, 74.1]	64.8 [64.2, 65.5]	27.3 [25.9, 28.8]	72.4 [71.5, 73.3]	0.687 [0.680, 0.695]
CareFlight	42.7 [41.1, 44.3]	94.3 [94.0, 94.6]	57.3 [55.7, 58.9]	42.0 [40.2, 43.8]	0.685 [0.677, 0.693]
JumpSTART	45.5 [43.9, 47.0]	91.2 [90.9, 91.6]	54.5 [53.0, 56.1]	51.1 [49.4, 52.7]	0.684 [0.675, 0.692]
MIMMS Triage Sieve	43.0 [41.4, 44.6]	92.6 [92.2, 93.0]	57.0 [55.4, 58.6]	48.2 [46.5, 50.0]	0.678 [0.670, 0.686]
MPTT	52.3 [50.7, 53.9]	57.1 [56.4, 57.8]	47.7 [46.1, 49.3]	81.6 [80.9, 82.4]	0.547 [0.538, 0.555]
MPTT-24	50.5 [48.9, 52.1]	61.6 [60.9, 62.2]	49.5 [47.9, 51.1]	80.5 [79.7, 81.3]	0.560 [0.552, 0.569]
MSTART	57.6 [56.1, 59.2]	88.9 [88.5, 89.3]	42.4 [40.8, 43.9]	51.0 [49.6, 52.5]	0.733 [0.725, 0.741]
NARU Triage Sieve	47.1 [45.5, 48.7]	87.5 [87.0, 87.9]	52.9 [51.3, 54.5]	59.0 [57.6, 60.5]	0.673 [0.665, 0.681]
RAMP	39.6 [38.1, 41.2]	94.4 [94.1, 94.7]	60.4 [58.8, 61.9]	43.4 [41.5, 45.3]	0.670 [0.662, 0.678]
START	54.4 [52.8, 55.9]	90.7 [90.3, 91.1]	45.6 [44.1, 47.2]	48.1 [46.5, 49.6]	0.725 [0.717, 0.733]
65+ years subgroup					
BCD Triage Sieve	58.1 [55.7, 60.4]	72.6 [72.1, 73.1]	41.9 [39.6, 44.3]	89.6 [89.0, 90.2]	0.653 [0.642, 0.665]
CareFlight	33.7 [31.4, 36.0]	94.6 [94.4, 94.9]	66.3 [64.0, 68.6]	74.5 [72.6, 76.3]	0.642 [0.630, 0.653]
JumpSTART	35.8 [33.5, 38.1]	92.6 [92.3, 92.9]	64.2 [61.9, 66.5]	79.1 [77.6, 80.6]	0.642 [0.630, 0.653]
MIMMS Triage Sieve	34.9 [32.6, 37.2]	91.7 [91.3, 92.0]	65.1 [62.8, 67.4]	81.4 [80.0, 82.8]	0.633 [0.621, 0.644]
MPTT	46.5 [44.1, 48.9]	65.9 [65.4, 66.5]	53.5 [51.1, 55.9]	93.1 [92.6, 93.5]	0.562 [0.550, 0.574]
MPTT-24	43.7 [41.3, 46.1]	69.6 [69.1, 70.1]	56.3 [53.9, 58.7]	92.7 [92.2, 93.2]	0.567 [0.555, 0.579]
MSTART	48.9 [46.5, 51.3]	88.0 [87.7, 88.4]	51.1 [48.7, 53.5]	81.8 [80.6, 82.9]	0.685 [0.673, 0.697]
NARU Triage Sieve	33.7 [31.5, 36.0]	88.9 [88.6, 89.3]	66.3 [64.0, 68.5]	85.8 [84.7, 86.8]	0.613 [0.602, 0.625]
RAMP	31.6 [29.4, 33.8]	94.8 [94.6, 95.1]	68.4 [66.2, 70.6]	75.1 [73.2, 76.9]	0.632 [0.621, 0.643]
START	46.2 [43.8, 48.6]	89.5 [89.1, 89.8]	53.8 [51.4, 56.2]	80.7 [79.5, 81.9]	0.678 [0.666, 0.690]

Ledger: BCD Triage Sieve=Battlefield Casualty Drills Triage Sieve (UK Military), CareFlight (Australia), JumpSTART (US paediatric triage tool), MIMMS Triage Sieve=Major Incident Medical Management and Support Triage Sieve, MPTT=Modified Physiological Triage Tool (tool modelled in UK military casualties), MPTT-24 (modification of MPTT, 2017), START=Simple Triage and Rapid Treatment (US adult tool), MSTART=modified START, NARU Triage Sieve=National Ambulance Resilience Unit Triage Sieve (Current UK civilian tool), RAMP=Rapid Assessment of Mentation and Pulse (New York Fire Department).

Supplementary Tables 4A-C: See landscape format document

Supplementary Table 5: External validation of shortlisted models and the Battlefield Casualty Drills Triage Sieve (comparator) using the Joint Theatre Trauma Registry (n=5455)

Method	Sensitivity	Specificity	Under-triage	Over-triage	AUC
Comparator (best existing tool):					
BCD Triage Sieve	0.802 [0.784, 0.819]	0.578 [0.561, 0.595]	0.198 [0.181, 0.216]	0.474 [0.456, 0.492]	0.690 [0.678, 0.702]
Primary tool candidate models					
rpart_1	0.330 [0.310, 0.351]	0.892 [0.881, 0.902]	0.670 [0.649, 0.690]	0.360 [0.331, 0.390]	0.618 [0.606, 0.629]
rpart_3	0.330 [0.310, 0.351]	0.892 [0.881, 0.902]	0.670 [0.649, 0.690]	0.360 [0.331, 0.390]	0.618 [0.607, 0.630]
rpart_37	0.330 [0.310, 0.351]	0.892 [0.881, 0.902]	0.670 [0.649, 0.690]	0.360 [0.331, 0.390]	0.618 [0.607, 0.630]
rpart_52	0.479 [0.457, 0.501]	0.752 [0.737, 0.766]	0.521 [0.499, 0.543]	0.470 [0.447, 0.494]	0.611 [0.598, 0.624]
rpart_124	0.437 [0.415, 0.459]	0.879 [0.868, 0.890]	0.563 [0.541, 0.585]	0.322 [0.297, 0.348]	0.668 [0.656, 0.680]
Secondary tool candidate models					
xgb_1	0.667 [0.646, 0.688]	0.788 [0.774, 0.802]	0.333 [0.312, 0.354]	0.353 [0.332, 0.374]	0.755 [0.743, 0.768]
xgb_3	0.667 [0.646, 0.688]	0.788 [0.774, 0.802]	0.333 [0.312, 0.354]	0.353 [0.332, 0.374]	0.755 [0.743, 0.768]
xgb_37	0.973 [0.964, 0.979]	0.199 [0.186, 0.213]	0.027 [0.021, 0.036]	0.585 [0.571, 0.599]	0.780 [0.768, 0.792]
xgb_38	0.667 [0.646, 0.688]	0.788 [0.774, 0.802]	0.333 [0.312, 0.354]	0.353 [0.332, 0.374]	0.755 [0.743, 0.768]
xgb_41	0.667 [0.646, 0.688]	0.788 [0.774, 0.802]	0.333 [0.312, 0.354]	0.353 [0.332, 0.374]	0.782 [0.769, 0.795]
xgb_42	0.667 [0.646, 0.688]	0.788 [0.774, 0.802]	0.333 [0.312, 0.354]	0.353 [0.332, 0.374]	0.748 [0.734, 0.761]
xgb_43	0.675 [0.654, 0.695]	0.791 [0.777, 0.805]	0.325 [0.305, 0.346]	0.347 [0.326, 0.367]	0.776 [0.762, 0.790]
xgb_44	0.973 [0.964, 0.979]	0.199 [0.186, 0.213]	0.027 [0.021, 0.036]	0.585 [0.571, 0.599]	0.780 [0.768, 0.792]
xgb_53	0.667 [0.646, 0.688]	0.788 [0.774, 0.802]	0.333 [0.312, 0.354]	0.353 [0.332, 0.374]	0.748 [0.734, 0.761]
xgb_54	0.675 [0.654, 0.695]	0.791 [0.777, 0.805]	0.325 [0.305, 0.346]	0.347 [0.326, 0.367]	0.777 [0.763, 0.790]
xgb_121	0.973 [0.964, 0.979]	0.199 [0.186, 0.213]	0.027 [0.021, 0.036]	0.585 [0.571, 0.599]	0.780 [0.768, 0.792]
xgb_124	0.975 [0.966, 0.981]	0.174 [0.162, 0.187]	0.025 [0.019, 0.034]	0.592 [0.578, 0.606]	0.777 [0.764, 0.790]
xgb_125	0.973 [0.964, 0.979]	0.199 [0.186, 0.213]	0.027 [0.021, 0.036]	0.585 [0.571, 0.599]	0.778 [0.766, 0.790]
xgb_130	0.667 [0.646, 0.688]	0.788 [0.774, 0.802]	0.333 [0.312, 0.354]	0.353 [0.332, 0.374]	0.748 [0.734, 0.761]
xgb_131	0.674 [0.653, 0.695]	0.793 [0.779, 0.807]	0.326 [0.305, 0.347]	0.344 [0.324, 0.365]	0.774 [0.760, 0.788]
xgb_141	0.676 [0.655, 0.696]	0.792 [0.778, 0.805]	0.324 [0.304, 0.345]	0.346 [0.325, 0.367]	0.768 [0.754, 0.782]
xgb_154	0.965 [0.955, 0.972]	0.171 [0.159, 0.184]	0.035 [0.028, 0.045]	0.596 [0.582, 0.610]	0.695 [0.680, 0.709]
xgb_165	0.684 [0.663, 0.704]	0.792 [0.778, 0.805]	0.316 [0.296, 0.337]	0.343 [0.323, 0.364]	0.788 [0.774, 0.801]
xgb_166	0.676 [0.655, 0.696]	0.791 [0.777, 0.805]	0.324 [0.304, 0.345]	0.346 [0.326, 0.367]	0.770 [0.757, 0.784]
xgb_250	0.973 [0.964, 0.979]	0.199 [0.186, 0.213]	0.027 [0.021, 0.036]	0.585 [0.571, 0.599]	0.781 [0.769, 0.794]
xgb_251	0.972 [0.964, 0.979]	0.203 [0.190, 0.217]	0.028 [0.021, 0.036]	0.584 [0.570, 0.599]	0.792 [0.779, 0.804]
xgb_270	0.722 [0.702, 0.742]	0.701 [0.685, 0.716]	0.278 [0.258, 0.298]	0.416 [0.396, 0.435]	0.785 [0.772, 0.798]
xgb_271	0.670 [0.649, 0.691]	0.800 [0.786, 0.813]	0.330 [0.309, 0.351]	0.338 [0.318, 0.359]	0.770 [0.756, 0.784]
xgb_289	0.971 [0.962, 0.977]	0.195 [0.182, 0.209]	0.029 [0.023, 0.038]	0.587 [0.573, 0.601]	0.778 [0.765, 0.791]
xgb_311	0.721 [0.701, 0.741]	0.704 [0.689, 0.719]	0.279 [0.259, 0.299]	0.413 [0.393, 0.433]	0.781 [0.768, 0.795]
xgb_380	0.971 [0.962, 0.977]	0.196 [0.183, 0.210]	0.029 [0.023, 0.038]	0.587 [0.573, 0.601]	0.795 [0.782, 0.807]
xgb_392	0.976 [0.968, 0.982]	0.174 [0.162, 0.187]	0.024 [0.018, 0.032]	0.592 [0.578, 0.606]	0.800 [0.788, 0.812]
xgb_402	0.709 [0.688, 0.729]	0.732 [0.717, 0.747]	0.291 [0.271, 0.312]	0.394 [0.374, 0.414]	0.783 [0.769, 0.796]
xgb_417	0.976 [0.968, 0.982]	0.181 [0.168, 0.194]	0.024 [0.018, 0.032]	0.590 [0.576, 0.604]	0.799 [0.787, 0.812]

Ledger: *this is a reduced JTTR dataset for which a complete set of physiological data exists for application of the BCD Triage Sieve (the best performing existing tool, selected as a comparator) can be applied. To allow direct comparison, the models shortlisted as candidates for triage tools are applied to the reduced dataset. rpart=decision tree (Recursive Partitioning And Regression Trees), xgb= extreme gradient boosting.

References:

- [1] Van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. "Super learner." *Statistical applications in genetics and molecular biology* 6.1 (2007).
- [2] Eric Polley, Erin LeDell, Chris Kennedy and Mark van der Laan (2019). SuperLearner: Super Learner Prediction. R package version 2.0-26. <https://CRAN.R-project.org/package=SuperLearner>
- [3] Brandon Greenwell, Bradley Boehmke, Jay Cunningham and GBM Developers (2022). gbm: Generalized Boosted Regression Models. R package version 2.1.8.1. <https://CRAN.R-project.org/package=gbm>].

Supplementary Table 4A: Models shortlisted as candidates for primary and secondary tools: Description of model variables

Task ID	Model variables	Variables
	Candidate primary tools (Decision Tree (RPART) models)	
1	Breathing status at scene Chest injury present GCS Motor Score	3
3	Breathing status at scene Chest injury present GCS Verbal Score	3
37	Breathing status at scene Chest injury present GCS Motor Score Injury type	4
52	Breathing status at scene Chest injury present GCS Verbal Score Respiratory rate	4
124	Breathing status at scene Chest injury present GCS Motor Score Injury type Respiratory rate	5
	Candidate secondary tools (Extreme Gradient Boosting (XGB) models)	
1	Breathing status at scene Chest injury present GCS Motor Score	3
3	Breathing status at scene Chest injury present GCS Verbal Score	3
37	Breathing status at scene Chest injury present GCS Motor Score Injury type	4
38	Breathing status at scene Chest injury present GCS Motor Score GCS Verbal Score	4
41	Breathing status at scene Chest injury present GCS Motor Score Respiratory rate	4
42	Breathing status at scene Chest injury present GCS Motor Score Head injury present	4
43	Breathing status at scene Chest injury present GCS Motor Score Systolic blood pressure	4
44	Breathing status at scene Chest injury present Injury type GCS Verbal Score	4
53	Breathing status at scene Chest injury present GCS Verbal Score Head injury present	4
54	Breathing status at scene Chest injury present GCS Verbal Score Systolic blood pressure	4
121	Breathing status at scene Chest injury present GCS Motor Score Injury type GCS Verbal Score	5
124	Breathing status at scene Chest injury present GCS Motor Score Injury type Respiratory rate	5
125	Breathing status at scene Chest injury present GCS Motor Score Injury type Head injury present	5
130	Breathing status at scene Chest injury present GCS Motor Score GCS Verbal Score Head injury present	5
131	Breathing status at scene Chest injury present GCS Motor Score GCS Verbal Score Systolic blood pressure	5
141	Breathing status at scene Chest injury present GCS Motor Score Head injury present Systolic blood pressure	5
154	Breathing status at scene Chest injury present Injury type Respiratory rate Head injury present	5
165	Breathing status at scene Chest injury present GCS Verbal Score Respiratory rate Systolic blood pressure	5
166	Breathing status at scene Chest injury present GCS Verbal Score Head injury present Systolic blood pressure	5
250	Breathing status at scene Chest injury present GCS Motor Score Injury type GCS Verbal Score Head injury present	6
251	Breathing status at scene Chest injury present GCS Motor Score Injury type GCS Verbal Score Systolic blood pressure	6
270	Breathing status at scene Chest injury present GCS Motor Score GCS Verbal Score Respiratory rate Systolic blood pressure	6
271	Breathing status at scene Chest injury present GCS Motor Score GCS Verbal Score Head injury present Systolic blood pressure	6
289	Breathing status at scene Chest injury present Injury type GCS Verbal Score Respiratory rate Head injury present	6
311	Breathing status at scene Chest injury present GCS Verbal Score Respiratory rate Head injury present Systolic blood pressure	6
380	Breathing status at scene Chest injury present GCS Motor Score Injury type GCS Verbal Score Respiratory rate Head injury present	7
392	Breathing status at scene Chest injury present GCS Motor Score Injury type Respiratory rate Head injury present Systolic blood pressure	7
402	Breathing status at scene Chest injury present GCS Motor Score GCS Verbal Score Respiratory rate Head injury present Systolic blood pressure	7
417	Breathing status at scene Chest injury present Injury type GCS Verbal Score Respiratory rate Head injury present Systolic blood pressure	7

Supplementary Table 4B: Performance characteristics of models shortlisted as primary and secondary tool candidates

Task ID	Internal validation using TARN testing dataset (all adult patients)					External validation in JTTR (n=5956)														
	Sensitivity		Specificity		Under-triage	Over-triage	AUC		Sensitivity	Specificity	Under-triage	Over-triage	AUC							
RPART:																				
1	73.0	[71.8, 74.2]	73.9	[73.5, 74.3]	27.0	[25.8, 28.2]	77.0	[76.4, 77.7]	0.772	[0.765, 0.779]	34.0	[32.1, 36.0]	89.0	[87.9, 90.0]	66.0	[64.0, 67.9]	34.4	[31.7, 37.2]	0.622	[0.611, 0.633]
3	72.3	[71.1, 73.5]	74.5	[74.1, 74.8]	27.7	[26.5, 28.9]	76.9	[76.2, 77.5]	0.775	[0.768, 0.782]	34.0	[32.1, 36.0]	89.0	[87.9, 90.0]	66.0	[64.0, 67.9]	34.4	[31.7, 37.2]	0.623	[0.612, 0.634]
37	73.0	[71.8, 74.2]	73.9	[73.5, 74.3]	27.0	[25.8, 28.2]	77.0	[76.4, 77.7]	0.782	[0.775, 0.789]	49.8	[47.7, 51.9]	74.5	[73.0, 75.9]	50.2	[48.1, 52.3]	45.4	[43.3, 47.6]	0.616	[0.604, 0.629]
52	72.3	[71.1, 73.5]	74.5	[74.1, 74.8]	27.7	[26.5, 28.9]	76.9	[76.2, 77.5]	0.780	[0.773, 0.787]	34.0	[32.1, 36.0]	89.0	[87.9, 90.0]	66.0	[64.0, 67.9]	34.4	[31.7, 37.2]	0.623	[0.612, 0.634]
124	73.0	[71.8, 74.2]	73.9	[73.5, 74.3]	27.0	[25.8, 28.2]	77.0	[76.4, 77.7]	0.777	[0.770, 0.783]	46.1	[44.0, 48.2]	86.3	[85.2, 87.4]	53.9	[51.8, 56.0]	32.5	[30.2, 34.9]	0.671	[0.659, 0.683]
XGB:																				
1	73.0	[71.8, 74.2]	73.9	[73.5, 74.3]	27.0	[25.8, 28.2]	77.0	[76.4, 77.7]	0.783	[0.776, 0.790]	70.5	[68.6, 72.4]	73.7	[72.2, 75.1]	29.5	[27.6, 31.4]	37.7	[35.9, 39.6]	0.754	[0.742, 0.766]
3	72.3	[71.1, 73.5]	74.5	[74.1, 74.8]	27.7	[26.5, 28.9]	76.9	[76.2, 77.5]	0.796	[0.789, 0.803]	70.5	[68.6, 72.4]	73.7	[72.2, 75.1]	29.5	[27.6, 31.4]	37.7	[35.9, 39.6]	0.754	[0.742, 0.766]
37	77.9	[76.8, 79.0]	73.1	[72.7, 73.5]	22.1	[21.0, 23.2]	76.4	[75.8, 77.0]	0.817	[0.810, 0.824]	97.6	[96.8, 98.2]	18.6	[17.4, 19.9]	2.4	[1.8, 3.2]	57.5	[56.2, 58.9]	0.778	[0.766, 0.790]
38	73.8	[72.6, 75.0]	73.6	[73.2, 73.9]	26.2	[25.0, 27.4]	77.1	[76.5, 77.7]	0.798	[0.792, 0.805]	70.5	[68.6, 72.4]	73.7	[72.2, 75.1]	29.5	[27.6, 31.4]	37.7	[35.9, 39.6]	0.754	[0.742, 0.766]
41	73.0	[71.8, 74.2]	73.9	[73.5, 74.3]	27.0	[25.8, 28.2]	77.0	[76.4, 77.7]	0.808	[0.801, 0.815]	70.5	[68.6, 72.4]	73.7	[72.2, 75.1]	29.5	[27.6, 31.4]	37.7	[35.9, 39.6]	0.779	[0.767, 0.792]
42	73.0	[71.8, 74.2]	73.9	[73.5, 74.3]	27.0	[25.8, 28.2]	77.0	[76.4, 77.7]	0.798	[0.791, 0.805]	70.5	[68.6, 72.4]	73.7	[72.2, 75.1]	29.5	[27.6, 31.4]	37.7	[35.9, 39.6]	0.749	[0.736, 0.762]
43	70.0	[68.7, 71.2]	78.5	[78.2, 78.9]	30.0	[28.8, 31.3]	74.3	[73.6, 75.0]	0.809	[0.802, 0.816]	71.2	[69.3, 73.1]	74.0	[72.5, 75.4]	28.8	[26.9, 30.7]	37.2	[35.4, 39.1]	0.775	[0.762, 0.788]
44	77.3	[76.2, 78.4]	73.6	[73.3, 74.0]	22.7	[21.6, 23.8]	76.2	[75.6, 76.8]	0.830	[0.824, 0.837]	97.6	[96.8, 98.2]	18.6	[17.4, 19.9]	2.4	[1.8, 3.2]	57.5	[56.2, 58.9]	0.778	[0.766, 0.790]
53	72.3	[71.1, 73.5]	74.5	[74.1, 74.8]	27.7	[26.5, 28.9]	76.9	[76.2, 77.5]	0.802	[0.795, 0.809]	70.5	[68.6, 72.4]	73.7	[72.2, 75.1]	29.5	[27.6, 31.4]	37.7	[35.9, 39.6]	0.749	[0.736, 0.762]
54	70.7	[69.5, 71.9]	77.7	[77.4, 78.1]	29.3	[28.1, 30.5]	74.7	[74.0, 75.4]	0.816	[0.810, 0.823]	71.2	[69.3, 73.1]	74.0	[72.5, 75.4]	28.8	[26.9, 30.7]	37.2	[35.4, 39.1]	0.776	[0.763, 0.789]
121	78.7	[77.6, 79.8]	72.7	[72.4, 73.1]	21.3	[20.2, 22.4]	76.5	[75.9, 77.1]	0.833	[0.827, 0.839]	97.6	[96.8, 98.2]	18.6	[17.4, 19.9]	2.4	[1.8, 3.2]	57.5	[56.2, 58.9]	0.778	[0.766, 0.790]
124	71.8	[70.6, 73.0]	81.9	[81.6, 82.2]	28.2	[27.0, 29.4]	70.3	[69.5, 71.1]	0.834	[0.827, 0.841]	97.8	[97.0, 98.3]	16.3	[15.1, 17.5]	2.2	[1.7, 3.0]	58.2	[56.8, 59.5]	0.775	[0.763, 0.787]
125	77.6	[76.5, 78.7]	73.6	[73.2, 74.0]	22.4	[21.3, 23.5]	76.2	[75.5, 76.8]	0.836	[0.829, 0.842]	97.6	[96.8, 98.2]	18.6	[17.4, 19.9]	2.4	[1.8, 3.2]	57.5	[56.2, 58.9]	0.778	[0.766, 0.790]
130	73.8	[72.6, 74.9]	73.6	[73.2, 74.0]	26.2	[25.1, 27.4]	77.1	[76.5, 77.7]	0.804	[0.797, 0.811]	70.5	[68.6, 72.4]	73.7	[72.2, 75.1]	29.5	[27.6, 31.4]	37.7	[35.9, 39.6]	0.749	[0.736, 0.762]
131	71.3	[70.1, 72.5]	77.7	[77.4, 78.1]	28.7	[27.5, 29.9]	74.6	[73.9, 75.3]	0.819	[0.812, 0.826]	71.2	[69.3, 73.0]	74.2	[72.7, 75.6]	28.8	[27.0, 30.7]	37.1	[35.2, 39.0]	0.771	[0.758, 0.783]
141	71.0	[69.8, 72.2]	77.9	[77.5, 78.2]	29.0	[27.8, 30.2]	74.6	[73.9, 75.2]	0.819	[0.813, 0.826]	71.3	[69.4, 73.2]	74.0	[72.6, 75.4]	28.7	[26.8, 30.6]	37.2	[35.3, 39.1]	0.768	[0.755, 0.781]
154	71.3	[70.1, 72.5]	78.0	[77.7, 78.4]	28.7	[27.5, 29.9]	74.4	[73.7, 75.0]	0.829	[0.823, 0.835]	96.7	[95.9, 97.4]	17.1	[15.9, 18.4]	3.3	[2.6, 4.1]	58.2	[56.8, 59.5]	0.702	[0.688, 0.715]
165	68.7	[67.4, 69.9]	81.7	[81.3, 82.0]	31.3	[30.1, 32.6]	71.5	[70.7, 72.3]	0.826	[0.820, 0.833]	72.0	[70.1, 73.9]	74.0	[72.6, 75.4]	28.0	[26.1, 29.9]	36.9	[35.1, 38.8]	0.784	[0.772, 0.797]
166	71.9	[70.7, 73.1]	76.2	[75.9, 76.6]	28.1	[26.9, 29.3]	75.7	[75.0, 76.3]	0.822	[0.815, 0.828]	71.3	[69.4, 73.2]	74.0	[72.5, 75.4]	28.7	[26.8, 30.6]	37.2	[35.3, 39.1]	0.772	[0.758, 0.785]
250	78.6	[77.5, 79.7]	73.0	[72.6, 73.4]	21.4	[20.3, 22.5]	76.4	[75.7, 77.0]	0.842	[0.835, 0.848]	97.6	[96.8, 98.2]	18.6	[17.4, 19.9]	2.4	[1.8, 3.2]	57.5	[56.2, 58.9]	0.779	[0.767, 0.790]
251	73.5	[72.3, 74.6]	80.1	[79.8, 80.5]	26.5	[25.4, 27.7]	71.8	[71.0, 72.5]	0.845	[0.839, 0.851]	97.5	[96.8, 98.1]	19.0	[17.7, 20.3]	2.5	[1.9, 3.2]	57.4	[56.1, 58.8]	0.788	[0.776, 0.801]
270	69.4	[68.1, 70.6]	81.6	[81.3, 82.0]	30.6	[29.4, 31.9]	71.3	[70.6, 72.1]	0.829	[0.822, 0.835]	75.4	[73.6, 77.2]	65.5	[63.9, 67.0]	24.6	[22.8, 26.4]	42.6	[40.8, 44.4]	0.782	[0.770, 0.794]
271	70.5	[69.3, 71.7]	78.3	[78.0, 78.7]	29.5	[28.3, 30.7]	74.3	[73.6, 75.0]	0.824	[0.817, 0.830]	70.8	[68.9, 72.7]	74.8	[73.4, 76.2]	29.2	[27.3, 31.1]	36.6	[34.7, 38.5]	0.771	[0.758, 0.784]
289	72.1	[70.9, 73.3]	82.6	[82.2, 82.9]	27.9	[26.7, 29.1]	69.4	[68.6, 70.2]	0.850	[0.844, 0.856]	97.4	[96.6, 98.0]	18.2	[17.0, 19.5]	2.6	[2.0, 3.4]	57.7	[56.3, 59.0]	0.779	[0.767, 0.791]
311	69.1	[67.8, 70.3]	81.5	[81.1, 81.8]	30.9	[29.7, 32.2]	71.6	[70.8, 72.4]	0.831	[0.825, 0.838]	75.3	[73.5, 77.1]	65.8	[64.3, 67.4]	24.7	[22.9, 26.5]	42.4	[40.6, 44.2]	0.781	[0.769, 0.794]
380	71.3	[70.1, 72.5]	83.9	[83.6, 84.2]	28.7	[27.5, 29.9]	68.0	[67.1, 68.8]	0.851	[0.845, 0.857]	97.4	[96.6, 98.0]	18.4	[17.1, 19.7]	2.6	[2.0, 3.4]	57.6	[56.3, 59.0]	0.793	[0.782, 0.805]
392	71.4	[70.2, 72.6]	84.5	[84.2, 84.8]	28.6	[27.4, 29.8]	67.1	[66.2, 67.9]	0.854	[0.848, 0.860]	97.9	[97.2, 98.4]	16.3	[15.1, 17.5]	2.1	[1.6, 2.8]	58.1	[56.8, 59.5]	0.798	[0.787, 0.810]
402	69.6	[68.4, 70.8]	81.6	[81.3, 81.9]	30.4	[29.2, 31.6]	71.3	[70.5, 72.0]	0.833	[0.827, 0.840]	74.2	[72.4, 76.0]	68.4	[66.9, 69.9]	25.8	[24.0, 27.6]	40.9	[39.0, 42.7]	0.781	[0.769, 0.794]
417	72.0	[70.8, 73.2]	83.7	[83.4, 84.0]	28.0	[26.8, 29.2]	68.0	[67.2, 68.8]	0.856	[0.850, 0.862]	97.9	[97.2, 98.4]	16.9	[15.7, 18.2]	2.1	[1.6, 2.8]	58.0	[56.6, 59.3]	0.799	[0.787, 0.811]

Ledger: Values shown are percentages (except AUC), accompanied by 95% confidence intervals. rpart=Decision tree, XGB=eXtreme Gradient Boosting

Supplementary Table 4C: Performance characteristics by age subgroup of models shortlisted as tool candidates using the internal validation (TARN) dataset

Task ID	16-64 years										65+ years									
	Sensitivity		Specificity		Under-triage		Over-triage		AUC		Sensitivity		Specificity		Under-triage		Over-triage		AUC	
RPART																				
1	76.0	[74.6, 77.3]	71.8	[71.2, 72.4]	24.0	[22.7, 25.4]	66.8	[65.8, 67.8]	0.783	[0.775, 0.791]	66.3	[64.0, 68.6]	75.3	[74.8, 75.8]	33.7	[31.4, 36.0]	87.2	[86.5, 87.9]	0.738	[0.726, 0.751]
3	75.3	[73.9, 76.6]	72.4	[71.8, 73.0]	24.7	[23.4, 26.1]	66.6	[65.5, 67.5]	0.787	[0.778, 0.795]	65.7	[63.4, 67.9]	75.8	[75.4, 76.3]	34.3	[32.1, 36.6]	87.1	[86.4, 87.8]	0.741	[0.728, 0.754]
37	76.0	[74.6, 77.3]	71.8	[71.2, 72.4]	24.0	[22.7, 25.4]	66.8	[65.8, 67.8]	0.794	[0.786, 0.803]	66.3	[64.0, 68.6]	75.3	[74.8, 75.8]	33.7	[31.4, 36.0]	87.2	[86.5, 87.9]	0.746	[0.733, 0.759]
52	75.3	[73.9, 76.6]	72.4	[71.8, 73.0]	24.7	[23.4, 26.1]	66.6	[65.5, 67.5]	0.792	[0.784, 0.800]	65.7	[63.4, 67.9]	75.8	[75.4, 76.3]	34.3	[32.1, 36.6]	87.1	[86.4, 87.8]	0.745	[0.732, 0.758]
124	76.0	[74.6, 77.3]	71.8	[71.2, 72.4]	24.0	[22.7, 25.4]	66.8	[65.8, 67.8]	0.788	[0.780, 0.796]	66.3	[64.0, 68.6]	75.3	[74.8, 75.8]	33.7	[31.4, 36.0]	87.2	[86.5, 87.9]	0.742	[0.729, 0.754]
XGB																				
1	76.0	[74.6, 77.3]	71.8	[71.2, 72.4]	24.0	[22.7, 25.4]	66.8	[65.8, 67.8]	0.795	[0.787, 0.804]	66.3	[64.0, 68.6]	75.3	[74.8, 75.8]	33.7	[31.4, 36.0]	87.2	[86.5, 87.9]	0.747	[0.734, 0.759]
3	75.3	[73.9, 76.6]	72.4	[71.8, 73.0]	24.7	[23.4, 26.1]	66.6	[65.5, 67.5]	0.809	[0.801, 0.817]	65.7	[63.4, 67.9]	75.8	[75.4, 76.3]	34.3	[32.1, 36.6]	87.1	[86.4, 87.8]	0.763	[0.750, 0.776]
37	82.6	[81.4, 83.8]	70.1	[69.4, 70.7]	17.4	[16.2, 18.6]	66.3	[65.3, 67.2]	0.839	[0.831, 0.847]	67.3	[65.0, 69.5]	75.1	[74.6, 75.6]	32.7	[30.5, 35.0]	87.1	[86.4, 87.8]	0.752	[0.739, 0.765]
38	76.8	[75.4, 78.1]	71.5	[70.9, 72.2]	23.2	[21.9, 24.6]	66.8	[65.8, 67.7]	0.811	[0.803, 0.819]	67.2	[64.9, 69.4]	74.9	[74.4, 75.4]	32.8	[30.6, 35.1]	87.3	[86.6, 88.0]	0.765	[0.752, 0.778]
41	76.0	[74.6, 77.3]	71.8	[71.2, 72.4]	24.0	[22.7, 25.4]	66.8	[65.8, 67.8]	0.821	[0.813, 0.829]	66.3	[64.0, 68.6]	75.3	[74.8, 75.8]	33.7	[31.4, 36.0]	87.2	[86.5, 87.9]	0.766	[0.752, 0.779]
42	76.0	[74.6, 77.3]	71.8	[71.2, 72.4]	24.0	[22.7, 25.4]	66.8	[65.8, 67.8]	0.806	[0.798, 0.815]	66.3	[64.0, 68.6]	75.3	[74.8, 75.8]	33.7	[31.4, 36.0]	87.2	[86.5, 87.9]	0.778	[0.766, 0.791]
43	74.4	[73.0, 75.7]	74.0	[73.4, 74.6]	25.6	[24.3, 27.0]	65.4	[64.4, 66.4]	0.815	[0.807, 0.824]	60.0	[57.7, 62.4]	81.5	[81.1, 81.9]	40.0	[37.6, 42.3]	85.0	[84.1, 85.8]	0.760	[0.746, 0.774]
44	82.0	[80.8, 83.2]	70.6	[70.0, 71.2]	18.0	[16.8, 19.2]	66.0	[65.0, 67.0]	0.852	[0.845, 0.859]	66.7	[64.4, 68.9]	75.7	[75.2, 76.1]	33.3	[31.1, 35.6]	87.0	[86.3, 87.7]	0.769	[0.756, 0.781]
53	75.3	[73.9, 76.6]	72.4	[71.8, 73.0]	24.7	[23.4, 26.1]	66.6	[65.5, 67.5]	0.813	[0.805, 0.821]	65.7	[63.4, 67.9]	75.8	[75.4, 76.3]	34.3	[32.1, 36.6]	87.1	[86.4, 87.8]	0.780	[0.767, 0.792]
54	74.9	[73.5, 76.3]	73.6	[73.0, 74.2]	25.1	[23.7, 26.5]	65.6	[64.6, 66.7]	0.825	[0.817, 0.833]	61.2	[58.9, 63.5]	80.5	[80.1, 80.9]	38.8	[36.5, 41.1]	85.4	[84.5, 86.2]	0.773	[0.760, 0.786]
121	83.4	[82.2, 84.5]	69.8	[69.2, 70.4]	16.6	[15.5, 17.8]	66.2	[65.3, 67.2]	0.855	[0.848, 0.862]	68.1	[65.9, 70.3]	74.7	[74.2, 75.2]	31.9	[29.7, 34.1]	87.2	[86.5, 87.9]	0.770	[0.758, 0.783]
124	77.5	[76.1, 78.8]	78.8	[78.2, 79.3]	22.5	[21.2, 23.9]	59.7	[58.6, 60.8]	0.854	[0.847, 0.862]	59.1	[56.7, 61.4]	84.0	[83.6, 84.4]	40.9	[38.6, 43.3]	83.3	[82.3, 84.2]	0.772	[0.758, 0.785]
125	82.3	[81.1, 83.5]	70.5	[69.9, 71.2]	17.7	[16.5, 18.9]	66.0	[65.0, 66.9]	0.854	[0.846, 0.861]	67.2	[64.9, 69.4]	75.7	[75.2, 76.1]	32.8	[30.6, 35.1]	86.9	[86.2, 87.6]	0.784	[0.772, 0.797]
130	76.6	[75.3, 77.9]	71.6	[71.0, 72.2]	23.4	[22.1, 24.7]	66.7	[65.7, 67.7]	0.814	[0.806, 0.822]	67.3	[65.0, 69.5]	74.9	[74.4, 75.4]	32.7	[30.5, 35.0]	87.2	[86.5, 87.9]	0.783	[0.771, 0.795]
131	75.6	[74.2, 76.9]	73.7	[73.1, 74.3]	24.4	[23.1, 25.8]	65.4	[64.3, 66.4]	0.828	[0.820, 0.836]	61.7	[59.4, 64.0]	80.4	[80.0, 80.8]	38.3	[36.0, 40.6]	85.3	[84.5, 86.1]	0.775	[0.761, 0.788]
141	74.8	[73.4, 76.1]	73.7	[73.1, 74.3]	25.2	[23.9, 26.6]	65.6	[64.6, 66.6]	0.824	[0.816, 0.833]	62.4	[60.0, 64.7]	80.6	[80.2, 81.1]	37.6	[35.3, 40.0]	85.1	[84.2, 85.9]	0.788	[0.775, 0.800]
154	75.5	[74.1, 76.9]	77.1	[76.5, 77.7]	24.5	[23.1, 25.9]	62.2	[61.1, 63.3]	0.847	[0.839, 0.854]	61.8	[59.4, 64.1]	78.6	[78.2, 79.1]	38.2	[35.9, 40.6]	86.4	[85.6, 87.1]	0.780	[0.768, 0.793]
165	72.9	[71.5, 74.3]	78.3	[77.8, 78.9]	27.1	[25.7, 28.5]	61.7	[60.6, 62.8]	0.838	[0.830, 0.846]	59.2	[56.8, 61.5]	83.9	[83.5, 84.3]	40.8	[38.5, 43.2]	83.3	[82.4, 84.2]	0.781	[0.768, 0.794]
166	75.9	[74.5, 77.2]	72.2	[71.5, 72.8]	24.1	[22.8, 25.5]	66.5	[65.5, 67.5]	0.828	[0.820, 0.836]	62.8	[60.5, 65.1]	78.9	[78.5, 79.4]	37.2	[34.9, 39.5]	86.0	[85.2, 86.8]	0.790	[0.778, 0.803]
250	83.2	[82.0, 84.3]	70.1	[69.4, 70.7]	16.8	[15.7, 18.0]	66.1	[65.1, 67.1]	0.860	[0.853, 0.868]	68.3	[66.0, 70.4]	74.9	[74.4, 75.4]	31.7	[29.6, 34.0]	87.1	[86.4, 87.8]	0.789	[0.777, 0.801]
251	79.9	[78.6, 81.2]	75.2	[74.6, 75.8]	20.1	[18.8, 21.4]	62.7	[61.7, 63.8]	0.863	[0.856, 0.871]	58.9	[56.6, 61.3]	83.4	[83.0, 83.8]	41.1	[38.7, 43.4]	83.8	[82.8, 84.7]	0.783	[0.770, 0.796]
270	73.7	[72.3, 75.1]	78.2	[77.6, 78.7]	26.3	[24.9, 27.7]	61.6	[60.5, 62.7]	0.840	[0.833, 0.848]	59.6	[57.2, 61.9]	83.9	[83.5, 84.3]	40.4	[38.1, 42.8]	83.2	[82.2, 84.1]	0.783	[0.770, 0.796]
271	74.5	[73.1, 75.9]	74.6	[74.0, 75.2]	25.5	[24.1, 26.9]	64.9	[63.8, 65.9]	0.830	[0.822, 0.838]	61.6	[59.2, 63.9]	80.8	[80.3, 81.2]	38.4	[36.1, 40.8]	85.1	[84.3, 85.9]	0.791	[0.779, 0.804]
289	77.5	[76.1, 78.8]	80.2	[79.7, 80.7]	22.5	[21.2, 23.9]	58.1	[56.9, 59.2]	0.869	[0.862, 0.876]	60.2	[57.8, 62.5]	84.1	[83.7, 84.5]	39.8	[37.5, 42.2]	82.9	[81.9, 83.8]	0.797	[0.785, 0.810]
311	73.1	[71.6, 74.5]	78.1	[77.5, 78.6]	26.9	[25.5, 28.4]	61.9	[60.8, 63.0]	0.840	[0.832, 0.847]	60.1	[57.7, 62.4]	83.7	[83.3, 84.1]	39.9	[37.6, 42.3]	83.3	[82.3, 84.2]	0.797	[0.784, 0.809]
380	76.9	[75.5, 78.2]	81.6	[81.0, 82.1]	23.1	[21.8, 24.5]	56.5	[55.3, 57.7]	0.870	[0.863, 0.877]	58.7	[56.3, 61.0]	85.4	[85.1, 85.8]	41.3	[39.0, 43.7]	82.0	[80.9, 83.0]	0.798	[0.786, 0.810]
392	77.2	[75.9, 78.5]	80.8	[80.3, 81.3]	22.8	[21.5, 24.1]	57.4	[56.2, 58.6]	0.869	[0.862, 0.876]	58.3	[55.9, 60.6]	87.0	[86.6, 87.4]	41.7	[39.4, 44.1]	80.3	[79.2, 81.4]	0.801	[0.788, 0.813]
402	73.6	[72.2, 75.0]	78.2	[77.6, 78.7]	26.4	[25.0, 27.8]	61.6	[60.5, 62.8]	0.842	[0.834, 0.849]	60.6	[58.2, 62.9]	83.9	[83.5, 84.3]	39.4	[37.1, 41.8]	83.0	[82.0, 83.9]	0.797	[0.785, 0.810]
417	77.9	[76.5, 79.2]	80.2	[79.7, 80.8]	22.1	[20.8, 23.5]	57.9	[56.7, 59.0]	0.873	[0.866, 0.880]	58.8	[56.4, 61.1]	86.0	[85.6, 86.4]	41.2	[38.9, 43.6]	81.4	[80.3, 82.4]	0.803	[0.790, 0.815]

Ledger: Values shown are percentages (except AUC), accompanied by 95% confidence intervals - rpart=Decision tree, XGB=eXtreme Gradient Boosting