Validating performance of TRISS, TARN and NORMIT survival prediction models in a Norwegian trauma population

N. O. Skaga[1,2], T. Eken[1,2,3] and S. Søvik[3,4]

[1]Division of Emergencies and Critical Care, Department of Anaesthesiology, Oslo University Hospital Ullevål, Oslo, Norway
[2]Division of Emergencies and Critical Care, Oslo University Hospital Trauma Registry, Oslo University Hospital Ullevål, Oslo, Norway
[3]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
[4]Department of Anaesthesia and Critical Care, Akershus University Hospital, Lørenskog, Norway

Correspondence
N. O. Skaga, Division of Emergencies and Critical Care, Department of Anaesthesiology, Oslo University Hospital Ullevål, PO box 4956 Nydalen, NO-0424 Oslo, Norway
E-mail: noskaga@online.no

Location: Patients were recruited at Oslo University Hospital Ullevål, Oslo, Norway.

The first two authors contributed equally to this work.

Conflicts of interest
The authors declare that they have no conflicts of interest.

Funding
The study was funded by the authors' institutions.

Submitted 4 September 2017; accepted 11 October 2017; submission 22 February 2017.

Citation
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

doi: 10.1111/aas.13029

Introduction: Anatomic injury, physiological derangement, age, injury mechanism and pre-injury comorbidity are well-founded predictors of trauma outcome. Statistical prediction models may have poorer discrimination, calibration and accuracy when applied in new locations. We aimed to compare the TRISS, TARN and NORMIT survival prediction models in a Norwegian trauma population.

Methods: Consecutive patients admitted to Oslo University Hospital Ullevål within 24 h after injury, with Injury Severity Score $\geq$ 10, proximal penetrating injuries, or received by trauma team, were studied. Original NORMIT coefficients were updated in a *derivation* dataset (NORMIT 2; n = 5923; 2005–2009). TRISS, TARN and NORMIT prediction models were evaluated in the *validation* dataset (n = 6348; 2010–2013) using two different AIS editions for injury coding. Exclusion due to missing data was 0.26%. Outcome was 30-day mortality. Validation included AUROC, scaled Brier statistics, and calibration plots.

Results: The NORMIT models had significantly better discrimination, calibration, and overall fit than the TRISS 09, TARN 09 and TARN 12 models. The updated NORMIT 2 had higher numerical values of AUROC and scaled Brier than the original NORMIT, but with overlapping 95%CI. Overlapping 95%CI for AUROCs and Discrimination slopes indicated that the TARN and TRISS models performed similarly. Calibration plots showed tight and consistent predictions over all *P*s strata for NORMIT 2 run on AIS'98 coded data, and only little deterioration when AIS'08 data was substituted.

Conclusions: In a Norwegian trauma population, the updated Norwegian survival prediction model in trauma (NORMIT 2) performed better than well-established British and US alternatives. External validation of these three models in other Nordic populations is warranted.

Editorial comment
Prognostic scoring systems are established in trauma care worldwide. In this retrospective study of a large regional cohort of trauma patients from the Oslo area in Norway, a locally developed score (NORMIT) was found to perform better than traditional scoring methods.

During the past decade, mortality rates following major trauma have steadily declined in patients admitted to our institution.[1,2] However, comparison of crude mortality rates without adjusting for patient risk profiles is of limited value.[3] Robust statistical prediction models are needed to benchmark treatment performance, as suboptimal or inappropriately applied models may yield misleading estimations.

Outcome following injury results from many factors. A statistical prediction model of good quality and with high face validity includes all important variables that significantly affect survival, well defined and appropriately modelled. The prediction model must fit the data well, must display good discrimination between survivors and non-survivors, and show adequate calibration over the whole spectrum of survival predictions. The aim is that such risk models should adjust for all sources of variation that are institution independent, so the residual effect represents deviations in the quality of care compared to the average care in the trauma system and population where the model was derived. Relatively few studies have evaluated the performance of multiple prediction models in the same trauma population.[4]

Quality of care assessment in trauma has traditionally used a methodology that determined the *statistical significance* of the difference between observed and predicted outcomes for a specific hospital.[5] A problem with this methodology is that the predictive power (discrimination, calibration and accuracy) of the model used to assign the risk score for each trauma victim may be low in the new location. This can result in erroneous conclusions on performance.[6] Importantly, increasing the number of patients in the study population will not increase the accuracy. The *magnitude* of the difference in performance between institutions is quantified by the W statistic,[7] which expresses the difference between predicted and observed survival rates per 100 patients. A further development is the W*s* statistic,[8] which is standardized with respect to injury severity case mix.

Several prediction models for survival after trauma exist.[9–13] The TRISS model (Trauma Score Injury Severity Score) was developed in 1987[9] and has been in worldwide use since. The TRISS coefficients were last updated in 2010.[12] The UK Trauma Audit and Research Network (TARN) database was the foundation for the national UK trauma prediction model. The TARN Ps04 model, derived from a huge dataset, was launched in 2006[10] after studies had revealed that TRISS performed unsatisfactorily in the UK trauma system. Specifically, TRISS exclusion rates were high due to patients with missing physiological variables.[10] The TARN model has since been regularly updated.[14,15]

In view of the spectrum of injury mechanisms with particularly few penetrating injuries and the widespread advanced pre-hospital physician-manned Emergency Medical Systems (EMS) in the Nordic countries, we introduced the Norwegian prediction Model in Trauma (NORMIT) in 2014,[16] derived from injury data obtained August 2000 through July 2006. NORMIT addressed several weaknesses we had experienced with the TRISS and TARN models.

The aim of the present study was (1) to generate an updated version of the NORMIT model (NORMIT 2) based on injury data from January 2005 through December 2009, and (2) to perform temporal validation of the NORMIT 2 model and external validation of several editions of the US and UK trauma survival prediction models, in a dataset from January 2010 through December 2013. Each model was evaluated using data coded with two different editions of the prevailing injury scoring system. We aimed to adhere to the TRIPOD Guidelines in our reporting.[17]

Patients and methods

Population and study participants

Oslo University Hospital Ullevål (OUH-U) is the major trauma hospital for more than 660,000 citizens and the trauma referral centre for 2.8 million people. Currently, approximately 1850 patients are enrolled in the hospital based trauma registry (OUH-TR) each year, including nearly 270 children < 16 years old. Over the study inclusion period, the yearly number of patients registered in the database increased from 1028 in 2005 to 1784 in 2013. Nearly 40% of the trauma victims had severe injury, i.e., ISS ≥ 16.[7] On average, 90% suffered from blunt injury and 10% from penetrating injury.

This retrospective, non-interventional trial was based on anonymized registry data only. The Oslo University Hospital Data Protection Officer, in this matter representing the Regional Committee for Medical and Health Research Ethics and the Norwegian Data Protection Authority, therefore considered the study exempt from patient consent requirements (H. Thorstensen, 26 March 2014).

Inclusion and exclusion criteria

We studied the OUH-TR population from 1 January 2005 to 31 December 2013. Eligible[16] were all patients received at OUH-U with trauma team activation (TTA), except those suffering from medical conditions or injuries unsuited for Abbreviated Injury Scale (AIS) scoring (i.e., drowning, hypothermia, asphyxia, spontaneous subarachnoid haemorrhage, or cardiac arrest, entered in OUH-TR because of TTA). Included were also all patients with documented anatomical injury ISS ≥ 10 according to the Abbreviated Injury Scale 1990, Update 1998 (AIS'98),[18] and/or with head injuries scored as AIS ≥ 3, and/or with penetrating injuries towards the head, neck, torso, and/or proximal to elbow or knee irrespective of ISS. Patients with an isolated single extremity fracture were excluded unless the trauma team was activated. All patients declared dead on arrival (DOA) according to the Utstein template definition[19] were included.

Coding, data extraction and outcome assessment

Anatomical injury was classified according to AIS'98, and from year 2009 also according to AIS edition 2005, Update 2008 (AIS'08).[20] Injury Severity Score (ISS)[21] and New Injury Severity Score (NISS)[22] were calculated for data derived with both AIS coding editions.

Physiological derangement on arrival was classified according to the Triage Revised Trauma Score (T-RTS).[16,23] The T-RTS range (0–12) is defined as the sum of the clinical category values of Glasgow Coma Scale (GCS) score, Systolic Blood Pressure (SBP), and Respiratory Rate (RR) (Table 1). Such scoring of physiological data into clinical categories, based on information from text in addition to numerical raw data, substantially reduces the number of patient exclusions due to missing data.[16] For patients arriving at OUH-U intubated and in general anaesthesia, GCS and RR were scored based on values documented immediately prior to intubation. In cases of missing T-RTS data elements, all available information in the patient records was used to estimate pre-intubation RTS clinical category. To avoid positive performance biasing, the value closest to normal was chosen when doubt existed. Normal values were used as final default.[16]

Outcome was defined as survival or death 30 days after injury, independent of whether the patient at that point was admitted or discharged from hospital. Survival was verified from the Norwegian Population Registry. Foreign citizens repatriated alive to their home country earlier than 30 days after injury (37 patients, 0.3% of the total material) were defined as survivors.

Data were coded and extracted from OUH-TR by registrars who were certified nurse anaesthetists with trauma team experience, formally trained in injury coding according to AIS'98 and AIS'08. Before data extraction, all data elements were thoroughly screened for inconsistencies and non-logical values, in compliance with the OUH-TR data validation protocol.

Table 1 Clinical categories for the Revised Trauma Score (RTS) elements constituting the Triage-RTS (T-RTS).

RTS category scale	Respiratory rate	Systolic blood pressure	GCS score
4	10–29 (normal)	> 89 (good radial pulse)	13–15
3	> 29 (fast)	76–89 (weak radial pulse)	9–12
2	6–9 (slow)	50–75 (femoral pulse)	6–8
1	1–5 (gasp)	1–4 (only carotid pulse)	4–5
0	0 (no respiration)	0 (no carotid pulse)	3

Clinical categories for the RTS elements, from Pillgram-Larsen J, Initial treatment of the severely injured at Ullevål hospital, May 1999. Triage RTS (T-RTS) is defined as the sum of a patient's three RTS clinical category values and thus ranges 0–12.

Comparison of prediction models

The NORMIT model was developed based on injury data from OUH-U in the period 1 August 2000 through 31 July 2006.[16] First, the original NORMIT model coefficients[16] were updated in the *derivation* dataset ($n = 5923$), comprising the patients admitted 1 January 2005 through 31 December 2009. The new NORMIT 2 coefficients were derived with injury data coded with AIS'98. The three prediction models of study were then evaluated in the *validation* dataset ($n = 6348$), comprising patients admitted 1 January 2010 through 31 December 2013.

Each model was assessed with its most recently updated regression coefficients from where the model was derived. Table 2 lists the evaluated survival prediction models, their derivation datasets, and the AIS editions ('98 or '08) used during derivation. Trauma databases worldwide currently vary regarding which AIS edition is employed. Therefore, to 'stress' the prediction models, the performance of each model was evaluated both with injury data coded according to the AIS edition the model was originally derived with, and with the other AIS edition.

Statistical methods

Prediction model performance is routinely evaluated by measures of discrimination, calibration and overall accuracy.

Discrimination between survivors and non-survivors in each model was evaluated by calculation of the area under the receiver operating characteristic curve (AUROC). AUROC is mathematically equivalent to the *c* statistic, which denotes the proportion of all possible pairs of patients drawn from the population, one a survivor and one a non-survivor, where the patient who survived had the higher *P*s. AUROC with 95% confidence intervals (95% CI) for all models were compared, and non-overlapping 95% CIs were considered a significant difference in discrimination ability. We calculated the *discrimination slope* with 95% CI for each model, i.e., the absolute difference between the mean predicted probability of survival (*P*s) for survivors and for non-survivors.[24] We also calculated the *median P*s with 95% CI for survivors and for non-survivors, as the distributions of *P*s values were highly skewed.

Calibration was assessed through calibration plots, which show the fraction of patients who actually survived for each decile of predicted survival.[24,25]

Overall model performance was evaluated with the scaled Brier score.[26] This is a quadratic scoring rule, and it is analogous to R^2.[24,27,28] First, the differences between each patient's predicted probability of survival (*P*s) and that patient's observed outcome (*D*; survival = 1, non-survival = 0) are squared. The mean of these squared differences in the entire population is divided by the mean squared difference between observed outcome and predicted outcome for a totally uninformed model, i.e., a model where the outcome for all individuals is

Table 2 Trauma survival prediction models evaluated.

Prediction model	Parent model	Updated coefficients	Derivation dataset	Derivation Injury coding system
NORMIT 2	NORMIT*	Yes†	OUH–TR 2005–2009	AIS Edition 1990 Update 98
TRISS 09	TRISS‡	Yes§	US NTDB 2002–2006	AIS Edition 1990 Update 98
TARN 12	TARN *Ps* 04¶	Yes**	UK TARN 2005–2010	AIS Edition 2005 Update 08
TARN 09	TARN *Ps* 04¶	Yes††	UK TARN 2002–2008	AIS Edition 2005 Update 08

NORMIT, Norwegian prediction Model in Trauma; NORMIT 2, update in present study; TRISS, Trauma Score Injury Severity Score; TRISS 09, updated 2009; TARN, UK Trauma Audit and Research Network prediction model; TARN 12, updated 2012; TARN 09, updated 2009; TARN Ps04, original model; OUH–TR, Oslo University Hospital Trauma Registry; US NTDB, US National Trauma Data Bank; AIS, Abbreviated Injury Score; *Reference 16, †Present study, ‡Reference 7, §Reference 12, ¶ Reference 10, **Reference 15, ††Reference 14.

predicted to be equal to the average outcome in that population ($\bar{p}$):

$$\mathrm{sBrier} = 1 - \frac{\frac{1}{n} \times \sum_{i=1}^{n} (D_i - Ps_i)^2}{\frac{1}{n} \times \sum_{i=1}^{n} (D_i - \bar{p})^2}$$

The scaled Brier score is independent of the prevalence of the outcome in the population.[24,27,29]

Assessment of importance of the individual predictor variables in the NORMIT 2 model was performed as variance-based sensitivity analysis. Importance indices were constructed from observed combinations of factor values, since predictor variables were generally correlated.

Data analysis was undertaken using JMP 11.2.1 (SAS Institute, Cary, NC, USA). AUROCs were calculated with the ROC-Curve & partial Area Under the Curve Analysis JMP add-in module created by Sebastian Hoffmeister (https://community.jmp.com/docs/DOC-7500). Bootstrap 95% CIs (1000 repetitions) were reported for AUROCs. The significance level was set to 0.05.

Results

Of the 12,303 trauma patients included in the OUH-TR during the study period, 32 patients (11 in the derivation dataset and 21 in the validation dataset; 8 non-survivors) were excluded due to lack of complete registrations for validation of all prediction models. The final study population, with complete information on all prognostic and calculated variables of interest, formed the derivation (n = 5923) and validation (n = 6348) datasets. For population characteristics, see Table 3a and b.

In the *validation* dataset (Table 3b), 71% were males, 90% had blunt injury, 36% had severe injury defined as ISS ≥ 16,[7] and 95.5% survived to 30 days post-injury. Sixty-four per cent of the patients were primary admissions to OUH-U, and 12% were intubated and in general anaesthesia prior to arrival.

Updated NORMIT 2 coefficients

The derivation dataset (Table 3a) was utilized to generate updated coefficients for the predictors in the original NORMIT model: NISS, T-RTS, age represented as an upward-slanting cubic function, *pre-injury* ASA-PS (American Association of Anesthesiologists Physical Status classification system) score indicating comorbidity on a four-level ordinal scale (there were no pre-injury ASA5 or ASA6 patients), and an interaction between NISS and pre-injury ASA-PS. All predictors in the model were highly significant ($P < 0.001$). The estimated relative importance of the individual predictor variables in NORMIT 2 were: T-RTS 0.398, NISS 0.306, pre-injury ASA-PS score 0.262, and age 0.157. The NORMIT 2 model, i.e., with the updated coefficients, is shown in Fig. 1.

Model performance

Table 4 summarizes performance measures for the studied trauma survival prediction models. Judged by non-overlapping 95% CIs, the NORMIT and NORMIT 2 models showed better discrimination between survivors and non-survivors than the TRISS and TARN models, both as measured by AUROC and by Discrimination slopes. The NORMIT models also showed best overall fit measured by higher scaled Brier scores.

NORMIT 2 had higher numerical values of AUROC and scaled Brier score than the original NORMIT model, though AUROC 95% CIs overlapped. 'Stressing' NORMIT 2 by using injury data coded with AIS'08 did not result in poorer performance; the scaled Brier score actually increased.

Overlapping 95% CIs for both AUROCs and Discrimination slopes indicated that the performance of the TARN and TRISS models were not significantly different. However, scaled Brier scores indicated that TRISS 09, especially run on AIS'08 injury data, had better overall fit than TARN 12 and TARN 09.

Visual inspection of calibration plots (Fig. 2) showed tight and consistent predictions over all *P*s strata for NORMIT 2 run on AIS'98 data. Use of AIS'08 data resulted in only slightly larger deviations between predicted and observed survival. TARN 12 run on AIS'08 data (for which it was derived) showed poorer calibration, especially in *P*s bands 0.2–0.4, and overall was pessimistically biased. TRISS 09 run on AIS'08

Table 3 Population characteristics. (a) Derivation dataset; (b) Validation dataset.

	Total *N*	*N* Dead	% Mortality	OR	95% CI for OR	*P*
(a) Derivation dataset						
Overall	5923	368	6.21			
Injury Mechanism						
Blunt	5403	342	6.33			
Penetrating	520	26	5.00	0.779	0.517–1.173	0.25
Age						
< 55 years	4582	183	3.99			
≥ 55 years	1341	185	13.80	3.847	3.105–4.767	**
Gender						
Male	4309	248	5.76			
Female	1614	120	7.43	1.315	1.049–1.649	< 0.02
ASA-PS score						
ASA 1	4305	178	4.13			
ASA 2	1017	70	6.88	1.714	1.281–2.269	**
ASA 3	535	87	16.26	4.503	3.411–5.907	**
ASA 4	66	33	50.00	23.19	13.96–38.52	**
Intubated						
Not intubated	4238	76	1.79			
Intubated in ER	697	92	13.20	8.328	6.08–11.44	**
Arrived intubated	988	200	20.24	13.90	10.61–18.39	**
RR RTS Category						
4	5549	253	4.56			
3	203	22	10.84	2.544	1.565–3.946	*
2	71	14	19.72	5.141	2.722–9.087	**
1	32	15	46.88	18.47	9.015–37.47	**
0	68	64	94.12	334.9	136.99–1108	**
SBP RTS Category						
4	5567	251	4.51			
3	153	15	9.80	2.302	1.279–3.854	*
2	137	43	31.39	9.688	6.558–14.12	**
1	21	18	85.71	127.1	42.67–544.9	**
0	45	41	91.11	217.1	86.83–726.6	**
GCS RTS Category						
4	4648	78	1.68			
3	438	33	7.53	4.774	3.101–7.196	**
2	384	61	15.89	11.06	7.75–15.74	**
1	168	49	29.17	24.13	16.1–35.95	**
0	285	147	51.58	62.41	45.37–86.54	**
ISS						
1–8	2025	7	0.35			
9–15	1418	31	2.19	6.443	3.000–15.97	**
16–24	1156	45	3.89	11.68	5.603–28.43	**
25–34	884	154	17.42	60.82	30.61–143.9	**
35–49	302	72	23.84	90.25	44.00–217.8	**
50–75	138	59	42.75	215.3	101.6–531.3	**
NISS						
1–8	1935	7	0.36			
9–15	924	19	2.06	5.782	2.531–14.8	**
16–24	954	19	1.99	5.60	2.45–14.37	**

Table 3 (Continued)

	Total *N*	*N* Dead	% Mortality	OR	95% CI for OR	*P*
25–34	1117	51	4.57	13.18	6.373–31.93	**
35–49	485	64	13.20	41.87	20.4–101.0	**
50–75	508	208	40.94	191.0	96.05–452.19	**
(b) Validation dataset						
Overall	6348	287	4.52			
Injury Mechanism						
Blunt	5738	260	4.53			
Penetrating	610	27	4.43	0.976	0.651–1.464	0.91
Age						
< 55 years	4770	118	2.47			
≥ 55 years	1578	169	10.71	4.729	3.71–6.027	**
Gender						
Male	4499	210	4.67			
Female	1849	77	4.16	0.887	0.68–1.159	0.38
ASA-PS score						
ASA 1	4163	114	2.74			
ASA 2	1344	45	3.35	1.230	0.859–1.734	0.25
ASA 3	766	92	12.01	4.848	3.635–6.452	**
ASA 4	75	36	48.00	32.79	20.05–53.57	**
Intubated						
Not intubated	5018	67	1.34			
Intubated in ER	576	63	10.94	9.075	6.35–12.96	**
Arrived intubated	754	157	20.82	19.43	14.49–26.34	**
RR RTS Category						
4	5889	180	3.06			
3	312	23	7.37	2.524	1.571–3.877	**
2	65	22	33.85	16.23	9.358–27.42	**
1	19	9	47.37	28.55	11.21–71.66	**
0	63	53	84.13	168.10	87.86–355.7	**
SBP RTS Category						
4	6007	184	3.06			
3	158	20	12.66	4.586	2.73–7.332	**
2	130	36	27.69	12.12	7.951–18.14	**
1	17	12	70.59	75.95	27.86–240.7	**
0	36	35	97.22	1107.6	237.5–19,725	**
GCS RTS Category						
4	5387	57	1.06			
3	338	36	10.65	11.15	7.176–17.11	**
2	257	36	14.01	15.23	9.756–23.51	**
1	124	31	25.00	31.17	19.07–50.27	**
0	242	127	52.48	103.3	72.23–149.4	**
ISS						
1–8	2602	8	0.31			
9–15	1462	17	1.16	3.815	1.692–9.369	**
16–24	1083	35	3.23	10.83	5.271–25.19	**
25–34	814	131	16.09	62.19	32.36–138.9	**
35–49	277	47	16.97	66.26	32.69–153.0	**
50–75	110	49	44.55	260.5	124.7–616.1	**

Table 3 (Continued)

	Total *N*	*N* Dead	% Mortality	OR	95% CI for OR	*P*
NISS						
1–8	2505	8	0.32			
9–15	1019	13	1.28	4.033	1.695–10.22	**
16–24	932	17	1.82	5.799	2.57–14.25	**
25–34	966	33	3.42	11.04	5.342–25.77	**
35–49	443	35	7.90	26.78	12.98–62.48	**
50–75	483	181	37.47	187.1	97.43–417.6	**

ASA-PS, American Association of Anestehsiologists' Physical status Score; RTS, Revised trauma score; RR, respiratory rate; SBP, Systolic blood pressure; GCS, Glascow Coma Scale score; ISS, Injury Severity Score; NISS, New Injury Severity Score; 95% CI, 95% confidence interval; Fisher's Exact test. ******< 0.0005, *****< 0.01.

$$P_S = \frac{1}{1 + e^{-b}}$$

$$b = (0.5562 \times \text{T-RTS}) - 4.3234 \times \left(\frac{\text{age} + 1}{100}\right)^3 + \begin{cases} \text{ASA1:} & (-0.0713 \times \text{NISS}) + 0.6266 \\ \text{ASA2:} & (-0.0565 \times \text{NISS}) - 0.2142 \\ \text{ASA3:} & (-0.0487 \times \text{NISS}) - 0.8971 \\ \text{ASA4:} & (-0.0081 \times \text{NISS}) - 3.8748 \end{cases}$$

Fig. 1. The NORMIT 2 trauma survival model equation, i.e., with updated coefficients. Predicted probability of survival for an individual trauma victim is calculated by inserting the patient's T-RTS value, age, and NISS value. The adequate NISS expression is selected depending on the patient's pre-injury ASA-PS classification. *Ps*, Probability of survival; T-RTS, Triage Revised Trauma Score; age, years; ASA1, ASA2, ASA3 and ASA4, individual pre-injury American Society of Anesthesiologists Physical Status Classification System (ASA-PS) categories; NISS, New Injury Severity Score.

data (for which it was derived) showed better calibration than TARN 12, but with high variability between *P*s bands.

Discussion

This study performed in the same dataset an external validation of the TRISS and TARN models with their most recent regression coefficients and a temporal validation of the NORMIT 2 model. Temporal validation represents a prospective evaluation of a model, is independent of the original data, and may thus be considered an external validation in time.[30] In this Norwegian population, the NORMIT 2 survival prediction model had significantly better discrimination, calibration, and overall fit than the TRISS and TARN models. NORMIT 2 may therefore supplement the long-standing TRISS model when trauma care is evaluated within the Nordic countries. The face validity of NORMIT 2 is high, as its predictors are few and intuitive: Anatomical injury (NISS), physiological derangement on arrival (T-RTS), age, and a four-level comorbidity scale (pre-injury ASA-PS score).

Validation measures of model performance

Discrimination is the ability of a prediction model to separate subjects with and without the outcome of study.[29] A much used measure is the area under the receiver operating characteristic curve (AUROC), which describes how well the model rank-orders survivors and non-survivors.[31] Both original NORMIT and NORMIT 2 had significantly higher AUROCs than the British and US models.

Importantly, the AUROC, or *c* statistic, is not a function of the actual magnitude of the predicted probabilities,[32] since in its calculation any patient pair where the survivor has a higher *P*s than the non-survivor is considered a 'concordant pair'. Improving a model so that it assigns survivors somewhat higher *Ps*'s and non-survivors somewhat lower *Ps*'s will not improve the AUROC unless the correct *P*s calculations result in a higher proportion of concordant pairs. AUROC may therefore be less sensitive than measures based on likelihood ratio tests or other global measures of fit[32] and will not necessarily detect small differences in discriminative ability between two models.

Table 4 Trauma survival prediction models: performance measures.

Prediction model and AIS Edition	AUROC	Scaled Brier score	Discrimination slope	Median *P*s Non-Survivors	Median *P*s Survivors
NORMIT 2 AIS'08*	0.979 (0.974–0.985)	0.526	0.523 (0.487–0.560)	0.409 (0.356–0.492)	0.997 (0.997–0.997)
NORMIT 2 AIS'98	0.977 (0.972–0.983)	0.505	0.536 (0.501–0.572)	0.388 (0.329–0.461)	0.996 (0.996–0.997)
NORMIT AIS'98	0.973 (0.968–0.980)	0.428	0.578 (0.542–0.614)	0.317 (0.263–0.391)	0.998 (0.997–0.998)
TRISS 09 AIS'08*	0.956 (0.948–0.967)	0.383	0.399 (0.362–0.435)	0.601 (0.563–0.711)	0.994 (0.993–0.994)
TRISS 09 AIS'98	0.950 (0.941–0.963)	0.344	0.410 (0.373–0.448)	0.583 (0.509–0.674)	0.992 (0.992–0.993)
TARN 12 AIS'08	0.952 (0.944–0.963)	0.288	0.399 (0.367–0.430)	0.565 (0.536–0.627)	0.992 (0.992–0.992)
TARN 12 AIS'98*	0.947 (0.939–0.958)	0.212	0.407 (0.367–0.439)	0.559 (0.491–0.617)	0.991 (0.990–0.992)
TARN 09 AIS'08	0.952 (0.943–0.962)	0.238	0.443 (0.410–0.475)	0.502 (0.455–0.552)	0.987 (0.987–0.988)
TARN 09 AIS'98*	0.946 (0.937–0.957)	0.119	0.452 (0.420–0.485)	0.469 (0.404–0.539)	0.987 (0.986–0.987)

Numbers in brackets are 95% Confidence intervals (95% CI). The Scaled Brier score is a sum-of-squares R^2 statistic. The Discrimination slope is the difference between the mean values of *P*s among survivors and among non-survivors; 95% CI's assuming unequal variances (Welch *t*-test).
***Model 'stressed' by using injury data coded with an AIS Edition different from the one that the model was derived with.

A majority of trauma patients do not have life-threatening injuries and are easy to predict as survivors. Trauma prediction models therefore often have higher AUROC values than e.g. predictive laboratory tests. A better indicator of high discriminating ability may therefore be less overlap in *P*s values between trauma survivors and non-survivors.[24] For all evaluated models, we found very high mean and median *P*s values among trauma survivors (Table 4 and Fig. S1). Overall lower *P*s values among non-survivors indicated that the NORMIT models showed better discrimination than the TRISS and TARN models. Discrimination slope analyses substantiated this finding.

Calibration denotes agreement between survival predictions and observed outcomes over the full span of probabilities.[29] Since the lightly and the very severely injured patients are easier to predict as survivors and non-survivors, respectively, high model performance in the mid-bands of *P*s strata distinguishes a well-calibrated prediction model. Visual inspection of calibration plots from this Norwegian dataset indicated that NORMIT 2 had better calibration than the newest TRISS and TARN models, tested both with AIS'98 and AIS'08 coded data (Fig. 2).

The original NORMIT model and the TRISS 09 model recently underwent external validation in a Finnish dataset.[33] This study was in agreement with our present findings, i.e., that the original NORMIT had better discriminative ability than TRISS 09, but calibration was unsatisfactory for both models as predictions were too pessimistic. A possible contribution to this is that the original NORMIT was derived on a dataset from the 6-year period starting August 2000. Risk adjusted mortality at OUH-U has declined markedly from late 2004, primarily due to a sudden survival improvement in patients having at least one AIS 5 injury in the head/neck region.[2] Not unexpectedly, we found that NORMIT 2 with its regression coefficients derived from 2005 to 2009 showed improved calibration compared to the original NORMIT (Table 4).

Overall model performance was evaluated using the scaled Brier score, which is analogous to R^2. This measure most markedly differentiated the various prediction models (Table 4). In the

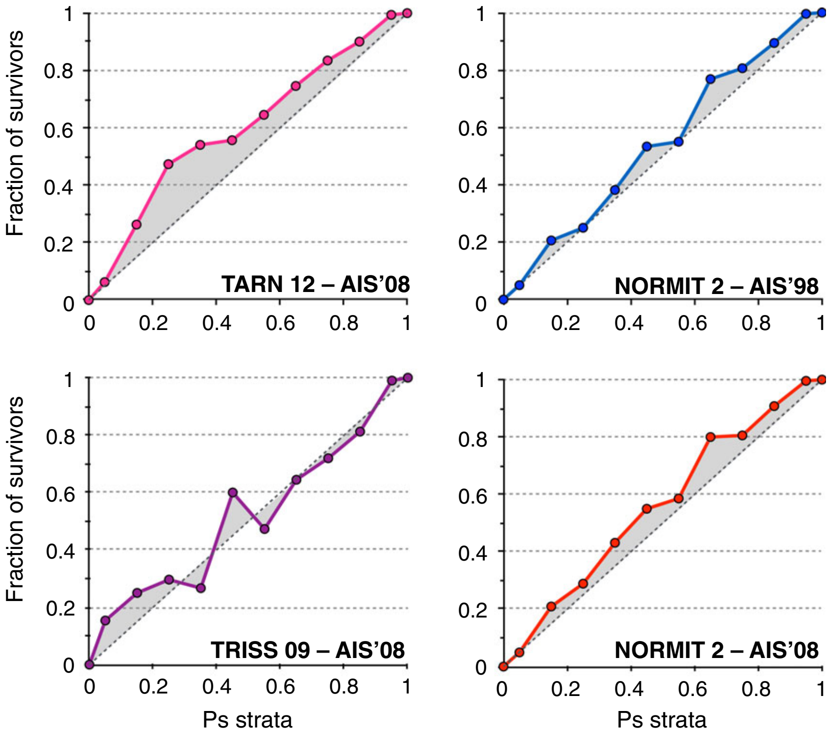

Fig. 2. Calibration plots for the most recent TARN and TRISS models, and the updated NORMIT 2 model with data based on AIS'98 and AIS'08. Observed survival for patients in each decile of predicted survival is plotted against predicted survival. Larger deviations from the line of unity denote poorer model calibration. Note pessimistic biasing in the lower *P*s strata for TARN 12 and variable predictions for TRISS 09. AIS'08, AIS edition 2005, Update 2008; AIS'98, AIS edition 1990, Update 1998; TARN 12, update TARN Ps12; TRISS 09, update TRISS 2009. [Colour figure can be viewed at wileyonlinelibrary.com]

scaled Brier score, every patient contributes with the squared difference between their outcome (0 = died, 1 = survived) and predicted probability of survival (a number between 0.0 and 1.0). Inaccurate models and models where specific patient groups are systematically mispredicted (high *P*s but dies, or low *P*s but survives) will have large squared errors for these patients, and a correspondingly low scaled Brier score (low 'R^2') indicating low overall fit.

The updated coefficients in NORMIT 2, reflecting the improved survival in trauma patients admitted to OUH-U,[2] probably resulted in more correct *P*s values in our validation dataset. NORMIT 2 was derived within a single trauma system, with cooperating hospitals and an extensive EMS system including anaesthesiologist-manned cars and helicopters delivering advanced emergency care at the site of injury and during patient transport. Variation in trauma outcome may have been larger in the huge derivation populations of the TARN and TRISS models. This may have resulted in larger deviations between *P*s values and outcomes, and thus poorer scaled Brier scores when TARN and TRISS models were applied on our dataset.

Because the expected mortality rate after various anatomical injuries has declined, in the AIS 2005 Update 2008 the severity grades of several injuries have been downscaled relative to that in the AIS'98 edition. Thus, many patients will receive lower injury severity score when coded according to AIS'08, reflecting the lower expected mortality rate for a given injury due to improved treatment of trauma patients. More realistic injury coding probably caused the improved scaled Brier scores found when the TRISS and NORMIT models, derived with AIS'98 data, were 'stressed' with AIS'08 data. The TARN models also showed much better overall fit with AIS'08 coded data than with AIS'98 data (Table 4).

Model characteristics possibly affecting performance

Outcome in this study was survival to 30 days post-injury, irrespective of hospitalization. The NORMIT models were derived for this WHO-recommended variable, which is independent of hospital transfer and discharge practices. In contrast, TRISS coefficients were derived from survival data evaluated at discharge from the trauma centre, while the TARN models used survival at discharge from hospital, or at 30 days for patients still hospitalized. Both strategies generate positive bias. Thus, neither the TRISS nor TARN models were tuned for the outcome measure used in the Nordic countries.

Dead on arrival (DOA) patient definition and whether DOA patients are excluded from analyses are crucial sources of bias. In Norwegian hospitals, all patients admitted to the ED are registered as admitted to the hospital. This is unlike the US system, where trauma patients treated and dying in the ER may never be 'admitted' to the hospital and therefore not included in any trauma registry. TRISS is based on such data.[12] Exclusion of DOA patients results in optimistic biasing. Our dataset included all patients classified as DOA according to the Utstein template,[19] suiting the NORMIT and TARN models, which had coefficients derived with DOA patients included.

Injury mechanism is fundamental to TRISS, which is in fact four distinct prediction models, with separate full coefficient sets for blunt and penetrating injury in both adult and paediatric patients. The original TRISS derivation population, the foundation for the 1990 TRISS coefficients, had more than three times as many penetrating traumas as OUH-TR (32.0%[7] vs. 9.4%), while the 2009 TRISS revision dataset had a prevalence of penetrating trauma (11.9%) more comparable to ours.[12] Paediatric penetrating trauma was too infrequent in the 2009 derivation dataset to obtain stable coefficients,[12] and a formal solution to this problem has to our knowledge not been published. Although a survival prediction model is designed to adjust for case mix, it is unclear whether the differences between the US and our dataset could have affected TRISS model performance. TARN and NORMIT models are not adjusted for injury mechanism.

Anatomical injury is represented by ISS in the TRISS and TARN models, and by NISS in NORMIT and NORMIT 2. We have previously shown that NISS had better predictive power than ISS in our trauma system.[16] The NISS would be superior to ISS in patients with several severe injuries in a single body compartment, e.g., penetrating injuries towards the torso or both blunt and penetrating head injury.[34–38] In all, 27% of patients in our validation dataset had head injury with AIS severity 3–6.

Physiological derangement on admission is represented by clinical categories for GCS, SBP and RR in both TRISS and NORMIT, but the scores for the categories are weighted separately in TRISS whereas their sum is used in the NORMIT models (Table 1). In contrast, the TARN models use GCS only. The latter strategy could lead to poorer predictions in e.g. patients in circulatory shock but still with normal or near-normal GCS.[39]

Patients with missing physiological values, e.g. those intubated before admission, were until recently excluded from TRISS derivation data. This is clearly suboptimal, since physiological data are not missing at random: In the derivation dataset for TRISS 09, almost one in five patients had incomplete RTS scoring, and those with incomplete information were more likely to die than the average trauma registry population.[6,12,40,41] Statistical imputation of missing physiological values was therefore employed for the derivation of TRISS 09. However, no official recommendation was given for how to score actual cases to obtain a probability of survival when physiological data is missing. The current TARN model does not include SBP and RR, and patients with missing GCS values due to intubation are handled similar to patients with GCS close to 4–5, regardless of the reason for intubation. In contrast, NORMIT models use actual pre-hospital data for T-RTS calculations in these cases, greatly reducing patient exclusion and not assuming artificially low GCS values e.g. in patients who are intubated before helicopter transport due to severe pain.

Age and comorbidity affect trauma survival independently,[16] but the studied prediction models differ markedly regarding these factors. TRISS employs full separate sets of coefficients for paediatric patients (< 14 years) and dichotomizes age (< 55 or ≥ 55 years) in adult patients, thus predicting an identical age-related reduction in trauma survival in a 56-year-old and an 86-year-old patient. Furthermore, TRISS has no mechanism to account for patient comorbidity. TARN 09 and TARN 12 adjust elaborately for increasing age, with an eight-level categorical variable additionally used in an interaction effect with gender. Comorbidity is however unaccounted for. These factors may have reduced TRISS's and TARN's ability to correctly predict non-survival in e.g. younger but sicker patients and resulted in poorer overall model fit. The recently published TARN Ps14 model uses a five-level ordinal scale for grouping of 21 categories of comorbidities represented by a modified version of the Charlson Comorbidity Index (mCCI).[42] Retrospective scoring of this large number of comorbidities would be impractical, maybe impossible, in large datasets. Also, the mCCI does not differentiate between light and severe cases of the same disease.

The NORMIT models adjust for increasing age with a single continuous, slowly upward-slanting cubic function, while functional limitation caused by any systemic disease is represented on a four-level comorbidity scale (ASA-PS). We believe this strategy gains important predictive power in an ageing trauma population, with increasing prevalence of e.g., cardiovascular and pulmonary disease and cancer treatment sequelae.

Study limitations

This was a single-institution study, external validation on data from other trauma centres is warranted. Our findings pertain to trauma populations and systems similar to those in the Nordic countries; the evaluated prediction models could thus perform differently in other settings. Specifically, our cohort had only 9% penetrating injury, and very few gunshot and blast injuries. Trauma data came from a single institution. This somewhat limits the generalizability of our study, though it ensured homogenous coding and a very low rate of missing values. Importantly, these prognostic models are designed to be used for institution benchmarking on a population level and should not be employed for prognostication in individual patients.

Conclusion

In our Norwegian trauma population, the NORMIT 2 survival prediction model with updated coefficients displayed very good calibration and significantly better discrimination and overall fit than the TRISS and TARN prediction models. Injury spectrum, pre-hospital treatment, patient inclusion, trauma scoring, and age and comorbidity classification all affect the performance of prediction models used outside their derivation population. Though TRISS will still be highly useful for international comparisons, NORMIT 2 may be well suited for evaluation of trauma care in the Nordic countries. However, external validation together with the most recent TRISS and TARN models is warranted.

Acknowledgements

We thank trauma registrars Morten Hestnes and Hans Johansson for their continuing efforts at the Oslo University Hospital Trauma Registry.

Authors' contributions

N. O. S. and T. E.: designed and built the Oslo University Hospital Trauma Registry.

N. O. S., S. S. and T. E.: planned and designed the study and carried out the statistical analyses.

N. O. S. and S. S.: drafted the manuscript.

S. S.: created the tables and figures.

All authors critically evaluated and discussed the ongoing analyses, critically revised the manuscript, and approved the final version.

References

1. Groven S, Eken T, Skaga NO, Roise O, Naess PA, Gaarder C. Long-lasting performance improvement after formalization of a dedicated trauma service. J Trauma 2011; 70: 569–74.
2. Søvik S, Skaga NO, Hanoa R, Eken T. Sudden survival improvement in critical neurotrauma: an exploratory analysis using a stratified statistical process control technique. Injury 2014; 45: 1722–30.
3. Roques F, Nashef SA, Michel P, Gauducheau E, de Vincentiis C, Baudet E, Cortina J, David M, Faichney A, Gabrielle F, Gams E, Harjula A, Jones MT, Pintor PP, Salamon R, Thulin L. Risk factors

and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg 1999; 15: 816–22.

4. de Munter L, Polinder S, Lansink KW, Cnossen MC, Steyerberg EW, de Jongh MA. Mortality prediction models in the general trauma population: a systematic review. Injury 2017; 48: 221–9.
5. Flora JD Jr. A method for comparing survival of burn patients to a standard survival curve. J Trauma 1978; 18: 701–5.
6. Clark DE. Comparing institutional trauma survival to a standard: current limitations and suggested alternatives. J Trauma 1999; 47: S92–8.
7. Champion HR, Copes WS, Sacco WJ, Lawnick MM, Keast SL, Bain LW Jr, Flanagan ME, Frey CF. The Major Trauma Outcome Study: establishing national norms for trauma care. J Trauma 1990; 30: 1356–65.
8. Hollis S, Yates DW, Woodford M, Foster P. Standardized comparison of performance indicators in trauma: a new approach to case-mix variation. J Trauma 1995; 38: 763–6.
9. Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score. J Trauma 1987; 27: 370–8.
10. 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. J Trauma 2006; 61: 701–10.
11. Huber-Wagner S, Lefering R, Qvick LM, Korner M, Kay MV, Pfeifer KJ, Reiser M, Mutschler W, Kanz KG, Working Group on Polytrauma of the German Trauma Society. Effect of whole-body CT during trauma resuscitation on survival: a retrospective, multicentre study. Lancet 2009; 373: 1455–61.
12. Schluter PJ, Nathens A, Neal ML, Goble S, Cameron CM, Davey TM, McClure RJ. Trauma and Injury Severity Score (TRISS) coefficients 2009 revision. J Trauma 2010; 68: 761–70.
13. Lefering R, Huber-Wagner S, Nienaber U, Maegele M, Bouillon B. Update of the trauma risk adjustment model of the TraumaRegister DGU: the Revised Injury Severity Classification, version II. Crit Care 2014; 18: 476.
14. Procedures manual 2009. Manchester, UK: The Trauma Audit & Research Network, 2009.
15. Procedures manual 2016. Manchester, UK: The Trauma Audit & Research Network, 2016.
16. 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 Anaesthesiol Scand 2014; 58: 303–15.
17. Moons KG, Altman DG, Reitsma JB, Collins GS, Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Development Initiative. New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: the TRIPOD statement. Adv Anat Pathol 2015; 22: 303–5.
18. The Abbreviated Injury Scale 1990 revision – Update 98. Des Plains, IL: Association for the Advancement of Automotive Medicine, 1998.
19. Ringdal KG, Coats TJ, Lefering R, Di Bartolomeo S, Steen PA, Røise O, Handolin L, Lossius HM. The Utstein template for uniform reporting of data following major trauma: a joint revision by SCANTEM, TARN, DGU-TR and RITG. Scand J Trauma Resusc Emerg Med 2008; 16: 7.
20. The Abbreviated Injury Scale 2005 revision – Update 2008. Des Plains, IL: Association for the Advancement of Automotive Medicine, 2008.
21. Baker SP, O'Neill B, Haddon W Jr, Long WB. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. J Trauma 1974; 14: 187–96.
22. Osler T, Baker SP, Long W. A modification of the injury severity score that both improves accuracy and simplifies scoring. J Trauma 1997; 43: 922–5.
23. Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the Trauma Score. J Trauma 1989; 29: 623–9.
24. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010; 21: 128–38.
25. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol 2012; 12: 82.
26. Steyerberg E. Clinical prediction models. Berlin, Germany: Springer, 2009.
27. Wu YC, Lee WC. Alternative performance measures for prediction models. PLoS ONE 2014; 9: e91249.
28. Hu B, Palta M, Shao J. Properties of R^2 statistics for logistic regression. Stat Med 2006; 25: 1383–95.
29. Austin PC, Steyerberg EW. Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. Stat Med 2013; 32: 661–72.
30. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009; 338: b605.

31. Harrell FJ. Regression modeling strategies. Cham, Switzerland: Springer International Publishing AG, 2015.
32. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007; 115: 928–35.
33. Raj R, Brinck T, Skrifvars MB, Handolin L. External validation of the Norwegian survival prediction model in trauma after major trauma in Southern Finland. Acta Anaesthesiol Scand 2016; 60: 48–58.
34. Osler TM, Cohen M, Rogers FB, Camp L, Rutledge R, Shackford SR. Trauma registry injury coding is superfluous: a comparison of outcome prediction based on trauma registry International Classification of Diseases-Ninth Revision (ICD-9) and hospital information system ICD-9 codes. J Trauma 1997; 43: 253–6.
35. Sullivan T, Haider A, DiRusso SM, Nealon P, Shaukat A, Slim M. Prediction of mortality in pediatric trauma patients: new injury severity score outperforms injury severity score in the severely injured. J Trauma 2003; 55: 1083–7.
36. Lavoie A, Moore L, LeSage N, Liberman M, Sampalis JS. The New Injury Severity Score: a more accurate predictor of in-hospital mortality than the Injury Severity Score. J Trauma 2004; 56: 1312–20.
37. Frankema SP, Steyerberg EW, Edwards MJ, van Vugt AB. Comparison of current injury scales for survival chance estimation: an evaluation comparing the predictive performance of the ISS, NISS, and AP scores in a Dutch local trauma registration. J Trauma 2005; 58: 596–604.
38. Lefering R. Development and validation of the Revised Injury Severity Classification Score for severely injured patients. Eur J Trauma Emerg Surg 2009; 35: 437–47.
39. Moore L, Lavoie A, Abdous B, Le Sage N, Liberman M, Bergeron E, Emond M. Unification of the revised trauma score. J Trauma 2006; 61: 718–22.
40. 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? J Trauma 2006; 60: 538–47.
41. Kirkham JJ. A comparison of hospital performance with non-ignorable missing covariates: an application to trauma care data. Stat Med 2008; 27: 5725–44.
42. Bouamra O, Jacques R, Edwards A, Yates DW, Lawrence T, Jenks T, Woodford M, Lecky F. Prediction modelling for trauma using comorbidity and ‘true’ 30-day outcome. Emerg Med J 2015; 32: 933–8.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site:

Fig. S1. Distribution of predicted probability of survival (*P*s) from the NORMIT 2, TRISS 09 and TARN 12 trauma survival prediction models, plotted for survivors and non-survivors at 30 days post-injury.