Predicting Opportunities for Improvement in Trauma Care Using Machine Learning

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

[1]Department of Global Public Health, Karolinska Institute, Stockholm, Sweden. [2]Perioperative Medicine and Intensive Care, Karolinska University Hospital, Solna, Stock- holm, Sweden.

[3]Department of Medical Epidemiology and Biostatistics Karolinska Institute, 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

Manuscript word count: 2982

1 Key point

Question: How does the performance of machine learning models compare to that of audit filters when screening for opportunities for improvement (OFI), errors in care with adverse outcomes, among adult trauma patients?

Findings: Our registry-based cohort study including 8,220 patients showed that machine learning models outperform audit filters, exhibiting greater area under the curve values and reduced false-positive rates. Compared to audit filters, these models can be calibrated to balance sensitivity against overall screening burden.

Meaning: Machine learning models have the potential to reduce false positives when screening for OFI in adult trauma patients and thereby enhancing trauma quality programs.

2 Abstract

2.1 Importance

Identifying opportunities for improvement (OFI), errors in care with adverse outcomes, through mortality and morbidity conferences is essential for improving trauma quality. To screen patients for such conferences, trauma quality improvement programs rely on labor-intensive human reviews and audit filters that exhibit high false positive rates.

2.2 Objective

This study was conducted to develop machine learning models that predicts OFI in trauma care and

compare the performances of these models to those of commonly used audit filters.

2.3 Design

In this registry-based cohort study, we developed eight binary classification models using different machine learning methods with 17 predictors. Development used data from 2013 to 2022, and performance was measured between 2017 and 2022 using a add-one-year-in expanding window approach. We used two calibration strategies: 95% sensitivity (High sensitivity) and optimizing the area under the curve (Balanced). A bootstrap estimated confidence intervals.

2.4 Setting

The setting is a level one equivalent trauma center with bimonthly mortality and morbidity conferences for identifying OFIs; a combination of human review of individual patient cases and audit filters is used to screen patients for these conferences.

2.5 Participants

A total of 8220 adult trauma patients were screened for OFI. All patients prompted trauma team activation or were later found to have an injury severity score greater than 9.

2.6 Main outcome measures

Outcome measures were the models and audit filter performances, measured as discrimination, calibration, true positive and false positive rates.

2.7 Results

OFI were identified in 496 (6%) patients. The best performing model was XGBoost (High sensitivity: [auc:0.75, sens:0.904, FPR: 0.599], and Balanced: [auc:0.75, sens:0.502, FPR: 0.186]) followed by

Random Forest (High sensitivity: auc:0.733, sens:0.888, FPR: 0.617), and Balanced: [auc:0.733, sens:0.519, FPR: 0.222]). All machine learning models showed higher AUC and lower FPRs compared to Audit filters (auc:0.616, sens:0.903, FPR: 0.671).

2.8 Conclusion and Relevance

Machine learning models generally outperformed audit filters in predicting OFI among adult trauma patients, balancing and reducing overall screening burden for trauma quality improvement programs while potentially identifying new OFI types.

3 Introduction

Trauma is a leading cause of death and disability worldwide (1,2). Mortality and morbidity conferences are crucial components of trauma quality improvement programs (3,4). During these conferences, representatives from the disciplines and professions involved in trauma care discuss the care provided to specific patients to identify opportunities for improvement (OFI) (3,5,6). OFI represents errors in care with adverse outcomes (7), often related to airway management and hemorrhage control (8–10).

Selecting patients for these conferences relies on audit filters, sometimes in combination with individual human reviews (11). Audit filters represent sentinel events in patient care that are associated with suboptimal care and potentially poor patient outcomes, such as delays in conducting key tests or treatments or unexpected deaths (3,12). When such an event occurs, it triggers review through mortality and morbidity conferences, and if OFI are identified, then corrective actions should be instituted (12). Using audit filters to select patients is associated with high false positive rates, ranging from 24% to 80% (8,9,13).

Replacing filters with trauma mortality prediction models has been proposed to potentially improve the precision, but their performances in this context are poor (13–15), likely because they were not developed to predict OFI. No published research has evaluated prediction models for OFI. We aimed to develop machine learning models for predicting OFI during trauma care and compare the performances of these models to those of commonly used audit filters.

4 Methods

4.1 Design

We conducted a registry-based cohort study using all trauma patients included in both the Karolinska University Hospital trauma registry and the trauma care quality database. We used data from 2013 to 2022 to develop eight supervised machine learning models and compared the performances of these models with the performances of locally used audit filters. The study was approved by the Swedish Ethical Review Authority (approval numbers 2021- 02541 and 2021-03531).

4.2 Study Setting and Population

The Karolinska University Hospital in Solna, Sweden, is equivalent to a level 1 trauma center and manages approximately 1500 acute trauma patients each year (16).

The Karolinska University Hospital trauma registry, part of the Swedish Trauma Registry (16), includes all patients admitted to the Karolinska University Hospital with trauma team activation, regardless of injury severity score (ISS), as well as patients admitted without trauma team activation but found to have ISS of more than 9. The registry includes data on vital signs, times, injuries and interventions and demographics according to the European consensus statement, the Utstein template (17). The care quality database includes data relevant to mortality and morbidity conferences, including audit filters, identified OFI, and proposed corrective actions.

The mortality and morbidity conferences occur regularly, approximately eight times each year. The conference invites representatives from all professions involved in trauma care— surgery, neurosurgery, orthopedics, anesthesia and intensive care, nursing, and radiology— and identifies the presence of OFI and appropriate corrective actions through consensus. Electronic health records, including medical notes, laboratory parameters, and imaging, are the basis for these discussions. Selecting patients for the conference is a multistage process with escalating levels of review, with separate tracks for patients who die

within 30 days and those who survive. Patients who die are always reviewed in a multidisciplinary conference, henceforth referred to as a mortality conference. The mortality conference evaluates the preventability of the death and determines the presence of any OFI.

The process of identifying OFI in patients who survived was refined during the study period. Between 2013 and 2017, audit filters with high false positive rates in combination with limited human resources made individual review of the entire cohort impossible. Instead, a specialized trauma nurse reviewed as many patients as possible each month. Any potential OFI were then discussed among a small group of clinicians involved in trauma care. If the identified OFI was recurrent or particularly crucial, the patient was reviewed again at a broader multidisciplinary conference, establishing a precursor to the current conferences.

Starting in 2017, the process was further formalized with a specialized nurse performing a brief initial individual review and applying a new set of audit filters (Supplement E1) to all patients. Two specialized nurses perform a second more in-depth review of all patients flagged, through audit filter violation or other concern from the first nurse, for potential OFI during the first review. If this second review identifies a potential OFI, the patient is reviewed at a multidisciplinary conference, henceforth referred to as the morbidity conference, where the final decision on the presence of OFI is made.

4.3 Eligibility Criteria

We included all patients screened for OFI from the trauma registry and trauma care quality database between 1 January 2013 and 31 December 2022. Patients younger than 15 years were excluded because their clinical and review pathways differ from those of adults.

4.4 Opportunities for improvement

Opportunities for improvement are identified as avoidable errors in care with direct adverse effects on patient outcomes or recurrent deviations from safe clinical practice, occurring from arrival at the hospital up until discharge or death within 30 days. The different OFI are categorized by the mortality and morbidity conference as judgment errors, delays in treatment or diagnosis, missed diagnoses, technical errors, preventable deaths, or other errors.

4.5 Outcome

The models' outcome is the presence of an OFI, as determined 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".

4.6 Predictors

We considered all variables from the trauma registry as potential predictors. The final predictors were based on current audit filters, standard demographics, previous publications in the setting and expert opinion (18). The categorical predictors were gender, type of emergency procedure, highest level of care, reprioritization, type of trauma alarm, discharge destination and death within 30 days. The continuous features included age, vital signs on arrival, time to CT and intervention, ISS and length of stay. These comprise 17 variables with 45 corresponding parameters. eTable 1 (Supplement E3) shows all 17 predictors.

4.7 Sample Size Considerations

The relationship between the number of predictors and required sample size for different learners has not been well researched, but recommendations exist for logistic regression (19,20). According to these guidelines, we estimated that with a sample size of 3452, which is equivalent to 80% of the available data

from 2017–2020, would support 45 parameters, assuming a 6% event rate, a r2 of 0.11 and a target shrinkage of 0.9.

4.8 Statistical Analysis Methods

The statistical analyses were conducted using R (21). To evaluate the models, we used an add-one-year-in (AOYI) expanding window as well as a leave-one-year-out (LOYO) approach. In the main results, we present the AOYI results because they best represent how the models would have performed if they had been implemented prospectively. For each year (referred to as a split), we generated a validation sample using the holdout year. For AOYI, all years prior to the current validation sample were used as training data; for LOYO, the training data constituted all years, prior and future, excluding the current validation sample. The training data were then split, and 80% of the data were used for training and 20% for calibration.

Data preprocessing and imputation. A preprocessor was developed, rescaling continuous features using Yeo-Johnson's power transformation (22) and recoding categorical features into dummy variables via one-hot encoding. Predictors with near-zero variances were excluded. Missing continuous predictors were imputed using the mean of the predictor, and a missing indicator feature was created for each. Categorical predictors were imputed by introducing an 'unknown' category. If blood pressure or respiratory rate data were missing but corresponding revised trauma score categorical values were available, we imputed the missing data using the mean of all patients in that category.

The preprocessor was initially run on the training sample for each split to learn metrics and prevent data leakage. The trained preprocessor model was then applied independently to both the training and validation samples. To balance the training samples, we used the adaptive synthetic algorithm (23),

which generates synthetic data, enabling us to upsample OFI outcomes at a balanced 1:1 ratio.

Model development. Eight ML models were built using the Tidymodels framework (24). We used logistic regression (LR), random forest (RF) (25), decision tree (DT) (26), support vector machine with a radial basis kernel (SVM) (27), XGBoost (XGB) (28), LightGBM (LGB) (29), CatBoost (CAT) (30), and k-nearest neighbor (k-NN) (31). Each learner is briefly described in the online supplement (Supplement E2). All model hyperparameters were optimized on the training sample of each split using five-fold cross-validation through iterative Bayesian optimization, encompassing all the parameters provided by the tidymodels framework.

Performance measurements. The models and audit filters performance were assessed and compared in terms of false negative rates, false positives rates, discrimination, and calibration in each validation samples. Discrimination was measured using the area under the receiver operating characteristic curve (AUC), and calibration was measured using the integrated calibration index (ICI) (32). The ICI was not calculated for the audit filter system due to its inability to produce a probability for OFI.

To determine the class probability cutoff for each model, we first calibrated each model using Platt scaling on a 20% holdout sample from the training samples. We then determined the cutoff that produced a 95% sensitivity on the calibration sample and applied it to the holdout validation sample, called “High sensitivity”. Additionally, we conducted an analysis to establish an “optimal” cutoff threshold by identifying the point on the ROC curve that maximizes the trade-off between sensitivity and specificity, called “balanced calibration”. We estimated 95% confidence intervals (CIs) for all performance metrics through a bootstrap of 1000 resamples for each validation sample.

Feature importance. We calculated the feature importance for all models and predictors using permutation feature importance on the nonresampled validation samples (33). The importance of a feature was thus calculated by taking the average AUC performance when shuffling a feature's data five times and comparing it to the model's performance on nonshuffled data.

Code availability. The code used in this study is publicly available online: https://github.com/noacs-io/predicting-ofi-in-trauma under the MIT License.

5 Results

5.1 Participants

Out of the 13879 patients in the registry collected between January 2013 and December 2022, 8220 (59.87%) patients had been reviewed regarding the presence of OFI, which were identified in 496 (6%) patients. The mortality conference reviewed 710 (8.54%) patients and classified 42 (5.92%) deaths as preventable (n=4, 0.56%) or possibly preventable (n=38, 5.35%) with OFI. Of the 7510 patients who survived for 30 days, 1760 (23.44%) patients were selected for inclusion at a morbidity conference, and OFI was identified in 454 (6%) patients. Figure 1 shows the flow of patients throughout the review process.

Patients with OFI (mean=49 years, SD=21) were older than patients without OFI (mean=45 years, SD=21). The ISS was greater in patients with OFI (mean=19, SD=11) than in patients without OFI (mean=12, SD=13), and patients with OFI had longer times (mean=271 minutes, SD=323) from hospital arrival to definitive treatment than patients without OFI (mean=251 minutes, SD=353). Treatment frequencies also differed, with the biggest difference being radiological interventions, where patients with OFI (n=32 [6%]) had more interventions than those without OFI (n=69 [1%]). Table 1

shows the characteristics of all reviewed patients.

The variable "Other emergency procedures" had the highest frequency of missing data (n=6783) followed by "Time to definitive treatment" (n=5898). This can be explained by the absence of a needed intervention, which is registered as a missing variable. Complete details are provided in eTable 1 (Supplement E3).

5.2 Model Development

The years 2017-2022 were all used as separate hold-out test sets. The OFI frequency varied, the highest occurring in 2017 (n=112, [9%]) and the lowest in 2018 (n=36, [3%]). Annual characteristics are provided in eTable 2 (Supplement E3). Figure 2 provides the number of patients and OFI for each year and corresponding training datasets for the AOYI-analysis.

5.3 Model Specification and Performance

Figure 3 shows the average feature importance for all years between 2017 and 2022 for all predictors. Overall, ISS was the most important predictor, followed by highest level of care. All models showed higher performances in terms of AUC and false positive rates, and a sensitivity within 0.029, compared to Audit filters (auc:0.616, sens:0.903, FPR: 0.671). The overall best model was XGBoost (auc:0.75, sens:0.904, FPR: 0.599, ICI: 0.033, pairwise difference from audit filter: auc:0.134, sens:0.001, spec:0.073, FPR: 0.073). Table 2 shows the AOYI analysis for the three best models, as well as logistic regression and audit filters, including pairwise performance differences compared to the audit filters. Figure 2 shows annual AUC values between 2017 and 2022. The LOYO-analysis showed additional increased performance and stable calibration, with several models outperforming audit filters in all aspects. See the supplementary material, eTable 3 (complete AOYI analysis), eTable 4 (complete LOYO analysis),

eFigure1 (Annual ROC-curves), eFigure 2 and eFigure 3 (annual specificity and sensitivity) for additional information.

6 Discussion

We found that machine learning models generally outperformed audit filters, with superior in predicting OFI among adult trauma patients, with the potential to reduce the screening burden in trauma quality improvement programs. Compared to audit filters, these models are scalable to meet different contexts and needs where we tested two calibration strategies; one with higher sensitivity and a moderate reduction in false positives, and one with a moderate loss in sensitivity and a substantial reduction in false positives. Hence, these models could systematize and quantify the balance between OFI identification and general screening burden.

Importantly, the models' performances are falsely low due to two limitations. First, the main analysis used the AOYI approach to simulate a prospective implementation. Hence, between 2017 and 2020, model performance decreased due to the small sample size. This theory is supported by the LOYO-analysis, which verified that the models could exhibit stable calibration and overall higher performance, with several models outperforming audit filters in all aspects. The current sample size calculations where a compromise across the AOYI-analysis and to present a realistic model for future implementation. Second, these models are only evaluated against OFI identified within the current screening system and not the actual OFI frequency. Hence, any OFI found with these models and not audit filters would yield increased model performance and decreased audit filter performance. The low OFI frequency compared to previous studies supports that theory (8,9,13).

While the performance differences between the models and audit filters could appear small and the potential sensitivity drop for some models could be discouraging, the low incidence of OFI generates

large differences in actual patient numbers with relatively small specificity changes and low numbers despite significant sensitivity loss. The model performance that most differed from that of the audit filters in the High sensitivity-analysis was that of XGBoost, which had a similar sensitivity but a reduced screening burden, yielding 90 (11%) false-positive patients annually. These numbers scale when calibrated toward balanced performance, where XGBoost reduced the screening burden by 572 (72%) with 28 (46%) fewer identified OFI annually. While a reduction in sensitivity is suboptimal, one must remember that the performance is falsely low, and many hospitals opt out of general screenings due to their false-positive volumes. Hence, this substantial reduction offers possibilities for settings with limited human resources.

While defined as a binary variable, the OFI includes a diverse set of outcomes ranging from preventable deaths to lacking communication. The heterogeneity of these outcomes represents a range of clinical events, each likely correlating to different predictors. In addition, machine learning models struggle to handle rare events, and despite being an aggregate of all previously identified errors, the OFI frequency is only 6%; as a result, OFI is a considerable predictive challenge.

Another potential risk is a "data shift". First, quality improvement efforts and clinical advances likely mean that different mistakes occurred for different reasons in 2022 compared to 2013. Second, due to feasibility, mortality and morbidity conferences and corresponding corrective actions can focus only on a subset of OFI at any given time, partly due to false positives, which causes models to learn older, less relevant OFIs. Hence, a correctly selected OFI might not be registered since the conference must prioritize other areas in need of correction. If human resources could be removed from basic screening tasks by reducing false positives, they could possibly be allocated toward more in-depth reviews, reducing the need to prioritize OFI subgroups and instead enabling the conference to be focused on all,

including potentially new, subtypes.

Previous research successfully applied some of the learners used in our study for predicting mortality and other more homogeneous outcomes in trauma, were equal performance where often found using logistic regression compared to other learners (35). While OFI presents a more di□icult prediction challenge, our study showed similar results where logistic regression had average performance. To substantially increase the performance, both higher quantity and quality of data are likely needed, e.g., higher-resolution data such as vital sign series or imaging information and higher OFI quality, e.g., defined, complete and consistent OFI classifications. However, that would sacrifice external validity and general feasibility compared to the study models that are easily applicable in settings with registries following the Utstein template.

Importantly, perfect performance is far from expected. Comparing these models to entire systems using a combination of quantitative screening and several human reviews, including a multidisciplinary review, is unfair and not the goal of this paper. Instead, we strive to facilitate efforts through a combination of human and artificial intelligence where this is the first time an alternative to audit filters has been presented. Future research should focus on prospective testing to estimate true performance, but importantly, our results are probably falsely low due to unknown false negatives and limitations in data quantity. Compared to audit filters, these models are scalable in the sense that one can balance and optimize screening burden and sensitivity goals, giving each trauma quality improvement program the potential to standardize and automate part of the review system while assisting human resources.

7 Acknowledgments

The thank all professionals who participated in the monthly mortality and morbidity conferences. The authors also thank Liselott Västerbo for her participation in collecting and recording the data and screening for OFI.

8 Footnotes

8.1 Contributors:

M.G.W. and J.A. obtained funding and conceptualized the study. M.G.W., J.A. and K.S. drafted the study protocol. M.G.W., J.A., K.S. and M.J. wrote the statistical analysis plan.

J.A. and K.S. performed the statistical analysis and model development. J.A. and K.S. drafted the manuscript, which was critically revised by all the authors. All the authors read and approved the final manuscript. J.A., K.S. and M.G.W. are guarantors. J.A. and K.S. contributed equally to this work. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

8.2 Funding:

This work was 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.

8.3 Competing interests:

All the authors have completed the ICMJE uniform disclosure form at http://www.icmje.or g/disclosure-of-interest/ and declare the following: M.G.W. and J.A. received grants related to this study from the

Swedish Society of Medicine and from "The Swedish Carnegie Hero Fund". The authors have no financial relationships with any organizations that might have an interest in the submitted work in the previous three years. The authors have no other relationships or activities could appear to have influenced the submitted work. The lead author (the manuscript's guarantors) a□irms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

8.4 Dissemination to participants and related patient and public communities:

The results will be disseminated through local and international conferences. To date, results have presented at the London Trauma Conference (December 2022). Additionally, code for replicating the results and models are publicly available: https://github.com/noacs- io/predicting-ofi-in-trauma

9 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 [Internet]. 2018 Nov [cited 2022 Dec 17];392(10159):1736–88. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673618322037

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 [Internet]. 2020 Oct [cited 2022 Dec 17];396(10258):1204–22. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673620309259

3. World Health Organization. Guidelines for trauma quality improvement programmes [Internet]. 2009 [cited 2022 Aug 24] p. 104. Available from: 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 [Internet]. 2014 Jan;259(1):186–92. Available from: https://doi.org/10.1097/sla.0b013e31828df98e

5. 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 [Internet]. 2014 Jun [cited 2022 Dec 17];40(3):279–85. Available from: http://link.springer.com/10.1007/s00068-013-0364-5

6. 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 [Internet]. 2014 Mar;22(1). Available from: https://doi.org/10.1186/1757-7241-22-18

7. 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 [Internet]. 2014 Aug [cited 2022 Dec 17];208(2):187–94. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002961014001688

8. 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 [Internet]. 2011 Apr [cited 2022 Dec 17];70(4):970–7. Available from: https://journals.lww.com/00005373-201104000-00032

9. 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 [Internet]. 2017 Dec [cited 2022 Dec 17];17(1):142. Available from: http://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-017-2085-

10. 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 [Internet]. 2013 Apr 2 [cited 2022 Dec 17];100(6):749–55. Available from: https://academic.oup.com/bjs/article/100/6/749-755/6138406

11. Hornor MA, Hoeft C, Nathens AB. Quality benchmarking in trauma: From the NTDB to TQIP. Curr Trauma Rep [Internet]. 2018 Jun [cited 2022 Dec 17];4(2):160–9. Available from: http://link.springer.com/10.1007/s40719-018-0127-1

12. Evans C, Howes D, Pickett W, Dagnone L. Audit filters for improving processes of care and clinical outcomes in trauma systems. Cochrane Injuries Group, editor. Cochrane Database of Systematic Reviews [Internet]. 2009 Oct 7 [cited 2022 Dec 17]; Available from: https://doi.wiley.com/10.1002/14651858.CD007590.pub2

13. Ghorbani P, Strömmer L. Analysis of preventable deaths and errors in trauma care in a scandinavian trauma level-i centre. Acta Anaesthesiol Scand [Internet]. 2018 Sep [cited 2022 Dec 17];62(8):1146–53. Available from: https://onlinelibrary.wiley.com/doi/10.1111/aas.13151

14. Radke OC, Heim C. Recognizing preventable death. Anesthesiology Clinics [Internet]. 2019 Mar [cited 2022 Dec 17];37(1):1–11. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1932227518300880

15. 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 [Internet]. 2016 Sep [cited 2022 Dec 17];47(9):1960–5. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0020138316302145

16. Årsrapporter | SweTrau [Internet]. [cited 2021 Feb 10]. Available from: http://rcsyd.se/swetrau/om-swetrau/arsrapporter

17. 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 [Internet]. 2008 [cited 2022 Dec 10];16(1):7. Available from: http://sjtrem.biomedcentral.com/articles/10.1186/1757-7241-16-7

18. Albaaj H, Attergrim J, Strömmer L, Brattström O, Jacobsson M, Wihlke G, et al. Patient and process factors associated with opportunities for improvement in trauma care: A registry-based study. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine [Internet]. 2023 Nov;31(1). Available from: http://dx.doi.org/10.1186/s13049-023-01157-y

19. 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 [Internet]. 2020 Mar 18 [cited 2022 Dec 17];m441. Available from: https://www.bmj.com/lookup/doi/10.1136/bmj.m441

20. 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 [Internet]. 2019 Aug [cited 2022 Dec 17];28(8):2455–74. Available from: http://journals.sagepub.com/doi/10.1177/0962280218784726

21. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/

22. Yeo I-K. A new family of power transformations to improve normality or symmetry. Biometrika [Internet]. 2000 Dec 1 [cited 2022 Dec 17];87(4):954–9. Available from: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/87.4.954

23. Haibo He, Yang Bai, Garcia EA, Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) [Internet]. Hong Kong, China: Ieee; 2008 [cited 2022 Dec 17]. p. 1322–8. Available from: http://ieeexplore.ieee.org/document/4633969/

24. Kuhn M, Wickham H. Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. [Internet]. 2020. Available from: https://www.tidymodels.org

25. Breiman L. Random forests. Machine Learning [Internet]. 2001 [cited 2022 Dec 17];45(1):5–32. Available from: http://link.springer.com/10.1023/A:1010933404324

26. Belson WA. Matching and prediction on the principle of biological classification. Ap-

plied Statistics [Internet]. 1959 Jun [cited 2022 Dec 17];8(2):65. Available from: https://www.jstor.org/stable/10.2307/2985543?origin=crossref

27. Cortes C, Vapnik V. Support-vector networks. Mach Learn [Internet]. 1995 Sep [cited 2022 Dec 17];20(3):273–97. Available from: http://link.springer.com/10.1007/BF00994018

28. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining [Internet]. Acm; 2016 [cited 2022 Dec 17]. p. 785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785

29. 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 [Internet]. Curran Associates, Inc.; 2017. Available from: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

30. 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 [Internet]. Curran Associates, Inc.; 2018. Available from: https://proceedings.neurips.cc/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

31. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory

[Internet]. 1967 Jan [cited 2022 Dec 17];13(1):21–7. Available from: http://ieeexplore.ieee.org/document/1053964/

32. Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine [Internet]. 2019 Sep 20 [cited 2023 Jan 4];38(21):4051–65. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.8281

33. 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.

34. 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 [Internet]. 2021 Oct [cited 2022 Dec 17];28(1):e100407. Available from: https://informatics.bmj.com/lookup/doi/10.1136/bmjhci-2021-100407

35. Zhang T, Nikouline A, Lightfoot D, Nolan B. Machine learning in the prediction of trauma outcomes: A systematic review. Annals of Emergency Medicine [Internet]. 2022 Nov [cited 2022 Dec 17];80(5):440–55. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0196064422003353

10 Figure Legends

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

Legend: OFI = Opportunity for improvement. * Survival/death was defined as death by day 30.

Figure 2. Annual AUC values for each model. Sample sizes and OFI number for each training and test set.

Legend: Year wise model performance and sample sizes for the expanding window add one year in analysis. A) Mean area under the curve (AUC) per model and year. Lines represent 95% confidence intervals. For any given year, the AUCs were calculated with that year as the test set and all preceding years, starting with 2013, as the training set. For example, for 2019 the AUCs were calculated using 2019 as the test set and 2013-2018 as the training set. B) Opportunities for improvement (OFI) and sample sizes per year. The training sample and the test sample sizes includes the OFI in each sample respectively. For any given year, the training OFI and training sample size rows are the total number of patients with OFI and total number of patients in all preceding years respectively. The test OFI test sample size rows are the number of patients with OFI and the total number of patients in a specific year. For example, for 2019 the training OFI is the total number of patients with OFI 2013-2018, the training sample size is the total number of patients 2013-2018, the test OFI is the number of patients with OFI in 2019 and the test sample size the number of patients in 2019.

Figure 3. Average permuted variable importance for each model during the AOYI-analysis.

Legend: The calculated, model-agnostic, permuted feature importance calculated using the AUC as the scoring metric. Variable importance is measured as AUC change when a variable is permuted. The model values are the average within that model between the years 2017 and 2022. The “Mean” value is the

mean of all models.

Definition of abbreviations: PH = pre-hospital; ED = emergency department, GCS = Glasgow Coma Scale, GOS = Glasgow Outcome Scale, RF=random forest, LR = logistic regression, XGB = XGBoost, CAT = CatBoost.

11 Tables

Table 1. Demographic and clinical characteristics of patients screened for OFI

	OFI	No OFI	Overall
	(N=496)	(N=7724)	(N=8220)
Age			
Mean (SD)	49 (22)	45 (21)	45 (21)
Median [Min, Max]	49 [15, 97]	43 [15, 100]	43 [15, 100]
Sex			
Female	136 (27%)	2388 (31%)	2524 (31%)
Male	360 (73%)	5336 (69%)	5696 (69%)
Dead at 30 days			
Yes	41 (8%)	677 (9%)	718 (9%)
No	453 (91%)	7038 (91%)	7491 (91%)
Missing	2 (<1%)	9 (<1%)	11 (<1%)
Highest level of care			
Emergency department	22 (4%)	1467 (19%)	1489 (18%)
General ward	116 (23%)	2920 (38%)	3036 (37%)
Operation Theatre	141 (28%)	1438 (19%)	1579 (19%)
Specialist ward/Intermediate ward	50 (10%)	336 (4%)	386 (5%)
Intensive care unit	167 (34%)	1563 (20%)	1730 (21%)
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%)	10 (<1%)	10 (<1%)
Respiratory rate			
Mean (SD)	19 (5)	18 (5)	18 (5)
Median [Min, Max]	18 [0, 40]	18 [0, 60]	18 [0, 60]
Missing	51 (10%)	812 (11%)	863 (10%)
ED GCS			
Mean (SD)	14 (3)	14 (2)	14 (2)
Median [Min, Max]	15 [3, 15]	15 [3, 15]	15 [3, 15]
Missing	49 (10%)	811 (11%)	860 (10%)
ED Systolic Blood Pressure			
Mean (SD)	133 (34)	133 (33)	133 (33)
Median [Min, Max]	135 [0, 237]	135 [0, 285]	135 [0, 285]
Missing	0 (0%)	13 (<1%)	13 (<1%)
Time to first CT			
Mean (SD)	76 (129)	71 (159)	71 (157)
Median [Min, Max]	40 [6, 1339]	33 [0, 7073]	33 [0, 7073]
Missing	42 (8%)	945 (12%)	987 (12%)
Time to definitive treatment			
Mean (SD)	271 (323)	251 (353)	253 (349)
Median [Min, Max]	143 [3, 1420]	102 [0, 2036]	106 [0, 2036]
Missing	230 (46%)	5668 (73%)	5898 (72%)
Emergency procedure			
Thoracotomy	8 (2%)	97 (1%)	105 (1%)
Laparotomy	28 (6%)	213 (3%)	241 (3%)
Pelvis Packing	0 (0%)	5 (<1%)	5 (<1%)
Revascularization	12 (2%)	37 (<1%)	49 (1%)
Radiological intervention	32 (6%)	69 (1%)	101 (1%)
Craniotomy	42 (8%)	240 (3%)	282 (3%)
Intracranial pressure measurement	13 (3%)	90 (1%)	103 (1%)
Other	131 (26%)	1305 (17%)	1436 (17%)
Missing	230 (46%)	5668 (73%)	5898 (72%)

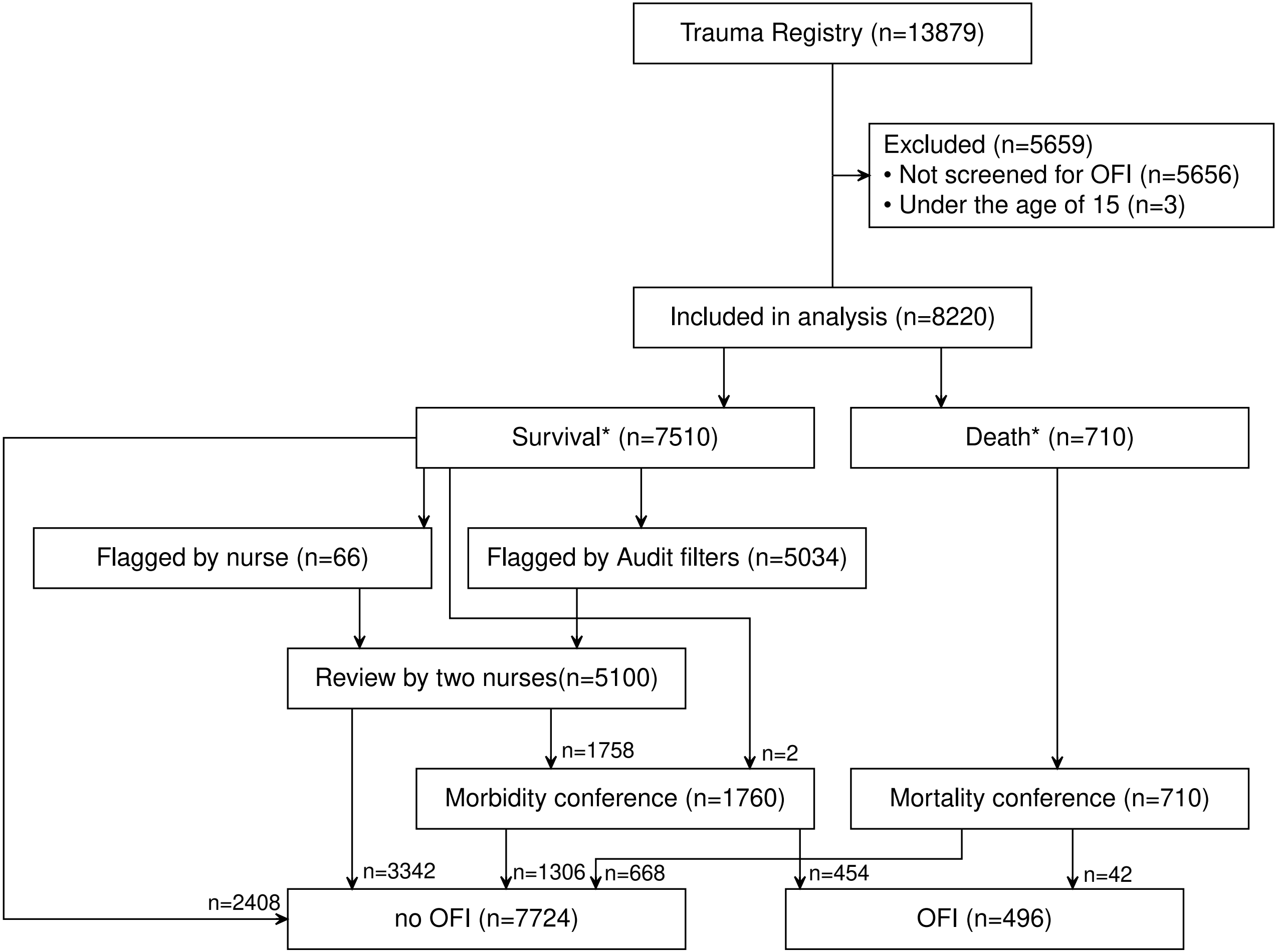
Definition of abbreviations: OFI = Opportunity for Improvement; ED = Emergency Department; GCS = Glasgow Coma Scale.
Time to first CT and time to definitive treatment: Measured in minutes from arrival at the hospital.

Table 2. Performance Metrics for AOYI-analysis

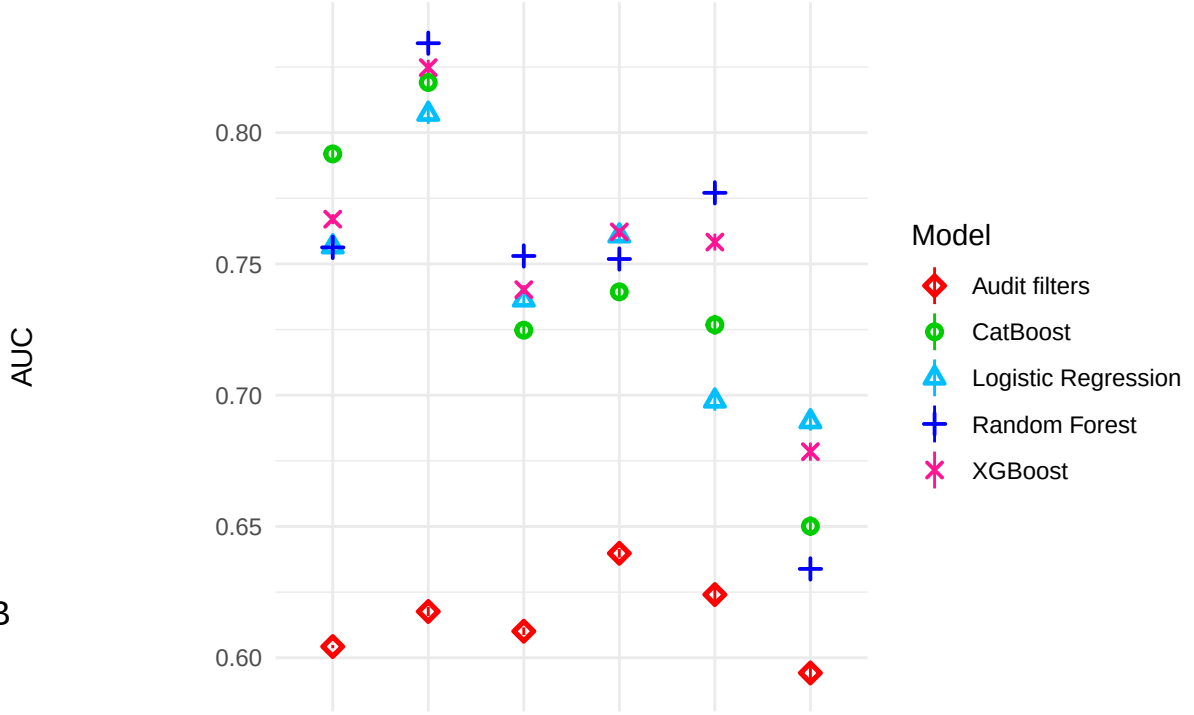
Model	AUC	Sensitivity	FPR	ICI
Calibrated to high sensitivity				
XGBoost	0.75(0.747-0.753)	0.904(0.901-0.907)	0.599(0.598-0.6)	0.033(0.032-0.033)
Random Forest	0.733(0.73-0.736)	0.888(0.884-0.891)	0.617(0.616-0.619)	0.031(0.031-0.032)
CatBoost	0.721(0.718-0.724)	0.874(0.871-0.878)	0.615(0.613-0.616)	0.029(0.028-0.029)
Logistic Regression	0.72(0.717-0.723)	0.885(0.881-0.888)	0.636(0.635-0.638)	0.032(0.032-0.032)
Audit filters	0.616(0.614-0.618)	0.903(0.9-0.906)	0.671(0.67-0.672)	-
Performance differences between audit filters and high sensitivity models				
XGBoost	0.134(0.132-0.137)	0.001(-0.003-0.005)	0.073(0.074-0.071)	-
Random Forest	0.117(0.115-0.12)	-0.016(-0.02--0.011)	0.054(0.055-0.052)	-
CatBoost	0.105(0.102-0.108)	-0.029(-0.033--0.024)	0.057(0.058-0.055)	-
Logistic Regression	0.104(0.101-0.107)	-0.019(-0.023--0.015)	0.035(0.036-0.033)	-
Balanced Calibration				
XGBoost	0.75(0.747-0.753)	0.502(0.496-0.507)	0.186(0.185-0.187)	0.033(0.032-0.033)
Random Forest	0.733(0.73-0.736)	0.519(0.514-0.524)	0.222(0.221-0.223)	0.031(0.031-0.032)
CatBoost	0.721(0.718-0.724)	0.401(0.396-0.407)	0.166(0.165-0.167)	0.029(0.028-0.029)
Logistic Regression	0.72(0.717-0.723)	0.501(0.496-0.507)	0.218(0.217-0.219)	0.032(0.032-0.032)
Audit filters	0.616(0.614-0.618)	0.903(0.9-0.906)	0.671(0.67-0.672)	-
Performance differences between audit filters and balanced models				
XGBoost	0.134(0.132-0.137)	-0.401(-0.407--0.396)	0.485(0.486-0.484)	-
Random Forest	0.117(0.115-0.12)	-0.384(-0.39--0.379)	0.449(0.451-0.448)	-
CatBoost	0.105(0.102-0.108)	-0.502(-0.507--0.496)	0.505(0.507-0.504)	-
Logistic Regression	0.104(0.101-0.107)	-0.402(-0.407--0.396)	0.453(0.455-0.452)	-

Average performance measures for the three best models, logistic regression, and audit filters for the expanding window add on year in approach. The performance differences are calculated by subtracting the audit filter performance values from the corresponding model value. ICI is not calculated for audit filters since they don't output prediction probabilities.

Definition of abbreviations: AUC = Area under the ROC Curve; FPR = False positive rate; ICI = Integrated calibration index


Trauma Registry (n=13879)
Excluded (n=5659)
• Not screened for OFI (n=5656)
• Under the age of 15 (n=3)
Included in analysis (n=8220)
Survival* (n=7510)
Death* (n=710)
Flagged by nurse (n=66)
Flagged by Audit filters (n=5034)
Review by two nurses(n=5100)
n=1758
n=2
Morbidity conference (n=1760)
Mortality conference (n=710)
n=3342
n=1306
n=668
n=454
n=42
n=2408
no OFI (n=7724)
OFI (n=496)

A



B

	2017	2018	2019	2020	2021	2022
Training OFI	95	207	243	342	413	450
Training sample size	761	2043	3363	4544	5868	7133
Number of OFI	112	36	99	71	37	46
number of patients	1282	1320	1181	1324	1265	1087

Year