Predicting Opportunities for Improvement in Trauma Using Machine Learning

A Registry based Study

Running title: Predicting Opportunities for Improvement in Trauma

Authors: Jonatan Attergrim[1,2,#,*], Kelvin Szolnoky[3,#], Lovisa Strömmer[4], Olof Brattström[2,5], Gunilla Whilke[2,6], Martin Jacobsson[7], Martin Gerdin Wärnberg[1,2]

[1]Department of Global Public Health, Karolinska Institutet, Stockholm, Sweden.

[2]Perioperative Medicine and Intensive Care, Karolinska University Hospital, Solna, Stockholm, Sweden.

[3]Department of Medical Epidemiology and Biostatistics Karolinska Institutet, Stockholm, Sweden.

[4]Department of Clinical Science, Intervention and Technology, Division of Surgery, Karolinska Institute, Karolinska University Hospital Huddinge, 14186 Stockholm, Sweden.

[5]Department of Physiology and Pharmacology, Karolinska Institute, Stockholm, Sweden.

[6]Trauma and Reparative Medicine, Karolinska University Hospital, Solna, 171 76, Stockholm, Sweden.

[7]Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Huddinge, Sweden.

[#]These authors contributed equally to this work.

[*]**Corresponding author**

Jonatan Attergrim (jonatan.attergrim@ki.se)

K9 Global folkhälsa, K9 GPH Stålsby Lundborg, 171 77 Stockholm, Sweden

Tel: (+46) 7287 31114

Author Contributions: All authors provided constructive criticism of the study manuscript. L.S., O.B. and G.W. collected data. J.A., K.S. and M.G.W. designed and interpreted the study. J.A. and K.S. analyzed the data.

Impact: Our research explores a novel system that predicts opportunities for improvement in trauma patients using machine learning and outperforms the established approach using audit filters. This could allow for reallocation and optimization of review resources as well as a means of possibly identifying new types of opportunities for improvement. The methodology uses state-of-the-art machine learning largely unexplored in a medical context adding to a broader general scientific knowledge reaching outside this project.

"This article has an online data supplement, which is accessible from this issue's table of content online at www.atsjournals.org"

Sources: Supported by the Swedish Society of Medicine, grant number SLS-973387, and by "The Swedish Carnegie Hero Fund." Parts of the results were presented orally and as an abstract at the London Trauma Conference.

Descriptor number: 4.6 ICU Management/Outcome

Manuscript word count: 3226

1 Abstract

1.1 Rationale

The multidisciplinary mortality and morbidity conference is the core of programs that aim to improve the quality of trauma care and is used to identify and address opportunities for improvement based on reviewing patient cases. Current systems rely on audit filters for review selection, a process that is hampered by high frequencies of false positives.

1.2 Objectives

To develop, validate, and compare the performance of different machine learning models for predicting opportunities for improvement.

1.3 Methods

We conducted a registry based study using all patients from the Karolinska university hospital that had been reviewed regarding the presence of opportunity for improvement, a binary consensus decision from the mortality and morbidity conference. We developed eight binary classification models using 45 predictors. Training used an 80%-20% train-test split and 1000 resamples without replacement estimated confidence intervals. Performance (sensitivity, specificity, integrated calibration index, Area under the receiver operating characteristics curve) was also compared to current audit filters.

1.4 Measurements and Main Results

The dataset included 6310 patients where opportunities for improvement were present among 431 (7%) patients. The audit filters (Area under the receiver operating characteristics curve: 0.624) was outperformed by all machine learning models. The best performing model was LightGBM (Area under the receiver operating characteristics curve: 0.789).

1.5 Conclusions

Machine learning models outperform the currently used audit filters and could prove to be valuable additions in the screening for opportunities for improvement. More research is needed on how to increase model performance and how to incorporate these models into trauma quality improvement programs.

Key words: Medical Audit, peer review, artificial intelligence, trauma care quality improvement

Abstract word count: 249

2 Introduction

Trauma is a leading cause of death and disease burden globally for people aged 10 to 49 years (1, 2). Multidisciplinary mortality and morbidity conferences are at the core of programs that aim to improve the quality of trauma care and thereby patient outcomes (3, 4). The mortality and morbidity conferences aim to identify and address opportunities for improvement (OFI) by reviewing patient cases. Conducting this review is associated with providing high-quality trauma care (5), including reduced preventable death rates, complication frequencies, and hospitalization duration (6).

A mortality and morbidity conference is conducted through meetings in which representatives from all disciplines and professions involved in trauma care participate (3). During these meetings the care provided in specific patient cases are discussed and compared with optimal trauma care under optimal treatment conditions (7, 8). OFIs have been associated with failures in inital care (9), specifically airway management, fluid resuscitation, hemorrhage control and chest injury management (10–12).

The review process uses audit filters and/or individual review to select trauma patients from a local registry (13). Trauma audit filters ”represent sentinel events in patient care which are associated with poor outcomes and/or sub-optimal care” (such as delays in performance of key tests or treatments, or unexpected deaths) (3, 14). When such a ”fall out” event occurs , it should ideally trigger review and if appropriate, a correction of a systematic errors and/or individual practitioner feedback (14). The use of trauma audit filters have been associated with high frequencies of false positives, in the range from 24% to 80% in different contexts (10, 11, 15). In order to improve the precision in the selection process, trauma mortality prediction models has been proposed. However, the performance of these models has been poor (15–17) most likely due to the fact that the models were developed to predict mortality and not morbidity or failures in trauma care.

Different machine learning (ML) based prediction models have been used for several outcomes

in trauma care (18), but the use of ML for OFI prediction has never been investigated. Therefore, our aim was to develop, validate, and compare the performance of different machine learning models for predicting OFI.

Some of the results of these studies were presented as an oral presentation at the London Trauma Conference 2022

3 Methods

We conducted a registry-based study using all trauma patients included in both the Karolinska University Hospital trauma registry and trauma care quality database between 2014 and 2021 to compare the performance of supervised ML models and audit filters in their ability to predict OFI. The study was approved by Stockholm Research Ethics Review Board, approval number 2021-02541 and 2021-03531.

3.1 Study Setting and Population

The Karolinska University Hospital in Solna, Stockholm, Sweden, is equivalent to a level 1 trauma center and manages approximatly 1500 acute trauma patients each year (19).

The Karolinska University Hospital trauma registry reports to the Swedish Trauma Registry (19) and includes all patients admitted to the Karolinska University Hospital with trauma team activation, regardless of injury severity score, as well as patients admitted without trauma team activation but found to have an injury severity score (ISS) of more than 9. The registry includes data on vital signs, times, injuries and interventions as well as patient demographics according to the european consensus statement, the utstein template (20). The Karolinska University Hospital trauma care quality database includes data relevant to the mortality and morbidity conferences, including audit filters, identified OFIs, and proposed corrective actions.

The mortality and morbidity conferences at Karolinska University hospital includes all

professions involved in trauma care, surgery, neurosurgery, orthopaedics, anaesthesia and intensive care, nursing, and radiology. The presence of an OFI is a consensus decision from the conference, which also includes appropriate corrective actions. The process is a multistage process with escalating levels of reviews. Mortality is directly escalated to the multidisciplinary conference which in addition to OFI also decides whether the death was preventable or possibly preventable, which is also classified as an OFI. To identify non-mortality poor outcomes the review process was subsequently improved and formalized during the study period. Between the years 2014-2017, trauma patients were individually reviewed by a specialized trauma nurse to identify possible OFIs to be escalated to the multidisciplinary conference, but the process was not formalized. From 2017, the process was formalized; a brief individual first review by a specialized nurse was performed when data was registered in the trauma registry and the trauma quality data base. Audit filters (supplement E1) were applied to the trauma quality data base. All ”fall outs” from audit filters and trauma patients with a possible failure of care identified by the nurse during the first review, were reviewed again in a second review by two specialized nurses. If the second review identified a possible OFI, the trauma patient´s process of care was reviewed in the multidisciplinary morbidity conference to decide upon whether or not there were a OFI in trauma care.”

3.2 Eligibility Criteria

We included all patients screened for OFI from the trauma registry and trauma care quality database between 2014 and 2021. Patients below the age of 15 were excluded due to their clinical pathway differing compared to adults.

3.3 Outcome

The models' outcome is the presence of an OFI, as decided by the mortality and morbidity conference, and defined as a binary variable with the levels “Yes - At least one OFI identified” and “No - No OFI identified.” Preventable or possible preventable deaths were also considered an OFI. The data for the outcome was extracted from the trauma care quality database.

3.4 Predictors

All variables from the trauma registry, and included in the revised Utstein template, were considered as potential predictors, including information from the pre-hospital setting, initial care, and subsequent in-hospital care. This information includes, but is not limited to, initial vital signs, times to and types of procedures and interventions, length and level of care, injuries, injury mechanisms and type as well as standard demographics. Both continuous and categorical predictors existed. The final models were built with 45 predictors. For a complete list of predictors, see Table **??** in the online data supplement.

While recommendations regarding the relationship between sample size and the number of predictors for logistic regression exist (21, 22), the optimal number of predictors for other learners are not well researched. To accurately compare all models, we chose to use all predictors regardless of learner, and used a pragmatic approach where we included all available data, leaving a sample size of 6310 patients.

3.5 Statistical Analysis Methods

All statistical analyses were done using R (23). We ran all analyses on an 80%-20% train-test split 1000 times using random resampling without replacement.

Data preprocessing and imputation. We developed a pre-processor to be used on each resample's data by rescaling continuous features using Yeo-Johnson's (24) power transformation and recoding categorical features using dummy variables via "one-hot encoding." Predictors with near-zero variance were removed. Missing continuous predictors were imputed using the mean of the predictor, while categorical predictors were imputed by introducing an unknown category. If the continuous values for blood pressure or respiratory rate were missing but the corresponding categorical "revised trauma score" values were present, imputation was instead done using the mean of all patients in that corresponding "revised trauma score" category. A missing indicator feature was created for each predictor.

The developed pre-processor was then run separately on the training and test sets for each resample. To balance the training sets, we used the Adaptive Synthetic (ADASYN) (25) algorithm, a method to generate synthetic data and thus upsample the OFI outcomes and obtain a 1:1 ratio between the outcome classes.

Model development. Eight ML models were built using the Tidymodels framework (26). We used logistic regression (LR), random forest (RF) (27), decision tree (DT) (28), support vector machine with a radial basis kernel (SVM) (29), XGBoost (30), LightGBM (31), CatBoost (32), and k-nearest neighbor (k-NN) (33). A short description of each learner is included in the online supplement under section E2. All model parameters were hyper-optimized on the first resample's training data, and the hyperparameters were saved for the following resamples. Hyperparameter optimization was run using a 5-fold validation through a simple grid search of size 30 on all available parameters provided by the Tidymodels framework.

Performance measurements. The performance of the different models, including the audit filters, was assessed and compared in terms of sensitivity, specificity, discrimination, and calibration in the test set. Discrimination was measured using the area under the receiver operating characteristics curve (AUC) and calibration was measured using the integrated calibration index (ICI) (34). ICI was not calculated for the audit filter system due to it not being able to output class probability. We chose the same probability cutoff for the ML mmodels that gave the same sensitivity as the audit filters on the first resample's test data. This cutoff was saved and used for the following resamples. We compared performance between models by calculating differences for each performance metric and model combination. We estimated 95% confidence intervals for all performance metrics as well as differences using the 1000 resamples.

Feature importance. We calculated the feature importance for all models on the first resample by using permutation feature importance (35) on the test set. Four predictors indicating another predictor not being done was incorporated in that predictor. The importance of a feature was thus calculated by taking the average AUC performance when shuffling a

feature's data 5 times and comparing it to the model's performance on non-shuffled data.

Code availability. The code used in this study is publicly available online at https://github.com/JonteAtter/Predicting-OFI-in-trauma under the MIT License.

4 Results

4.1 Participants

Out of the 11864 patients included in the Karolinska trauma registry 6310 patients have been reviewed regarding the presence of OFI using a combination of audit filters and individual review. This system selected 1316 patients to the morbidity conference and 596 patients to the mortality conference. (Figure 1). A total of 35 out of the 596 deaths reviewed at the mortality conference were considered preventable (n=4) or possible preventable (n=31) rendering OFI whereas 561 were non-preventable and without any identified OFI. Out of the 5714 alive patients, 1316 were selected for inclusion in a morbidity conference via the use of audit filters, individual review and via an unknown process (n=2) where OFI was identified in 396 (7%) patient cases.

Out of the 6310 included patients most were male (n=4383, 69%) with a mean age of 45 (SD 21) with an overall mortality of 599 (9%). The most common highest level of care was a general ward (36%).

Among the 431 patients with OFI compared to the 5879 patients without, the mean age was slightly higher (mean 48 vs 45 years old) and most patients were treated at a intensive care unit (35%) compared to general ward (37%). The mean ISS was also higher (mean:19 SD:11 vs mean:12 SD:13) as was the frequency of in hospital intubation (n=74 [17%] vs n=460 [8%]). Patients with OFI had longer times to definitive treatment compared to patients without OFI (median: 143 vs 99 minutes from hospital arrival). The definitive treatment also differed where patients with OFI had more interventions with the biggest difference being radiological interventions (7% vs 1%) compared to those without OFI. See Table 1 for details of selected

patient characteristics.

The variable "Time to normal BE" had the highest frequency of missing data (n=6107) followed by "Pre-hospital Intubation type" (n=5812) and "Emergency department intubation type" (n=5749). For complete details see Table **??** in the online supplement.

The 6310 patients was randomly split into a training set (n=5048, 80%) and a test set (n=1262, 20%) 1000 times. The first resamples training set had 338 (7%) OFI compared to its test set, which had 93 (7%) OFI. The mortality was 10% in the training set compared to 8% in the test set and the percentage of patients treated in the intensive care unit was 22% in the training set compared to 21% in the test set. This first resample was used for hyperoptimization and feature importance calculations. For further characteristics regarding the training and test set for the first resample see Table **??** in the online data supplement.

4.2 Model Specification

Figure 2 shows the calculated, model agnostic, permuted feature importance for the first resample. Overall, "Emergency procedure" was the most important predictor followed by "ISS." The highest feature importance for a single model was "ISS," accounting for 7% of model LR performance.

4.3 Model Performance

We used AUC as an overall performance measure since no model showed superior performance in all aspects where all ML models outperformed the use of audit filters (AUC: 0.624, Sens: 0.932, Spec: 0.932) where the highest AUC was found in LightGBM (AUC: 0.789, sens: 0.913, spec: 0.913, ICI: 0.036). Two models, RF (AUC: 0.788, sens: 0.956, spec: 0.956, ICI: 0.028) and XGB (AUC: 0.784, sens: 0.935, spec: 0.935, ICI: 0.017) showed superior performance in every performance metric compared to audit filters. Se Table 2 for all performance measures for all models as well as their calculated confidence intervals.

The calculated differences, delta values, in performance between each model are showed in

Table E3-E6 in the online data supplement. The largest point estimate difference where calculated between LightGBM and the audit filters för AUC: 0.165 (0.164-0.167), XGBoost and LR for ICI: -0.267 (-0.268–0.266), SVM and k-NN for sens: 0.108 (0.106-0.11) and between LR an k-NN för Spec: 0.107 (0.105-0.109).The width of all CI:s were narrow with values differentiating around a single or a few per mille.

5 Discussion

Our results highlight the possibility of incorporating machine learning into modern trauma care quality programs in hope of identifying and addressing OFI. All developed models outperformed the use of audit filters in terms of AUC. No single ML model outperformed in all metrics but LightGBM showed the highest AUC, followed by random forest. RF and XGB showed higher performance than audit filters in all performance measures. The methodology uses state-of-the-art machine learning largely unexplored in a medical context adding to a broader general scientific knowledge reaching outside this project.

In our study all ML learners showed relatively similar results however newer models showed a small, but significant, performance boost compared to more established learners such as k-NN, SVM or LR. While LR and SVM showed superior sensitivity it comes at the cost of specificity. A systematic review from 2022 on the use of ML for other trauma related predictions found a similar trend (18). They identified 25 studies comparing the performance of LR to other learners in which performance where similar in twelve studies, better for non LR learners in ten, and worse in three.

5.1 Limitations

Opportunity for improvement, while defined as a binary variable, includes a diverse set of outcomes ranging from preventable deaths to bad documentation. The heterogeneity of these outcomes represents a range of clinical events, each probably correlating to different predictors. In addition machine learning models struggle with rare events and despite being a aggregate of

all previously identified errors, the OFI frequency is only 6.83%. Hence, OFI is a considerable predictive challange. Future research should focus on identifying potential subgroups of OFI that proves to be relatively hard to predict and adjust accordingly.

The current screening system for OFI might also introduce bias since the filters would favor the identification of some, but not all, errors. This could skew the training data and subsequent models creating a self-reinforcing cycle in future identifications of OFIs. Fortunately, the current system also allow for patient selection without triggering audit filters reducing this issue. This leakage of OFIs is necessary in order to identify such a heterogeneous outcome as OFI. It is important to note that these models are highly dependent on the training data presented to them. If they were implemented in practice, without individual review and concurrent conference, another form of "leakage" would be needed in order to predict future unknown OFIs.

The selection system at the Karolinska University Hospital is also more likely to identify severe clinical outcomes, including preventable death, and it is also probable that complex patient cases with multiple interventions or longer hospital stays are selected for review at a higher rate. This theory is supported by the fact that "ISS" and "Days in hospital" are predictors in the higher range regarding of overall feature importance. This creates a possible selection bias where the measured OFI may not be representative of all errors or a true indicator of overall trauma care quality, but rather a combined measure of injury severity, previous patient morbidity, and adverse clinical outcomes. It is also likely that the nature of OFI changes over time, partly from general advances in trauma care and partly as a consequence of corrective actions originating from the peer review process itself.

5.2 Interpretation and Implications

While previous research successfully apply some of the learners used in our study to predict mortality and other more homogeneous outcomes (18), the heterogeneous nature of OFI presents a more difficult prediction challenge. It is possible that higher quality data with

reduced missing frequency and higher resolution, eg vital signs series instead of single values, additional predictors and complex machine learning systems such as ensembles are needed for more precise and better performing models. The issue with an ever-changing outcome also creates an issue, as future OFIs, at least in theory, change, which makes training on older data problematic. However, these models are built around basic and easy-to-record predictors that all exist in registries following the Utstein template (Uniform Reporting of Data following Major Trauma), allowing for easy application in other settings and systems. Demanding an unreasonable amount and quality of data would therefore come at the cost of reduced external validity and feasibility in broader settings.

It is important to note that a far from perfect performance is to be expected with a heterogeneous outcome such as OFI, and comparing these models to entire systems using a combination of quantitative screening and several human reviews, including a multidisciplinary review, is unfair and was never the goal. Instead, we strive to facilitate human efforts and this is the first time a viable alternative to audit filters is presented. Applying these models could possibly automate part of the resource intensive individual review that is needed when using audit filters allowing for improved resource allocation and optimization. The use of permuted feature importance also visualizes and highlights potentially new clinical patterns correlated to OFI. The models are also based on novel methodology and state-of-the-art machine learning adding to a broader increasing need reaching outside trauma care quality improvement. Hence, the positives of these models should not be understated.

While future research is needed on how to optimally implement and increase performance of these models, our results highlight the possibility of incorporating machine learning into modern trauma care quality programs in hope of identifying and addressing opportunities for improvement among trauma patients.

6 Acknowledgments

The authors thank Liselott Västerbo for her part in collecting and recording data and screening for OFI. They also thank all professionals taking part in the monthly mortality and morbidity conference.

7 References

1. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, *et al.* Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the global burden of disease study 2017. *The Lancet* 2018;392:1736–1788.

2. Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, *et al.* Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *The Lancet* 2020;396:1204–1222.

3. World Health Organization. *Guidelines for trauma quality improvement programmes.* 2009. at <https://www.who.int/publications/i/item/guidelines-for-trauma-quality-improvement-programmes>.

4. Santana MJ, Stelfox HT. Development and evaluation of evidence-informed quality indicators for adult injury care. *Annals of Surgery* 2014;259:186–192.

5. Stelfox HT, Straus SE, Nathens A, Bobranska-Artiuch B. Evidence for quality indicators to evaluate adult trauma care: A systematic review*: *Critical Care Medicine* 2011;39:846–859.

6. McDermott FT. Trauma Audit And Quality Improvement. *ANZ J Surg* 1994;64:147–154.

7. Kwon AM, Garbett NC, Kloecker GH. Pooled preventable death rates in trauma patients: Meta analysis and systematic review since 1990. *Eur J Trauma Emerg Surg* 2014;40:279–285.

8. Ghorbani P, Falkén M, Riddez L, Sundelöf M, Oldner A, Strömmer L. Clinical review is essential to evaluate 30-day mortality after trauma. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 2014;22:

9. Vioque SM, Kim PK, McMaster J, Gallagher J, Allen SR, Holena DN, *et al.* Classifying errors in preventable and potentially preventable trauma deaths: A 9-year review using the joint commission's standardized methodology. *The American Journal of Surgery* 2014;208:187–194.

10. Sanddal TL, Esposito TJ, Whitney JR, Hartford D, Taillac PP, Mann NC, *et al.* Analysis of preventable trauma deaths and opportunities for trauma care improvement in utah. *Journal of Trauma: Injury, Infection & Critical Care* 2011;70:970–977.

11. Roy N, Kizhakke Veetil D, Khajanchi MU, Kumar V, Solomon H, Kamble J, *et al.* Learning from 2523 trauma deaths in india- opportunities to prevent in-hospital deaths. *BMC Health Serv Res* 2017;17:142.

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

13. Hornor MA, Hoeft C, Nathens AB. Quality benchmarking in trauma: From the NTDB to TQIP. *Curr Trauma Rep* 2018;4:160–169.

14. Evans C, Howes D, Pickett W, Dagnone L. Audit filters for improving processes of care and clinical outcomes in trauma systems. In: Cochrane Injuries Group, editor. *Cochrane Database of Systematic Reviews* 2009;doi:10.1002/14651858.CD007590.pub2.

15. Ghorbani P, Strömmer L. Analysis of preventable deaths and errors in trauma care in a scandinavian trauma level-i centre. *Acta Anaesthesiol Scand* 2018;62:1146–1153.

16. Radke OC, Heim C. Recognizing preventable death. *Anesthesiology Clinics* 2019;37:1–11.

17. Heim C, Cole E, West A, Tai N, Brohi K. Survival prediction algorithms miss significant opportunities for improvement if used for case selection in trauma quality improvement programs. *Injury* 2016;47:1960–1965.

18. Zhang T, Nikouline A, Lightfoot D, Nolan B. Machine learning in the prediction of trauma outcomes: A systematic review. *Annals of Emergency Medicine* 2022;80:440–455.

19. Årsrapporter | SweTrau. at <http://rcsyd.se/swetrau/om-swetrau/arsrapporter>.

20. Ringdal KG, Coats TJ, Lefering R, Di Bartolomeo S, Steen PA, Roise O, *et al.* The utstein template for uniform reporting of data following major trauma: A joint revision by SCANTEM, TARN, DGU-TR and RITG. *Scand J Trauma Resusc Emerg Med* 2008;16:7.

21. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, *et al.* Calculating the sample size required for developing a clinical prediction model. *Bmj* 2020;m441.doi:10.1136/bmj.m441.

22. Smeden M van, Moons KG, Groot JA de, Collins GS, Altman DG, Eijkemans MJ, *et al.* Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res* 2019;28:2455–2474.

23. R Core Team. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2020. at <https://www.R-project.org/>.

24. Yeo I-K. A new family of power transformations to improve normality or symmetry. *Biometrika* 2000;87:954–959.

25. Haibo He, Yang Bai, Garcia EA, Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* Hong Kong, China: Ieee; 2008. p. 1322–1328.doi:10.1109/ijcnn.2008.4633969.

26. Kuhn M, Wickham H. *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.* 2020. at <https://www.tidymodels.org>.

27. Breiman L. Random forests. *Machine Learning* 2001;45:5–32.

28. Belson WA. Matching and prediction on the principle of biological classification. *Applied Statistics* 1959;8:65.

29. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273–297.

30. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* Acm; 2016. p. 785–794.doi:10.1145/2939672.2939785.

31. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al.* LightGBM: A highly efficient gradient boosting decision tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in neural information processing systems* Curran Associates, Inc.; 2017.

32. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: Unbiased boosting with categorical features. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. *Advances in neural information processing systems* Curran Associates, Inc.; 2018.

33. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inform Theory* 1967;13:21–27.

34. Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine* 2019;38:4051–4065.

35. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res* 2019;20:177.

36. Cardosi JD, Shen H, Groner JI, Armstrong M, Xiang H. Machine learning for outcome predictions of patients with trauma during emergency department care. *BMJ Health Care Inform* 2021;28:e100407.

8 Tables

Table 1. Demographic and Clinical Characteristics of patients screened for OFI.

	OFI	No OFI	Overall
	(N=431)	(N=5879)	(N=6310)
Age			
Mean (SD)	48 (21)	45 (21)	45 (21)
Median [Min, Max]	47 [15, 97]	42 [15, 100]	43 [15, 100]
Gender			
Female	114 (26%)	1813 (31%)	1927 (31%)
Male	317 (74%)	4066 (69%)	4383 (69%)
Dead at 30 days			
Yes	34 (8%)	565 (10%)	599 (9%)
No	395 (92%)	5304 (90%)	5699 (90%)
Missing	2 (<1%)	10 (<1%)	12 (<1%)
Highest level of care			
Emergency department	16 (4%)	1212 (21%)	1228 (19%)
General ward	91 (21%)	2163 (37%)	2254 (36%)
Surgical ward	134 (31%)	1083 (18%)	1217 (19%)
Specialist ward/Intermediate ward	38 (9%)	207 (4%)	245 (4%)
Intensive care unit	152 (35%)	1214 (21%)	1366 (22%)
Injury severity score			
Mean (SD)	19 (11)	12 (13)	12 (13)
Median [Min, Max]	17 [0, 75]	9 [0, 75]	9 [0, 75]
Missing	0 (0%)	8 (<1%)	8 (<1%)
Respiratory rate			
Mean (SD)	19 (5)	18 (5)	18 (5)
Median [Min, Max]	18 [0, 40]	18 [0, 60]	18 [0, 60]
Missing	46 (11%)	656 (11%)	702 (11%)
ED GCS			
Mean (SD)	14 (3)	14 (2)	14 (2)
Median [Min, Max]	15 [3, 15]	15 [3, 15]	15 [3, 15]
Missing	44 (10%)	655 (11%)	699 (11%)
ED Systolic Blood Pressure			
Mean (SD)	133 (34)	133 (35)	133 (35)
Median [Min, Max]	135 [0, 237]	135 [0, 285]	135 [0, 285]
Missing	0 (0%)	11 (<1%)	11 (<1%)
Time to first CT			
Mean (SD)	75 (123)	75 (169)	75 (166)
Median [Min, Max]	40 [6, 1339]	34 [0, 7073]	35 [0, 7073]
Missing	38 (9%)	731 (12%)	769 (12%)
Time to definitive treatment			
Mean (SD)	265 (321)	231 (337)	236 (335)
Median [Min, Max]	143 [3, 1420]	99 [0, 2036]	105 [0, 2036]
Missing	186 (43%)	4340 (74%)	4526 (72%)
Intubated			
Inhospital	74 (17%)	460 (8%)	534 (8%)
Not intubated	331 (77%)	4946 (84%)	5277 (84%)
Prehospital	26 (6%)	472 (8%)	498 (8%)
Missing	0 (0%)	1 (<1%)	1 (<1%)
Emergency procedure			
Thoracotomy	8 (2%)	88 (1%)	96 (2%)
Laparotomy	25 (6%)	157 (3%)	182 (3%)
Pelvis Packing	0 (0%)	5 (<1%)	5 (<1%)
Revascularization	11 (3%)	29 (<1%)	40 (1%)
Radiological intervention	32 (7%)	50 (1%)	82 (1%)
Craniotomy	39 (9%)	196 (3%)	235 (4%)
Intracranial pressure measurement	13 (3%)	63 (1%)	76 (1%)
Other	117 (27%)	951 (16%)	1068 (17%)
Missing	186 (43%)	4340 (74%)	4526 (72%)

Definition of abbreviations: OFI = Opportunity for Improvement; ED = Emergency Department; GCS = Glascow Coma Scale.
Time to first CT and Time to definitive treatment: Measured in minutes from arrival at the hospita.

Table 2. ML models and audit filters predictive performance for OFI using AUC, accuracy, ICI, and sensitivity.

	AUC	ICI	Sensitivity	Specitivity	FNR	FPR
CatBoost	0.786 (0.785-0.788)	0.077 (0.076-0.078)	0.925 (0.923-0.926)	0.433 (0.432-0.435)	0.075 (0.074-0.077)	0.567 (0.565-0.568)
DT	0.761 (0.76-0.763)	0.061 (0.06-0.061)	0.954 (0.952-0.956)	0.293 (0.287-0.299)	0.046 (0.044-0.048)	0.707 (0.701-0.713)
k-NN	0.726 (0.724-0.727)	0.241 (0.24-0.241)	0.862 (0.86-0.865)	**0.482 (0.481-0.483)**	0.138 (0.135-0.14)	**0.518 (0.517-0.519)**
LightGBM	**0.789 (0.788-0.791)**	0.036 (0.035-0.036)	0.913 (0.911-0.915)	0.462 (0.461-0.463)	0.087 (0.085-0.089)	0.538 (0.537-0.539)
LR	0.763 (0.761-0.764)	0.284 (0.284-0.285)	**0.97 (0.968-0.971)**	0.212 (0.21-0.213)	**0.03 (0.029-0.032)**	0.788 (0.787-0.79)
RF	0.788 (0.787-0.789)	0.028 (0.028-0.029)	0.956 (0.954-0.957)	0.341 (0.339-0.342)	0.044 (0.043-0.046)	0.659 (0.658-0.661)
SVM	0.774 (0.772-0.775)	0.274 (0.273-0.275)	**0.97 (0.969-0.972)**	0.238 (0.237-0.24)	**0.03 (0.028-0.031)**	0.762 (0.76-0.763)
XGBoost	0.784 (0.783-0.786)	**0.017 (0.017-0.018)**	0.935 (0.934-0.937)	0.364 (0.362-0.365)	0.065 (0.063-0.066)	0.636 (0.635-0.638)
Audit filters	0.624 (0.623-0.625)	-	0.932 (0.93-0.933)	0.316 (0.315-0.317)	0.068 (0.067-0.07)	0.684 (0.683-0.685)

Definition of abbreviations: ML = Machine learning; OFI = Opportunity for Improvement; RF=random forest; SVM = Support vector machine; LR = logistic regression; DT = decision tree; k-NN = k-nearest neighbors; AUC = area under the receiver operating characteristic curve; ICI = integrated calibration index; FNR = false negative rate; FPR = false positive rate.

9 Figure Legends

Figure 1. A flowchart describing the exclusions made and the process of trauma patient cases from arrival until opportunity for improvement decision. OFI = Opportunity for improvement. Survival/death defined as death by day 30.

Figure 2. The calculated, model agnostic, permuted feature importance for the first resample using AUC as the scoring metric. Overall, "Emergency procedure" was the most important predictor followed by "Injury Severity Score" The highest feature importance for a single model was "Injury Severity Score," accounting for 7% of model logistic regression performance. PH = Pre-hospital; ED = Emergency department, GCS = Glasgow Coma Scale, GOS = glascow outcome scale, RF=random forest, SVM = Support vector machine, LR = logistic regression, DT = decision tree, KNN = k-nearest neighbors, XGB = XGBoost, LGB = LightGBM, CAT = CatBoost.

10 Figures

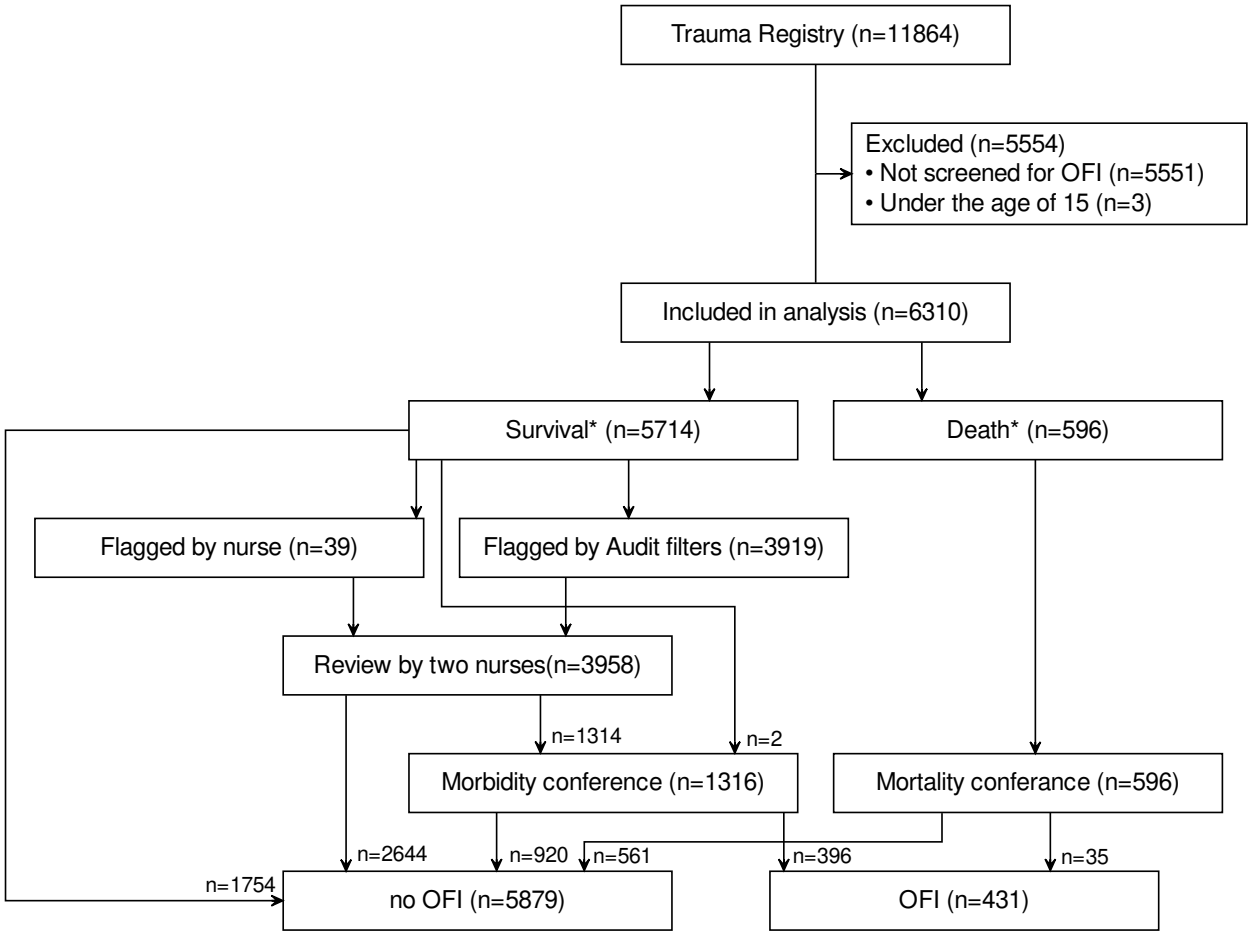

Figure 1. Flowchart describing the exclusions made and the process of trauma cases from arrival until OFI decision.

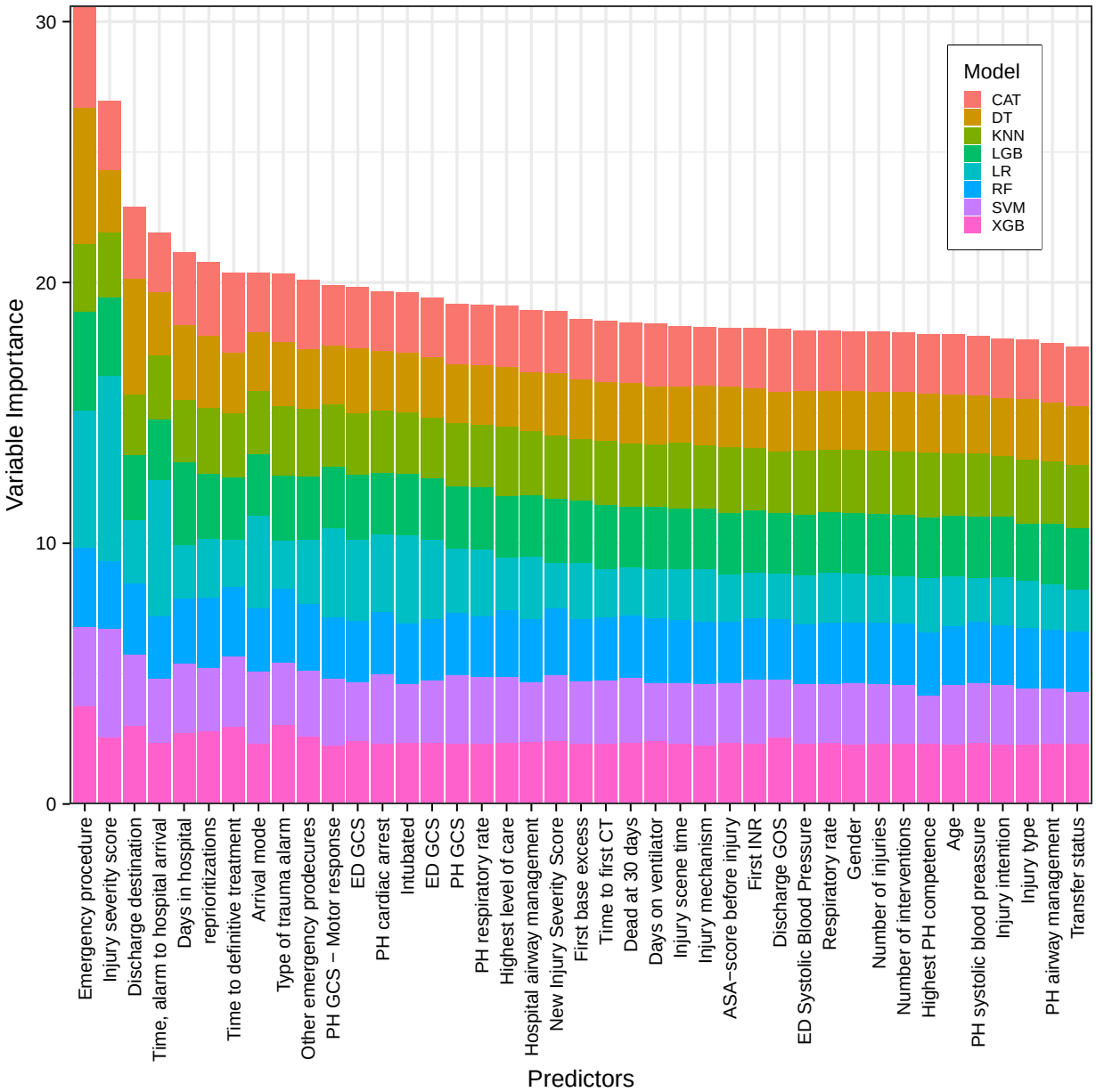

Figure 2. All predictors individual impact on each models classification. Calculated by using permuted feature importance.