Association between trauma severity models and opportunities for improvement: A retrospective cohort study

Victoria Witthoff

Contents

1 Abstract

Background: Trauma is a leading cause of death worldwide, and is the cause of a significant portion of the health burden of disability. Morbidity and mortality conferences are the cornerstone of trauma quality

improvement programs and there has been some focus in research on finding ways to improve the effectiveness of these conferences. One proposed method, that is already in use in many hospitals, is the use of trauma severity models as audit filters. The ability of these models to predict opportunities for improvement (OFI) is not well understood. *Aim:* To evaluate how the commonly used models TRISS, NORMIT and PS are associated with, and to what extent they can predict, OFI. *Methods:* This is a retrospective registry-based cohort study at Karolinska University Hospital covering the period 2014-2021. Probability of survival was calculated according to published algorithms. Logistic regression was used to estimate the association between the models and OFI. Receiver operating curves and the Integrated Calibration Index were used to assess discrimination and calibration. *Results:* 5320 participants were included from a seven-year period. Results were similar for all models. Lower probability of survival was associated with higher probability of OFI, with TRISS having the strongest association (OR: 0.35). Predictive performance and calibration was poor across the board. TRISS had the best AUC of 0.67. NORMIT had the best calibration with an ICI of 0.78. Accuracy [cannot be written because it's all wrong]. *Conclusions:* Low probability of survival is associated with increased odds of OFI, but none of the models were able to reliably predict OFI. The use of these models as audit filters in case selection for M&M conferences is not supported by our results and may lead to over-selection.

2   Abbreviations

- ASA-PS - American Society of Anesthesiology Physical Status
- ISS - Injury Severity Score
- M&M - Morbidity and mortality
- NISS - New Injury Severity Score
- NORMIT - Norwegian Survival Prediction Model in Trauma
- OFI - Opportunities for improvement
- PS - Probability of Survival score
- QI - Quality Improvement
- RTS - Revised Trauma Score
- TARN - Trauma Audit & Research Network
- TRISS - Trauma and Injury Severity Score
- WHO - World Health Organisation

3   Introduction

Trauma, defined as the clinical entity composed of physical injury and the body's associated response (1), is one of the leading causes of death worldwide and accounts for approximately 9% of global deaths annually (2). The global health burden of trauma has led to it being described as a worldwide pandemic. One of the characteristics of this particular pandemic is its demographic spread; trauma is a leading cause of death for people aged 10-49 years (2) and 90% of trauma-related deaths occur in low- to middle-income countries (3). Non-fatal outcomes also comprise a significant portion of the global health burden of disability (2), which can drastically reduce quality of life and health status in survivors of severe trauma (2,4).

The quality of trauma care is therefore a globally relevant issue with significance for a great number of people. Improvements in trauma outcomes benefit both the individual and also society at large by reducing the societal and healthcare costs of disability. The significance of the consequences and burden of trauma has been recognised by the UN, who in their Sustainable Development Goals included the aim of reducing the number of global deaths and injuries from road traffic accidents by half by 2020 (5). The importance of improving trauma care named as one of the strategies in achieving this goal. The goal was not met by 2020 (6), which is another signal that there is significant progress to be made in trauma care.

Efforts to advance the state of trauma care have been ongoing for many years. Today many trauma centres have implemented quality improvement (QI) programs, and it has been shown in the US that these programs improve outcomes (7). QI programs use many tools, including statistical analysis and audit filters (8), but the gold standard of trauma QI is the multidisciplinary morbidity and mortality (M&M) conference. These

conferences review selected patient cases and establish whether there was opportunity for improvement (OFI) in each case.

One area that is understudied is the use of trauma severity models as audit filters in case selection for M&M conferences. These models are algorithms that estimate the probability of survival. It has been proposed that using such models as audit filters to automatically flag cases for review could improve the effectiveness of M&M conferences while also reducing their resource burden (9,10). Many hospitals have already implemented this approach (9); however, the ability of trauma severity models to predict OFI is not well understood and will be investigated in this paper.

3.1   Morbidity and mortality conferences

The significance of M&M conferences in trauma QI is a consequence of the unique challenges inherent to trauma care. Management of severe trauma is highly time-sensitive and is also dependent on the actions of a chain of healthcare providers spanning multiple professions and disciplines. Outcomes in severe trauma are dependent on the quality of care received and errors in management are a common cause of preventable deaths (7,10–12). Since mistakes by any one provider can lead to severe outcomes, attempts to improve trauma outcomes must seek to improve all aspects of the entire chain of care. This can only be achieved through peer review and consequently, M&M conferences are the cornerstone of trauma QI programs.

The purpose of M&M conferences is to review selected patient cases, establish whether there were opportunities for improvement, and implement corrective action. The occurrence of OFI function in this manner as a proxy for incidence of suboptimal care. Regular conferences allow QI programs to identify patterns of errors and provide a basis for systemic quality improvement. To maximise effectiveness, conferences should be attended by all involved specialities and have a structured format. The WHO guidelines for trauma QI programs give specific recommendations as to the structure and format of M&M conferences (13). Despite these recommendations, there is a still a great degree of variation amongst clinics in how M&M conferences are carried out, with a significant number of clinics still using an informal and unstructured format (14). The World Health Organisation (WHO) themselves note in their guidelines that M&M conferences are not used to their fullest potential globally (13).

The value of M&M conferences in long-term quality improvement is largely dependent on selection of appropriate cases for review. For example, imagine a conference where every case reviewed was found to have no OFI. The value of that conference would effectively be zero, as it would fail to discover any flaws in trauma care. In an ideal world, only cases that were actually mismanaged would be selected for review; this is, of course, an impossible ideal as it cannot be known whether a case was mismanaged prior to it being reviewed. This reliance on case selection is a weakness of M&M review and has led to considerable efforts to develop strategies to standardise case selection (14). The WHO recommends that all "deaths, complications, adverse events and errors" should be reviewed (13). Some clinics have made further advances in standardising case election; most commonly by using quality indicators, or audit filters, to automatically flag cases for inclusion in M&M conferences (14,15). Examples of audit filters include time to first medical contact, adverse event rate, and, in some clinics, trauma severity models (13,15).

The benefit of standardising and automating case selection is twofold; it could increase identification of the cases that are most likely to have been mismanaged, and it could also save resources in the form of work hours. The latter benefit is not insignificant considering the large caseload found in many urban trauma centres. The process of identifying cases is time-consuming and trauma centres that lack resources may not be able to allocate enough staff to case selection, which in turn lowers the effectiveness of M&M review. Despite the benefits of standardised case selection, several studies have shown that the process by which patient cases are selected for M&M review differs between clinics and case selection is often non-standardised (14,16).

3.2   Trauma severity models

Trauma as a clinical entity is highly heterogeneous due to the variety in its underlying causes, which affect both the type and severity of the injury. One of several challenges posed by this heterogeneity is difficulty in expressing the severity of an injury in quantitative terms. For example, it is difficult to intuitively say who

was the more severely injured out of a patient with multiple gun shot wounds and a patient who was injured in a high speed road traffic accident. The ability to estimate trauma severity quantitatively is crucial for trauma QI programs as it allows trauma centres to assess outcomes and clinic performance.

Trauma severity models are algorithms that seek to address this need. These models translate anatomical and physiological parameters into a quantitive value that represents the totality of the injury to any given patient. Trauma severity models were initially intended for use in prehospital triage, and therefore were simplistic by design for ease of use (17), but over the years newer and more advanced algorithms have been developed to fulfil the needs of research, quality assurance, and clinical triage (17). Today there are hundreds of trauma severity models that differ in methodology but have the common aim of expressing trauma severity as a numerical value that either directly or indirectly represents the probability of survival.

Trauma severity models have become a cornerstone of modern trauma QI programs (18). These models allow clinics to compare their outcomes against an established benchmark. One example of their use is highlighting unexpectedly high mortality rates amongst patients with a high predicted probability of survival. (13). They also provide a common language for expressing trauma severity, which enables clinics to compare their performance with other, comparable clinics (13).

An inherent weakness of any trauma severity model is that estimates may be unreliable when applied to a population different from the one the model was developed from. This limitation in generalisability is a consequence of the heterogeneity of trauma, as the spectrum of traumatic injury varies across geographical regions (19). This weakness means that there is no single "best" model, as the performance of any model will be best in the population the model was sourced from. This is one reason why new models continue to be developed. Another reason is that efforts to establish which risk factors best predict mortality are still ongoing.

The most commonly used model worldwide is The Trauma and Injury Severity Score (TRISS) (20), first developed in the US in 1987 (21) and most recently updated in 2010 (22). TRISS is a combination index that incorporates patient age, the Injury Severity Score (ISS), the Revised Trauma Score (RTS), and whether the injury was blunt or penetrating (21). One weakness of TRISS is that the RTS requires GCS on arrival at hospital, which cannot be estimated in situations where the patient has already been intubated. This gives TRISS limited value in estimating probability of survival in certain populations (20). Another commonly used model is the Probability of Survival (PS) score, originally developed by the Trauma Audit & Research Network (TARN) in the UK in 2007 (23). In some ways the PS model is more complex than TRISS; it has coefficients for sex associated with age, and provides a formula that takes into account the non-linearity of ISS (23).

In 2014 researchers in Norway released a new model, the Norwegian Survival Prediction Model in Trauma (NORMIT), that attempts to address the weaknesses of TRISS and simultaneously provide a model based on Norwegian patient populations (24). NORMIT is based on the patient's age, New Injury Severity Score (NISS), comorbidities prior to traumatic injury based on the ASA physical status classification system (ASA-PS), and their RTS. NISS differs from ISS in that NISS allows for scoring of multiple injuries within the same anatomical region. Validation of NISS has shown that NISS is better at predicting mortality in trauma patients compared to ISS, especially in patients with multiple severe injuries within one anatomical region (25). NORMIT also provides a method to incorporate the GCS for intubated patients when calculating RTS. For patients who arrive at hospital and under general anaesthesia, the RTS is calculated based on the prehospital value prior to intubation (24). NORMIT was updated in 2018 (26) and a validation study found that the updated NORMIT reliably predicted survival at a Swedish trauma centre but both the old and new versions of NORMIT performed worse when used at the national level (27).

3.3   Knowledge gap

As previously mentioned, some trauma centres use trauma severity models as audit filters in selecting cases for M&M conferences (13). One example of their use as audit filters is in only reviewing deaths in cases that had a high predicted probability of survival. This strategy is based on the idea that deaths in cases with high predicted survival are more likely to have been preventable deaths. The factor that determines whether a death was preventable is whether an improvement in care could have prevented the death, i.e. whether there

was a significant OFI. The use of trauma severity models as audit filters for preventable deaths is therefore based upon the expectation that these models can predict OFI in mortalities. Research to date does not support that trauma severity models can predict OFI in deaths. Two major studies have looked at trauma severity models and preventable deaths. One study in Sweden investigated TRISS and NORMIT (10) and a study in the UK investigated TRISS and PS (9). Both of these studies showed that these studied models were unable to reliably discriminate between preventable and non-preventable deaths and applying cutoffs led to some preventable deaths being missed (9,10).

The studies mentioned signify that trauma severity models are not well suited to predicting OFI in fatalities and indicate that they should not be used for this purpose. However, the extent to which trauma severity models are associated with or can predict OFI is not well understood, as research thus far has only analysed their role in predicting preventable deaths. To date there is no published research that investigates whether trauma severity models can predict OFI across the board.

There is value in determining the effectiveness of trauma severity models in predicting OFI. If they can reliably predict OFI, their use as audit filters would be supported as they would improve case selection for M&M conferences. More accurate case selection will lead to M&M conferences becoming more effective at discovering flaws in care. There is also a secondary benefit in that effective audit filters allow case selection to become more automated, which reduces the large resource burden of identifying cases for review. Conversely, if trauma severity models cannot reliably predict OFI, we would know that their use as audit filters is harmful to the M&M process. Audit filters that are poor predictors can be harmful in two ways. On the one hand, they can over-predict OFI, causing more cases to be reviewed unnecessarily. This reduces the effectiveness of M&M conferences. On the other hand, they can miss cases with OFI, leading to flaws being undetected by M&M conferences. Given that some hospitals already use trauma severity models as audit filters, it is prudent to ascertain whether this is a helpful or harmful feature in the trauma QI process.

3.4   Aim

The aim of this study is to evaluate how the trauma severity models TRISS, NORMIT and PS are associated with and to what extent they can predict opportunities for improvement.

4   Methods

4.1   Study design

This is a single-centre retrospective cohort study, using data from two registries at Karolinska University Hospital: the trauma registry and the trauma care quality database. The trauma care quality database is a subset of the trauma registry and contains data on trauma patients who were selected for M&M reviews. These two databases were linked and then analysed to estimate the association between the common trauma severity models TRISS, NORMIT, and PS and OFI using logistic regression. The predictive performance of these models was evaluated using measures of discrimination and calibration.

4.2   Setting

Karolinska University Hospital is located in Solna, Sweden. The trauma centre at Karolinska University Hospital is equivalent to a level 1 trauma centre according to American College of Surgeons standards (28) and it is the primary trauma centre for the region of Stockholm. Karolinska University Hospital also has agreements with several other regions to provide trauma care for patients with severe trauma and the trauma centre has thereby a catchment population of almost 3 million residents (28).

The trauma registry at Karolinska University Hospital is part of SweTrau, Sweden's national trauma registry, and thereby follows SweTrau guidelines. All trauma patients meeting the SweTrau inclusion criteria are registered in the trauma registry by research nurses at the trauma centre. Data is collected from the patient's electronic medical records.

SweTrau follows the Utstein Trauma Template for reporting of data following major trauma. The registered data includes prehospital vital signs and management, vital signs and management on arrival at hospital, information about the type of injury, and the primary method of treatment. Follow-up variables include, but are not limited to, survival at 30 days, days in ventilator, Glasgow Outcome Scale score at discharge, highest level of care, and transfers to other units.

The trauma care quality database is an internal hospital registry and registers the outcome of M&M conferences. M&M conferences are held at Karolinska University Hospital ten times per year and are attended by all involved specialties and professions. Each case reviewed is registered in the trauma care quality register with data about the outcome of the review. The database contains data on patients from 2014 to 2021.

4.3 Participants

Participants were patients registered in both the trauma registry and the trauma care quality database at Karolinska University Hospital. The eligibility criteria for this study were the criteria for the two databases used in addition to some study-specific inclusion criteria.

4.3.1 Inclusion criteria for the trauma registry

The trauma registry includes all patients for whom the trauma team was activated after receiving a potentially traumatic injury, regardless of NISS score, and all patients with a NISS score >15, regardless of whether the trauma team was activated. Also included are patients who were transferred to the hospital within 7 days of injury and have a NISS score >15. Patients where the only traumatic injury is a chronic subdural haematoma and patients for whom the trauma team was activated without an underlying traumatic injury are excluded from the register.

4.3.2 Inclusion criteria for the trauma care quality database

The trauma care quality database contains patients from the trauma registry that were selected for M&M review. At Karolinska University Hospital, all fatalities are automatically selected for M&M review. In addition to this, the electronic medical records of all trauma patients are screened by research nursing staff to identify cases that potentially received sub-optimal care. These nurses also use audit filters that automatically highlight cases with abnormal parameters. Cases that two research nurses have evaluated and found to have potential for sub-optimal care are also selected for M&M review. Figure 1 gives an overview of the workflow that leads to determination of OFI.

4.3.3 Study-specific eligibility criteria

Patients <15 years of age were excluded as paediatric trauma is a separate clinical entity and is not part of the same review process. Patients that were dead on arrival were also excluded. This is a complete case study and all patients with missing data were excluded.

4.4 Variables

4.4.1 Outcome

The studied outcome was the binary variable "opportunities for improvement" (OFI), as identified by the multidisciplinary M&M conference at Karolinska University Hospital. OFI is registered in the trauma care quality database and is coded as either "Yes - at least one opportunity for improvement identified" or "No - no opportunities for improvement identified."

4.4.2 Exposures

The exposures of interest were the trauma severity models TRISS, NORMIT, and PS. All three of these models estimate the probability of death, expressed as a probability. The scores given by each of these models were calculated using physiological and anatomical parameters recorded in the trauma registry. Conversion and handling of these variables was carried out according to the SweTrau manual.
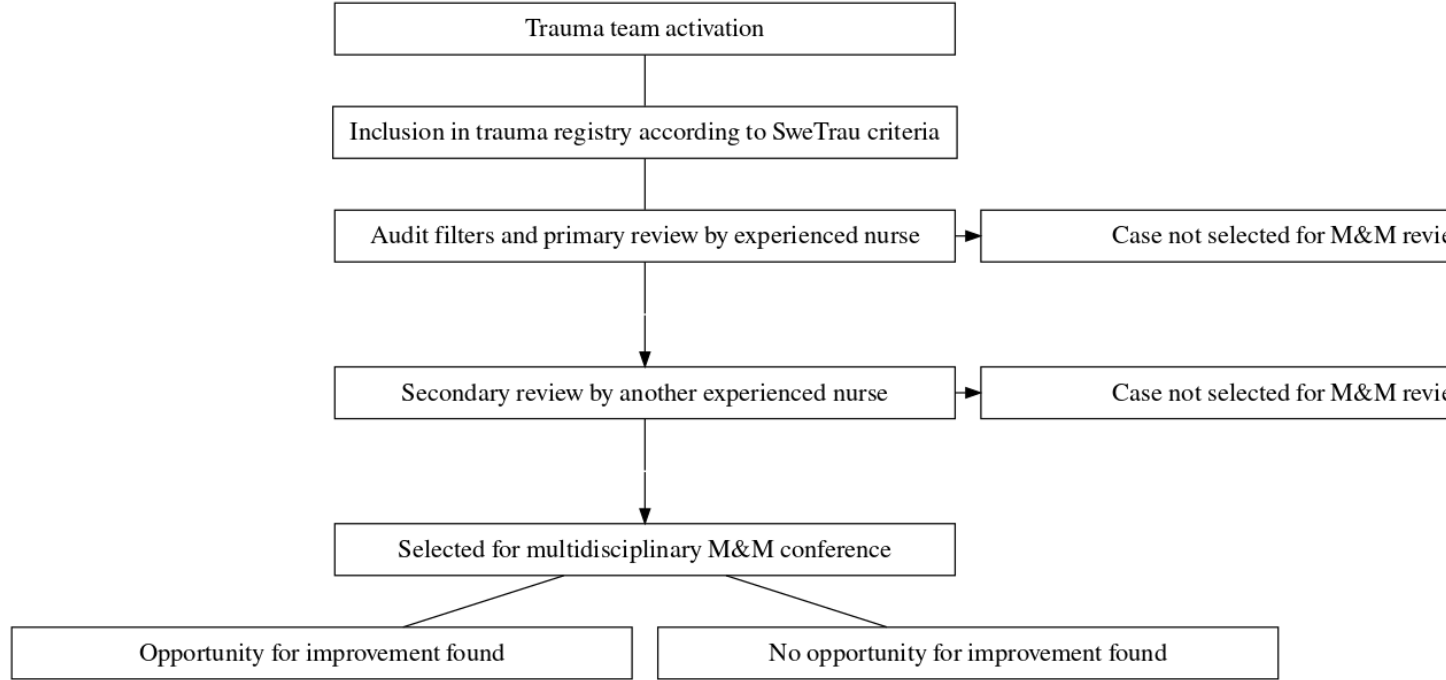
Figure 1: Workflow for case selection and determining OFI

4.4.2.1   TRISS   The TRISS model incorporates patient age, Injury Severity Score (ISS), GCS, systolic blood pressure, and respiratory rate. There are two sets of coefficients for these variables according to whether the trauma was blunt or penetrating.

The ISS is recorded in the trauma registry as part of original data collection. It is given by estimating the severity of physical injuries by anatomical region. Severity is assessed on a seven-point scale from "No injury" to "Unsurvivable injury." The values from the three anatomical regions with the most severe injury are combined into a score that is given is points on a scale from 0 to 75.

4.4.2.2   NORMIT   The NORMIT model incorporates the patient's age, New Injury Severity Score (NISS), comorbidities prior to traumatic injury based on the ASA physical status classification system (ASA-PS), and their RTS.

The NISS is recorded in the trauma registry as part of original data collection. NISS is based on the same variables as ISS. It differs from ISS in that it takes into account multiple injuries within one anatomical region and calculates a score based on the three most severe injuries regardless of anatomical region.

The RTS was computed using GCS, systolic blood pressure, and respiratory rate. For the systolic blood pressure and respiratory rate, values are taken from arrival at hospital. The GCS is also taken from arrival at hospital except in cases where the patient is intubated pre-hospitally; in these cases, the pre-hospital GCS is used. These parameters are weighted and expressed in a score from 0 to 8, where a lower score indicates more severe injury.

4.4.2.3   PS   The algorithm for PS includes the patient's age, gender, GCS and ISS. We chose to use the PS12 algorithm instead of the newer PS19 model due to our registry lacking data needed to calculate PS19.

4.5   Quantitative variables

RTS and PS12 were computed according to the original published algorithms (29,30). NORMIT was calculated with the revised 2018 algorithm (26). TRISS was calculated using the revised 2009 coefficients (22).

For all parameters other than GCS, the values used to compute scores were solely taken from arrival at hospital. For GCS, values were used based on arrival at hospital except in cases where the patient was intubated pre-hospitally. PS and NORMIT have methods of handling this scenario, as described above. TRISS does not have a published method to handle patients intubated pre-hospitally. In order to avoid excluding patients who were intubated pre-hospitally, the GCS parameter for TRISS was handled in the same manner as NORMIT, i.e. the pre-hospital GCS value was used.

4.6 Bias

The code for data analysis was written using simulated data to reduce the risk of researcher bias. Real-world data was only used once the data analysis model was completed and found to work correctly on simulated data.

4.7 Statistical methods

R, a programming language for statistical computing, was used for data collection and analysis (31). Logistic regression was used to estimate association between the estimates of the trauma severity models and OFI and the effect size was given as the OR. The CI for the OR was computed using the Wald estimator. Receiver operator characteristic (ROC) curves were used to assess diagnostic performance. The area under the curve (AUC) was calculated to compare the models with one another. The Integrated Calibration Index (ICI) was used to measure calibration. Accuracy was plotted and is reported as accuracy at a $>=0.5$ cutoff. Confidence intervals for the AUC, ICI and accuracy were computed with 1000 stratified bootstrap replicates. Results are presented with a confidence interval of 95%. A p-value of <0.05 was considered significant.

5 Ethical considerations

All patient information used in this study is pseudononymised and stored in a secure database. Patient medical records will not be accessed. Patients were not asked to give consent for inclusion in the Karolinska University Hospital trauma registry. The ability and ethical approval to collect patient data for this registry without pre-emptively asking for patient consent is well-established, as collecting this data is deemed to be in the public interest according to article 6.1 e of GDPR regulations. Collecting data is in the public interest due to the fact that this patient register is used to improve management of trauma, which is a leading cause of death. Patients have the right to be removed from the register on request. They also have the right to a copy of data that is stored about them on request.

This project is a retrospective cohort study and the project has thereby had no effect on patient treatment. There is no risk that patients are harmed by inclusion in the study. This study has the potential to improve patient care by analysing a potential factor that could predict suboptimal trauma care. It is true that the patient population for this study includes vulnerable populations, such as the deceased and the severely injured, but this is justified in the opinion of this author by the fact that this population stands to gain the most from improved trauma care.

Stockholm Research Ethics Review Board approval number 2021-02541 and 2021-03531.

6 Results

6.1 Participants

A total of 11864 patients were recorded in the trauma registry during the study period. Of these, 6313 were reviewed and therefore became eligible for study inclusion. 993 patients were excluded, due to either being of paediatric age (3), being dead on arrival at hospital (141), or due to incomplete data (849). Figure 2 shows the inclusion process for eligible participants.

Detailed totals of missing data for each required parameter in cases that were excluded for incomplete data are given in table 1. Some patients had missing data in several parameters and therefore the total number of
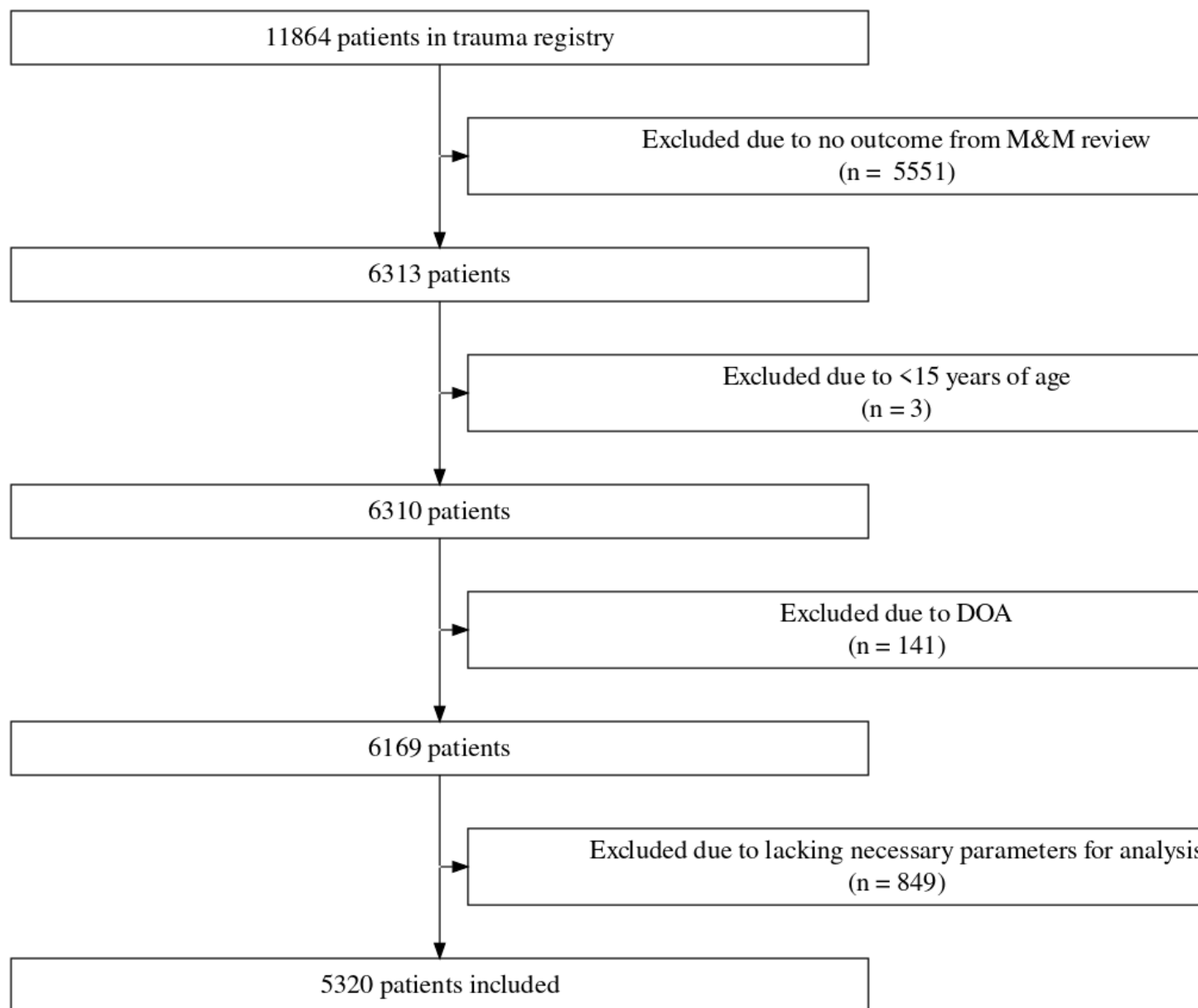
Figure 2: Flow diagram for application of exclusion criteria

Table 1: Parameters that were missing in cases excluded for incomplete data

Parameter	Total no. cases missing parameter
GCS	199
ASA	10
RR	583
Systolic BP	155
Dominant Injury Type	0
Age	0
ISS	3
NISS	3
Gender	0

missing parameters is greater than the total number of patients excluded due to incomplete data.

6.2 Descriptive data

A total of 5320 participants were identified as eligible for inclusion. Their demographical and clinical characteristics are described in table 2.

The median age was 42 (IQR: 26-60). 68.6% of participants were male. 84.8% of cases were due to blunt force trauma. The median ISS score was 9 (IQR: 1-16). Of cases where 30-day survival was available, there was a 6.1% 30-day mortality rate (11 cases with missing survival data). 17.8% of participants received intensive care as their highest level of care.

At least one OFI was found in 358 of cases (6.7%). In cases where death occurred within 30 days, at least one OFI was found in 0.2% of cases. The median scores for TRISS, NORMIT, and PS were 0.98 (IQR: 0.95-0.99), 0.96 (IQR: 0.85-0.99), and 0.99 (IQR: 0.96-0.99) respectively.

6.3 Main results

The fitted curves for the logistic regression models of each score are shown in figure 3. The odds ratios for TRISS, NORMIT and PS were 0.35 (95% CI: 0.22-0.54, p: 0.0001), 0.41 (95% CI: 0.3-0.57, p: 0.000007), and 0.38 (95% CI: 0.23-0.64, p: 0.002) respectively.
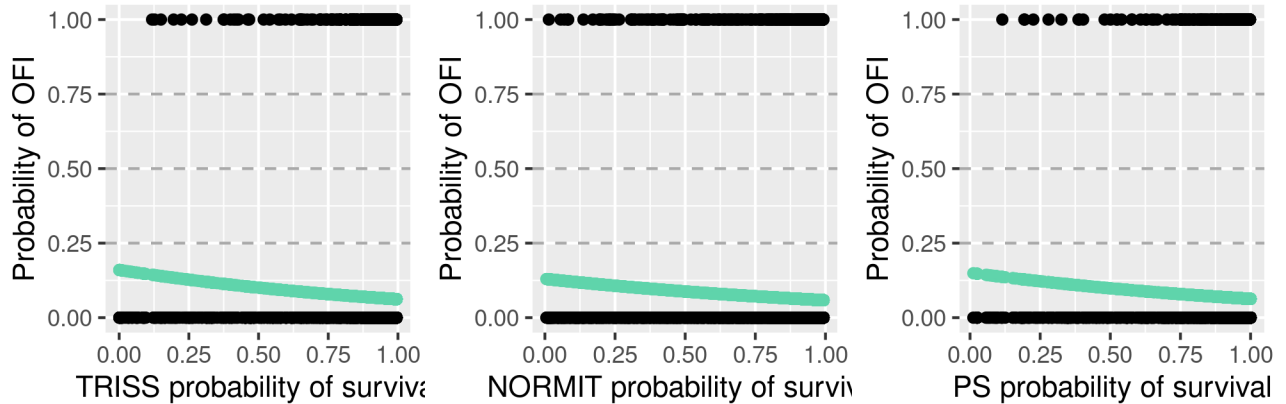

Figure 3: Fitted curves for the logistic regression models of each score

The receiver operating curves for all scores are plotted in figure 4. The areas under the curve for TRISS, NORMIT and PS were 0.67 (95% CI: 0.65-0.7), 0.66 (95% CI: 0.63-0.68), and 0.65 (95% CI: 0.62-0.67)

Table 2: Table 2. Characteristics of participants.

|	Opportunity for improvement	No opportunities for improvement	Total
	(N=358)	(N=4962)	(N=5320)
Age — yrs			
Mean (SD)	47.9 (21.7)	44.7 (21.3)	44.9 (21.3)
Median [Min, Max]	47.0 [15.0, 97.0]	42.0 [15.0, 100]	42.0 [15.0, 100]
Sex			
Male	259 (72.3%)	3390 (68.3%)	3649 (68.6%)
Female	99 (27.7%)	1572 (31.7%)	1671 (31.4%)
ISS			
Mean (SD)	17.9 (10.2)	10.2 (11.1)	10.7 (11.2)
Median [Min, Max]	17.0 [0, 66.0]	8.00 [0, 75.0]	9.00 [0, 75.0]
NISS			
Mean (SD)	22.9 (12.8)	13.4 (13.9)	14.1 (14.0)
Median [Min, Max]	22.0 [0, 75.0]	9.00 [0, 75.0]	10.0 [0, 75.0]
Dominating Type of Injury			
Blunt	313 (87.4%)	4200 (84.6%)	4513 (84.8%)
Penetrating	45 (12.6%)	762 (15.4%)	807 (15.2%)
GCS			
Mean (SD)	13.6 (2.92)	13.7 (3.00)	13.7 (2.99)
Median [Min, Max]	15.0 [3.00, 15.0]	15.0 [3.00, 15.0]	15.0 [3.00, 15.0]
Respiratory Rate			
Mean (SD)	23.6 (19.5)	23.2 (19.6)	23.2 (19.6)
Median [Min, Max]	18.0 [9.00, 99.0]	18.0 [3.00, 99.0]	18.0 [3.00, 99.0]
Systolic blood pressure (mmHg)			
Mean (SD)	136 (29.0)	137 (26.8)	137 (26.9)
Median [Min, Max]	135 [0, 237]	136 [0, 285]	136 [0, 285]
ASA score			
1	188 (52.5%)	2966 (59.8%)	3154 (59.3%)
2	99 (27.7%)	1188 (23.9%)	1287 (24.2%)
3	70 (19.6%)	767 (15.5%)	837 (15.7%)
4	1 (0.279%)	41 (0.826%)	42 (0.789%)
Revised Trauma Score			
Mean (SD)	7.44 (1.05)	7.48 (1.04)	7.47 (1.04)
Median [Min, Max]	7.84 [0.872, 7.84]	7.84 [0.872, 7.84]	7.84 [0.872, 7.84]
TRISS			
Mean (SD)	0.895 (0.151)	0.929 (0.155)	0.926 (0.155)
Median [Min, Max]	0.950 [0.118, 0.994]	0.981 [0.000569, 0.995]	0.980 [0.000569, 0.995]
NORMIT			
Mean (SD)	0.796 (0.233)	0.854 (0.236)	0.850 (0.236)
Median [Min, Max]	0.905 [0.0139, 0.993]	0.964 [0.00505, 0.993]	0.962 [0.00505, 0.993]
PS			
Mean (SD)	0.916 (0.141)	0.940 (0.140)	0.938 (0.140)
Median [Min, Max]	0.974 [0.115, 0.998]	0.991 [0.0117, 1.00]	0.990 [0.0117, 1.00]

respectively.
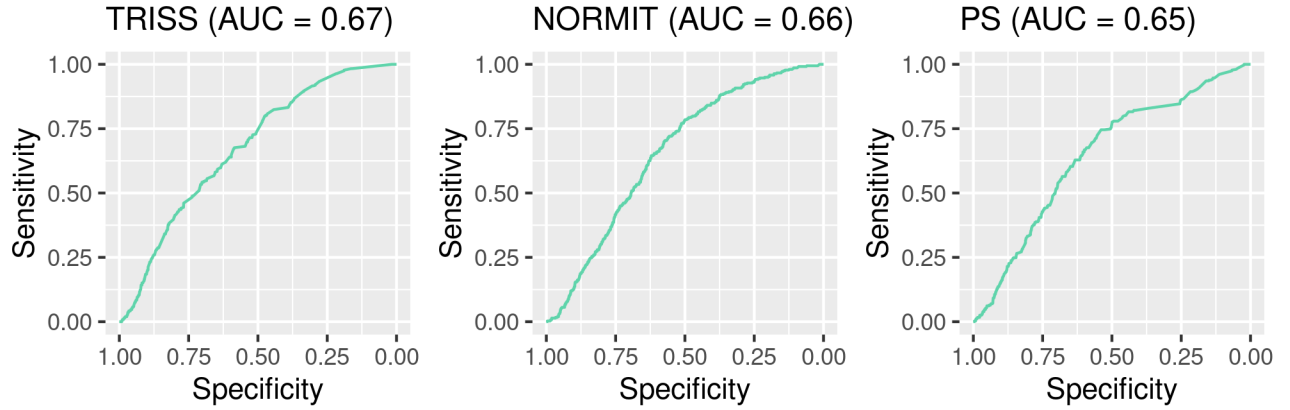
Figure 4: Receiver operating curves for each score

The integrated calibration index for TRISS, NORMIT and PS was 0.86 (95% CI: 0.85-0.87), 0.78 (95% CI: 0.78-0.79 and 0.87 (95% CI: 0.86-0.88 respectively. The accuracy at all cutoffs is plotted for all scores in figure 5. The accuracy at a cutoff of 0.5 for TRISS, NORMIT and PS was 0.1 (95% CI: 0.094-0.1), 0.15 (95% CI: 0.14-0.16) and 0.095 (95% CI: 0.089-0.1) respectively.
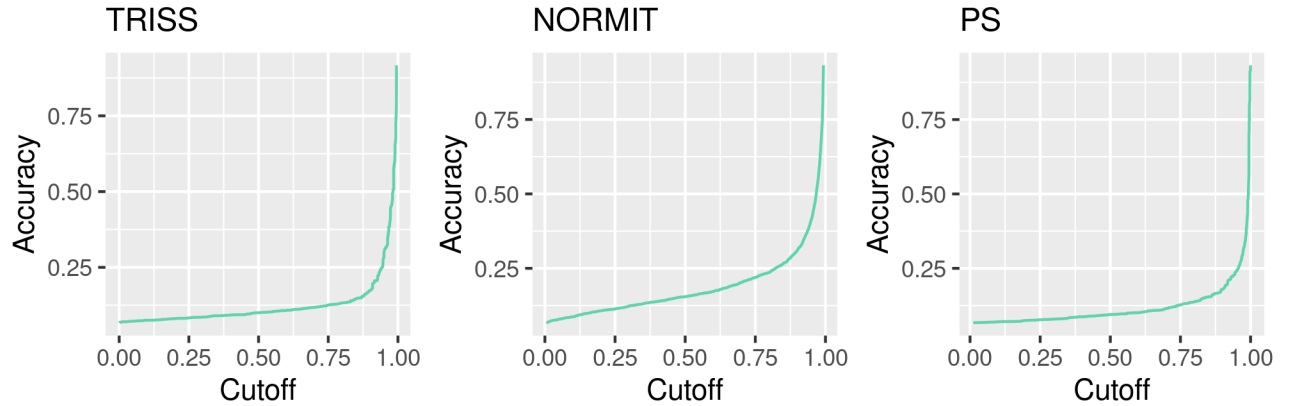

Figure 5: Accuracy at all cutoffs for each score

The differences between the scores for each result are shown in table 3. TRISS was superior to the other models with regards to the AUC. NORMIT was superior to the other models with regards to accuracy and to the ICI.

7 Discussion

7.1 Key results

This aim of this study was to evaluate whether several commonly used trauma severity models (TRISS, NORMIT, and PS) are associated with, and to what extent they can predict, OFI, with the goal of improving case selection for multidisciplinary morbidity and mortality conferences. The three models investigated did not differ significantly in their results. The main findings indicate that increased probability of survival was

Table 3: Comparison of model results

	AUC difference (95% CI)	ICI difference (95% CI)	Accuracy difference (95% CI)
TRISS - NORMIT	0.018 (0.0026-0.033)	0.08 (0.08-0.08)	-0.05 (-0.05–0.05)
TRISS - PS	0.028 (0.014-0.043)	-0.01 (-0.01–0.01)	0.005 (0.005-0.005)
NORMIT - PS	0.011 (-0.0024-0.024)	-0.09 (-0.09–0.09)	0.055 (0.055-0.055)

associated with lower odds for OFI, with odds ratios ranging from 0.35 to 0.41. The AUC values were almost identical for all three models (range: 0.65-0.67), indicating poor diagnostic ability in discriminating between OFI and no OFI. Calibration was also low for all models (ICI range: 0.78 - 0.87). Accuracy at a cutoff of 0.5 was low across the board (range: 0.095-0.15).

7.2 Strengths and limitations

One of the strengths of this study is the duration of enrolment and Karolinska University Hospital's catchment population of 3 million, which allowed for a large study population. The outcome data is strengthened by Karolinska University Hospital's extensive quality review process involving personnel from all concerned specialities and professions. All patients that met inclusion criteria were included in the study. Data collection, patient inclusion and data analysis was prepared on scrambled data by a single student research before being given access to the real data, in order to minimise researcher bias.

There were several limitations to this study. One of them is that selection of cases for review at conferences was in large part based upon the use of audit filters. There is as yet no widespread consensus on which audit filters are the most effective at identifying cases and this is an ongoing topic of research. The choice of audit filters used affects case selection and in turn introduces an element of selection bias into this study. It is possible, and perhaps even likely, that some cases with opportunities for improvement were never selected for case review and therefore weren't available for inclusion in this study. Another limitation is that, due to the long duration of enrolment in this study, the composition of the staff present at conferences will have differed over time. Determining whether OFI exists or not is an essentially subjective evaluation and the conclusions reached by any given conference may vary depending on which individuals are present. It is also possible that standards and expectations of care have changed over time as knowledge and technology improves, which would also in turn affect assignment of OFI.

7.3 Interpretation

Our results show an association between lower probability of survival and higher odds for assignment of OFI. This result should be evaluated in the context of potential selection bias, as described above. Cases with low trauma severity are less likely to trigger audit filters and are in turn less likely to be reviewed at M&M conferences, which will lead to severe trauma being over-represented in OFI assignment. However, it is unlikely that this selection bias is the main cause of the association found. Severe trauma is much more complex and time-dependent than mild trauma. Complexity and time pressure are two factors that increase the risk of management errors, and management errors have been found to be the main cause of preventable deaths (10–12). In this context, the association between trauma severity and OFI has an obvious causality and our findings on this association are unsurprising.

Despite this association, none of the trauma severity models we investigated are able to reliably predict OFI. The area under the curve of the receiver operating curves - a value that describes the ability of a model to discriminate between two outcomes - was around 0.65 for all three scores, which indicates that these models were not able to reliably predict OFI. We also showed that the models were poorly calibrated as to the outcomes. Intuitively, this makes sense, as these models are designed to predict death and not errors in care. These models do not take into account factors that might lead to suboptimal care - they are based on physiological and anatomical parameters that quantify the severity of an injury. Even though a higher trauma severity is associated with greater odds for OFI, these models cannot discriminate between a high trauma severity with OFI and one that has no OFI.

We showed that all three models have a low accuracy at a 0.5 cutoff. One reason for the low accuracy is the low incidence of OFI, which means that maximal accuracy is found when assigning "no OFI" to all cases, although this would obviously lead to false negatives. Either way, whether you use maximum accuracy or accuracy at a 0.5 cutoff as a metric, none of these models have significant accuracy. In real-world terms, there is no decision-making cutoff for predicted survival that would lead to an improvement in case selection - these approaches would lead to either a high number of false positives or false negatives.

These three models had essentially the same results, which is an expected outcome considering that they are all tools for estimating the same thing - the probability of survival. In addition, the models have a great degree of overlap in the parameters they incorporate, including GCS and age. The models we analysed are commonly-used and are sourced from three different geographical populations. The fact that they all had similar performance is a sign that investigation of other models is unlikely to lead to different results.

The results of this study agree with previous studies assessing the predictive power of trauma severity models with regards to preventable deaths. Although this study was not specifically about preventable deaths, the factor that determines whether a death was preventable is whether an improvement in care could have prevented the death, i.e. whether there was a significant opportunity for improvement, which makes the previous studies on preventable deaths broadly relevant to this study. Ghorbani et al found that both TRISS and NORMIT had poor performance in predicting preventable deaths (10). Heim et al found that PS and TRISS miss a significant proportion of preventable deaths (9). Our results also show that none of the three models investigated (TRISS, NORMIT and PS) were able to reliably predict OFI. The difference between our study and prior studies is that we looked at the trauma population across the board and not only deaths.

Even though our results were expected, it was still important to investigate this question scientifically, because trauma severity models are in use today as audit filters at many institutions. Our findings show that the use of trauma severity models as audit filters is unlikely to lead to improvements in case selection, and in fact could be harmful. Audit filters that do not improve case selection will either lead to cases being over-selected, causing resource wastage, or to cases being under-selected, leading to OFI being missed. Trauma severity models are useful when used on the population level, to allow institutions to compare themselves and to provide a framework for discussing trauma severity. They are much less useful when applied to individuals, and were not intended for that use.

7.4 Generalisability

This study is based on a patient population at a level 1 trauma centre in Stockholm, Sweden and the results are likely applicable for other level 1 trauma centres in the Nordic countries, as these countries share similarities in patient demographics, comorbidities, and injury types. This study is less generalisable for smaller trauma centres as those centres tend to manage less severe trauma and will have a different patient population. This study also has less external validity with regards to international populations that differ significantly from the one in this study, for example populations with a higher proportion of penetrating trauma.

7.4.1 Health equity

The demographical characteristics of the participants in this study reflect the well-known demographics of trauma patients worlwide. 68.6% of participants were male and the median age was 42, which matches with statistics both in Sweden and globally (2). The predominance of male patients invites one to question if female patients are less likely to receive quality care. In healthcare as in all areas, repetition leads to mastery of a skill, so an onlooker might be tempted to assume that the high volume of male patients in trauma will lead healthcare providers to superiority iin treating males, whilst treatment of female patients will be neglected. This demographical difference has precipitated research that seeks to investigate whether this presumption holds true, and several large multi-centre studies have been published that evaluated whether there is a correlation between sex and outcomes. None of these studies found a correlation beween sex and increased mortality or other serious outcomes (32–35), and in fact one study in the Netherlands found that female sex was correlated with decreased mortality (34).

This study has not compared differences between the sexes, and therefore cannot comment on any relation between sex and OFI at this centre. One of the trauma severity models we investigated, PS, does include patient sex as a coefficient in predicting mortality, but we did not examine whether patient sex affects the ability of PS to predict OFI. However, there is another currently ongoing project at this hospital that compares OFI between the male and female sex cohorts. The uneven distribution of sex in this study does not make for a limitation of this study in terms of generalisability since our demographics match the distribution of sex in trauma cases both globally and in the Swedish trauma population.

Another aspect of health equity is patient age. We know that peer review panels are more likely to classify trauma deaths as non-preventable in the geriatric population (36). This is unlikely to be purely due to discrimination, since several factors make the elderly less likely to survive severe trauma, including a higher prevalence of comorbidities and in general lower margins for tolerating extreme physiological stress (37). In fact, all three trauma severity models investigated in this study include coefficients for age, since age is a well-known predictor of mortality. Nevertheless, there is a possibility that healthcare providers are less likely to look for errors in older patients due to a predisposition to finding older patients less likely to be savable. The question of whether age is correlated with OFI was outside the scope of this study, but as with sex, is being investigated in other currently ongoing projects at this institution. The geriatric population differs from the average adult in terms of injury types, injury response, and management. This, in combination with worse trauma outcomes in the geriatric population, should prompt focus on this cohort in future research since this group has perhaps the most to gain from improvements in trauma care, especially improvements that recognise the unique challenges in this population and aim to specifically tailor their care to those challenges.

7.5 Future research

Future studies should seek to evaluate the predictive power of other audit filters that are currently used for case selection. As mentioned previously, trauma centres today use a large variety of audit filters and many of these have not yet been supported by research. Future studies could use logistic regression to identify factors associated with OFI, and investigate if any of these factors have predictive power. The ultimate goal of such a study study, as with this study, would be to improve case selection for review in conferences. As these conferences are the gold standard of trauma improvement programmes, improved case selection has the potential to directly improve trauma care, in addition to reducing the resource burden of identifying cases and holding said conferences.

In addition, future studies should aim to confirm the findings of this study within a different population. The population in this study is a typically Scandinavian one, with a low incidence of penetrating trauma. As mentioned previously, the external validity of trauma severity models is in some degree dependent on the similarity of the populations they were developed from. Ideally a new study would investigate the predictive power of the models used in this study, in a country with a higher incidence of penetrating trauma and potentially other different population characteristics, in order to confirm the results found in this study. Such a study could use a study design similar to this one, i.e. use logistic regression and measures of discrimination and calibration.

8 Conclusion

This aim of this study was to evaluate whether several commonly used trauma severity models (TRISS, NORMIT, and PS) are associated with, and to what extent they can predict, opportunities for improvement, with the goal of improving case selection for multidisciplinary morbidity and mortality conferences. We showed that, while low probability of survival is associated with increased odds of OFI, none of the models were able to reliably predict OFI. The use of these models as audit filters in case selection for M&M conferences is not supported by our results and may lead to over-selection.

References

1.   Gerdin M. The risk of dying: Predicting trauma mortality in urban indian hospitals [Internet]. Karolinska Institutet; 2015. Available from: http://hdl.handle.net/10616/44832

2.   Global burden of 369 diseases and injuries in 204 countries and territories, 1990: A systematic analysis for the Global Burden of Disease Study 2019. Lancet (London, England). 2020 Oct;396(10258):1204–22.

3.   WHO. Injuries and violence [Internet]. 2021 [cited 2022 Aug 30]. Available from: https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence

4.   Wihlke G, Strömmer L, Troëng T, Brattström O. Long-term follow-up of patients treated for traumatic injury regarding physical and psychological function and health-related quality of life. European Journal of Trauma and Emergency Surgery. 2021 Feb;47(1):129–35.

5.   UN General Assembly. Transforming our world : The 2030 agenda for sustainable development [Internet]. [cited 2022 Nov 2]. Available from: https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf

6.   United Nations: Department of Economic and Social Affairs. The sustainable development goals report 2020. New York, NY: United Nations; 2020.

7.   Hemmila MR, Cain-Nielsen AH, Jakubus JL, Mikhail JN, Dimick JB. Association of Hospital Participation in a Regional Trauma Quality Improvement Collaborative With Patient Outcomes. JAMA surgery. 2018 Aug;153(8):747–56.

8.   Hakkenbrak NAG, Mikdad SY, Zuidema WP, Halm JA, Schoonmade LJ, Reijnders UJL, et al. Preventable death in trauma: A systematic review on definition and classification. Injury. 2021 Oct;52(10):2768–77.

9.   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 Sep;47(9):1960–5.

10.   Ghorbani P, Strömmer L. Analysis of preventable deaths and errors in trauma care in a scandinavian trauma level-i centre. Acta Anaesthesiologica Scandinavica [Internet]. 2018;62(8):1146–53. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/aas.13151

11.   Montmany S, Pallisera A, Rebasa P, Campos A, Colilles C, Luna A, et al. Preventable deaths and potentially preventable deaths. What are our errors? Injury. 2016 Mar;47(3):669–73.

12.   Teixeira PGR, Inaba K, Hadjizacharia P, Brown C, Salim A, Rhee P, et al. Preventable or Potentially Preventable Mortality at a Mature Trauma Center. Journal of Trauma and Acute Care Surgery. 2007 Dec;63(6):1338–47.

13.   World Health Organization. Guidelines for trauma quality improvement programmes. World Health Organization; 2009.

14.   Stelfox HT, Straus SE, Nathens A, Gruen RL, Hameed SM, Kirkpatrick A. Trauma Center Quality Improvement Programs in the United States, Canada, and Australasia. Annals of Surgery. 2012 Jul;256(1):163–9.

15.   Santana MJ, Stelfox HT, Panel on behalf of the TQIC. Development and Evaluation of Evidence-Informed Quality Indicators for Adult Injury Care. Annals of Surgery. 2014 Jan;259(1):186–92.

16. Slater N, Sekhon P, Bradley N, Shariff F, Bedford J, Wong H, et al. Morbidity and mortality conferences in general surgery: A narrative systematic review. Canadian Journal of Surgery. 2020 Jun;63(3):E211–22.

17. Senkowski CK, McKenney MG. Trauma Scoring Systems: A Review. J Am Coll Surg. 1999;189(5):13.

18. Wisner DH. History and Current Status of Trauma Scoring Systems. Archives of Surgery. 1992 Jan;127(1):111–7.

19. Skaga NO, Eken T, Steen PA. Assessing Quality of Care in a Trauma Referral Center: Benchmarking Performance by TRISS-Based Statistics or by Analysis of Stratified ISS Data? Journal of Trauma and Acute Care Surgery. 2006 Mar;60(3):538–47.

20. Gabbe BJ, Cameron PA, Wolfe R. TRISS: Does It Get Better than This? Academic Emergency Medicine. 2004;11(2):181–6.

21. Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score. The Journal of Trauma. 1987 Apr;27(4):370–8.

22. Schluter PJ, Nathens A, Neal ML, Goble S, Cameron CM, Davey TM, et al. Trauma and Injury Severity Score (TRISS) coefficients 2009 revision. The Journal of Trauma. 2010 Apr;68(4):761–70.

23. Bouamra O, Wrotchford A, Hollis S, Vail A, Woodford M, Lecky F. A new approach to outcome prediction in trauma: A comparison with the TRISS model. Journal of Trauma and Acute Care Surgery [Internet]. 2006 Sep [cited 2022 Nov 15];61(3):701–10. Available from: https://journals.lww.com/jtrauma/Abstract/2006/09000/A_New_Approach_to_Outcome_Prediction_in_Trauma___A.30.aspx

24. Jones JM, Skaga NO, Søvik S, Lossius HM, Eken T. Norwegian survival prediction model in trauma: Modelling effects of anatomic injury, acute physiology, age, and co-morbidity. Acta Anaesthesiologica Scandinavica. 2014 Jan;58(3):303–15.

25. Osler T, Baker SP, Long W. A Modification of the Injury Severity Score That Both Improves Accuracy and Simplifies Scoring. Journal of Trauma and Acute Care Surgery. 1997 Dec;43(6):922–6.

26. Skaga NO, Eken T, Søvik S. Validating performance of TRISS, TARN and NORMIT survival prediction models in a Norwegian trauma population. Acta Anaesthesiologica Scandinavica. 2018;62(2):253–66.

27. Ghorbani P, Troëng T, Brattström O, Ringdal KG, Eken T, Ekbom A, et al. Validation of the Norwegian survival prediction model in trauma (NORMIT) in Swedish trauma populations. The British Journal of Surgery. 2020 Mar;107(4):381–90.

28. Karolinska University Hospital. Karolinska university hospital becomes sweden's first trauma centre [Internet]. 2020 [cited 2022 Sep 9]. Available from: https://www.karolinska.se/en/karolinska-university-hospital/news/2020/12/karolinska-university-hospital-becomes-swedens-first-trauma-centre

29. Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the Trauma Score. The Journal of Trauma. 1989 May;29(5):623–9.

30. The Trauma Audit and Research Network. TARN - PS12 - Calculations. https://www.tarn.ac.uk/Content.aspx?c=1895;

31. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/

32. Croce MA, Fabian TC, Malhotra AK, Bee TK, Miller PR. Does gender difference influence outcome? The Journal of Trauma. 2002 Nov;53(5):889–94.

33.   Magnotti LJ, Fischer PE, Zarzaur BL, Fabian TC, Croce MA. Impact of gender on outcomes after blunt injury: A definitive analysis of more than 36,000 trauma patients. Journal of the American College of Surgeons. 2008 May;206(5):984-991; discussion 991-992.

34.   Pape M, Giannakópoulos GF, Zuidema WP, de Lange-Klerk ESM, Toor EJ, Edwards MJR, et al. Is there an association between female gender and outcome in severe trauma? A multi-center analysis in the Netherlands. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine. 2019 Feb;27(1):16.

35.   Trentzsch H, Nienaber U, Behnke M, Lefering R, Piltz S. Female sex protects from organ failure and sepsis after major trauma haemorrhage. Injury. 2014 Oct;45 Suppl 3:S20–28.

36.   Radke OC, Heim C. Recognizing Preventable Death: Is There a Role of Survival Prediction Algorithms? Anesthesiology Clinics. 2019 Mar;37(1):1–1.

37.   Bardes JM, Benjamin E, Schellenberg M, Inaba K, Demetriades D. Old Age With a Traumatic Mechanism of Injury Should Be a Trauma Team Activation Criterion. The Journal of Emergency Medicine. 2019 Aug;57(2):151–5.