External validation of the Norwegian survival prediction model in trauma (NORMIT) in two Swedish trauma populations

Poya Ghorbani[1] (PG), Thomas Troëng[2] (TT), Olof Brattström[3] (OB), Kjetil Gorseth Ringdal[4,5] (KGR), Torsten Eken[6,7] (TE), Anders Ekbom[8] (AE), Lovisa Strömmer[1] (LS)

[1]Division of Surgery, Department of Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institute, Stockholm, Sweden

[2]Section of Vascular Surgery, Department of Surgical Sciences, Uppsala University, Uppsala, Sweden

[3]Section of Anaesthesiology and Intensive Care Medicine, Department of Physiology and Pharmacology, Karolinska Institute, Stockholm, Sweden

[4]Norwegian National Trauma Registry, Oslo University Hospital, Oslo, Norway

[5]Department of Anaesthesiology, Vestfold Hospital Trust, Tønsberg, Norway

[6]Department of Anaesthesiology, Division of Emergencies and Critical Care, Oslo University Hospital Ullevål, Oslo, Norway

[7]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

[8]Department of Medicine, Karolinska University Hospital – Solna, Stockholm, Sweden

CORRESPONDENCE

Poya Ghorbani, MD, PhD

Address: P9:03, Department of Surgical Gastroenterology, Karolinska University Hospital – Solna, SE-171 76 Stockholm, Sweden

Tel.: +46 (0) 76-2002066

E-mail: poya.ghorbani@sll.se

The authors have no conflicts of interest to declare.

This is an orginal article.

The paper is not based on a previous communication to a society or meeting.

The authors did not preregister the research in an independent, institutional registry.

ABSTRACT

BACKGROUND Trauma survival prediction models can be used for quality assessment in trauma populations. The Norwegian survival prediction model in trauma (NORMIT) has recently been updated and internally validated (NORMIT 2). The aim of this observational study was to compare the accuracy of NORMIT 1 and 2 in two Swedish trauma populations.

METHODS Eligible were adult patients registered in the national trauma registry during 2014-2016. The study populations consisted of (1) the total national trauma (NT) population, and (2) a subpopulation of patients admitted to a single Level I trauma centre (TC). The primary outcome was 30-day mortality. Model validation included receiver operating characteristic curves and GiViTI calibration belts. The calibration was also assessed in subgroups of severely injured patients (New Injury Severity Score [NISS] >15).

RESULTS 26504 patients were included. Exclusion due to missing data was 18.7% in the NT (n=21554) and 2.6% in the TC (n=3972) population. NORMIT 1 and 2 showed excellent ability to distinguish between survivors and non-survivors in both populations, but poor agreement between predicted and observed outcome in the NT population with overestimation of survival including in the subgroup of NISS >15. In the TC subpopulation, NORMIT 1 underestimated survival irrespective of injury severity, but NORMIT 2 showed good calibration both in the total subpopulation and the NISS >15 subgroup.

CONCLUSION NORMIT 2 is well suited to predict survival in a Swedish trauma centre population irrespective of injury severity, but both models perform poorly in a more heterogeneous national trauma population.

KEY WORDS Survival Analysis; Wounds and Injuries; Trauma Centers; Mortality; Validation studies; Registries; Area under Curve; Calibration.

INTRODUCTION

Trauma quality assessment and improvement are important components in all trauma systems and trauma centres[1]. The quality of care including mortality should be monitored continuously within the same institution over time and compared to other institutions[2]. However, comparisons within and between institutions must be adjusted for case-mix[3] and therefore several mortality prediction models have been developed for this purpose[4-8].

An accurate trauma prediction model serves dual purposes. Firstly, it is necessary to have a tool that can be used to follow expected vs. actual survival in each individual hospital over time. Secondly, a trauma prediction model is of great value and use in trauma performance quality improvement programmes, such as the US TQIP[2]. In clinically relevant subgroups (e.g. penetrating torso trauma, multisystem or single system blunt trauma, isolated traumatic brain injury (TBI) and geriatric trauma patients) mortality can be estimated and root causes of unexpected high mortality between institutions or over time, identified and corrected.

The most widely used model for the calculation of the probability of survival (Ps) in trauma populations is the TRISS methodology[4, 9] introduced in the 1980s based on a North American trauma population. Because of the widespread use of TRISS, many studies, including from the authors own institution (Karolinska University Hospital – Solna [KUH])[10], have identified several major limitations of the method[11-13]. One important limitation is that the application of the TRISS model on trauma populations other than the one from which the model was derived may result in selection differences[14]. Further, outcome is scored as dead or alive at end of acute care, i.e. an administrative time point; age is categorised into three groups with cut-off at 16 and 55 years; and physiology is scored only in the emergency department, effectively excluding patients who arrive intubated – a group with generally worse prognosis. The resulting models for different age groups and mechanisms of injury

contain 24 coefficients[7]. The UK Trauma Audit and Research Network (TARN)[5] has developed an alternative probability of survival model, using survival status at 30 days as endpoint. The model has altogether 26 coefficients, with complex transformations of Injury Severity Score (ISS), seven age categories, and interactions between age categories and gender.

In an attempt to address limitations and complexity in previous models, Jones et al. introduced the Norwegian survival prediction model in trauma (NORMIT)[15] in 2014; the first prediction model developed and validated in a Scandinavian Level I trauma centre. In NORMIT, the anatomic injury is represented by the New Injury Severity Score (NISS)[13]. The NISS has been shown to better predict trauma mortality compared to ISS[13, 16-18], in particular in patients with several severe injuries in a single body region, such as penetrating injuries towards the torso, and in both blunt and penetrating TBI[19]. In contrast to TRISS, NORMIT also accounts for the patient's pre-injury health status according to the American Society of Anesthesiologists Physical Status (ASA-PS) classification system[20]. Additionally, NORMIT incorporates age as a continuous variable, includes an unweighted physiological scoring (Revised Trauma Score for Triage; T-RTS)[21], defines rules for handling of missing physiological data that allow inclusion of intubated patients, and utilises mortality at 30 days after injury as endpoint. The model is also fairly simple, containing only eight coefficients. It might therefore be a viable alternative to older models for case-mix adjustment.

An external validation of NORMIT in a population of severely injured patients in Finland showed good discrimination but poor calibration, and suggested that the model should be re-calibrated to better fit the more severely injured patients (NISS >15)[22]. NORMIT was recently updated (NORMIT 2)[23] and a temporal validation of the model showed that NORMIT 2 performed better than both the TRISS 09 and two TARN prediction models. The authors concluded that external validations in other populations were warranted. The aim of

the present study was to perform an external validation of the NORMIT models by comparing NORMIT 1 and 2 with regard to accuracy in two Swedish trauma populations: (1) A national trauma (NT) population consisting of all patients registered in the national Swedish trauma registry (SweTrau)[24] and (2) a subpopulation consisting of patients admitted only to a single designated trauma centre (TC). The hypothesis was that NORMIT 2 would outperform NORMIT 1 and demonstrate good performance with regard to discrimination and calibration in a Swedish NT population and a TC subpopulation. The study is also intended to contribute towards a first formal evaluation of trauma care quality between Swedish hospitals.

METHODS

The reporting of the study conforms to the guideline for Transparent Reporting of a multivariable Prediction model for Individual Prognosis or Diagnosis (TRIPOD)[25] and to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement guidelines for reporting observational studies[26].

Inclusion criteria in SweTrau

The inclusion criteria in SweTrau are (a) all trauma patients admitted through trauma team activation (TTA) irrespective of injury severity, (b) all patients with NISS >15 who did not receive TTA, and (c) all patients who were transferred from another hospital (secondary admissions) to the reporting hospital within seven days after trauma who had a NISS >15. Exclusion criteria are (a) patients with isolated subdural hematoma, and (b) patients with TTA who were not exposed to a prior traumatic event. The registry is based on the revised Utstein Trauma Template which is the current European core dataset[27], and was introduced in Sweden in 2011. To date, it includes more than 50 000 trauma patients from 48 out of 52 Swedish hospitals with emergency surgical units that admit trauma patients of all ages, 24/7, 365 days a year[24].

Inclusion criteria in the current study

The study inclusion period was January 1st 2014 to December 31st 2016. Eligible were all adult patients (age ≥ 15years) registered in SweTrau who were primarily (directly) admitted to the reporting hospital. The rationale behind including only primary admissions was to limit the effect of case-mix and the differences in trauma care processes between primarily and secondary admitted (transferred) patients, which have been demonstrated in a previous study[10]. The extracted data were anonymized to ensure confidentiality of patients, physicians

and participating hospitals. Patients were excluded if data was missing and it was impossible to impute data for the different components of the prediction models (as described below), or if 30-day outcome could not be determined.

The study populations

The NT population consisted of all patients registered in SweTrau who fulfilled the inclusion criteria for the current study. The patients in this population were admitted to designated trauma hospitals, university hospitals as well as regional and local hospitals. The TC subpopulation consisted of patients fulfilling the same criteria, who were admitted to KUH in Stockholm. KUH is equivalent to a Level I trauma centre with a catchment population of 2.2 million inhabitants[28]. Trauma infrastructure including care processes as well as patient characteristics at KUH has been described previously[10, 29, 30]. To further explore the impact of case-mix with regard to injury severity, NORMIT's calibration was also assessed in subgroups with NISS >15 in both study populations.

Coding, scoring and outcome

Anatomic injury severity was scored by Association for the Advancement of Automotive Medicine (AAAM)-certified registrars according to the Abbreviated Injury Scale (AIS) 2005 – Update 2008 (AIS 08)[31]. Physiological derangement on arrival was classified according to the T-RTS[15, 21]. The T-RTS (range 0–12) is defined as the sum of the clinical category values of Glasgow Coma Scale (GCS)[32] score, systolic blood pressure (SBP), and respiratory rate (RR). The scoring of physiological data into clinical categories, based on information from text in medical records in addition to numerical raw data, substantially reduces the number of patient exclusions due to missing data[15]. In accordance with NORMIT coding rules, the most recent pre-hospital values were used when admission data were unavailable[15]. Thus, for

patients arriving intubated and in general anaesthesia, GCS, SBP and RR were scored based on values documented immediately prior to intubation. If pre-hospital values were also unavailable, the patient was excluded. Outcome was defined as survival or death 30 days after injury, independent of whether the patient was admitted or discharged from hospital.

Comparisons of NORMIT 1 and 2

The NORMIT 1 model was developed based on trauma data from Oslo University Hospital Ullevål (OUH) from a six-year period (2000-2006) and validated in a two-year dataset from 2006-2008[15]. In NORMIT 2, the original NORMIT model coefficients were updated in a derivation dataset with patients admitted 2005-2009 and evaluated in a validation dataset with patients admitted 2010-2013[23], also to OUH. The coefficients for both versions were derived with injury data coded according to AIS 1990 - Update 98 (AIS 98)[33].

Statistical methods

Data are presented as medians with quartiles. Comparisons of continuous data were performed using the Mann-Whitney *U* test or the Wilcoxon Signed Rank test depending on the distribution of the data. Normality was tested using the Shapiro-Wilk test. Differences between categorical variables were evaluated using the chi-square test (two-tailed). Statistical significance was assumed for two-sided P-values <0.05.

The performance of NORMIT 1 and 2 was evaluated by measuring their discrimination and calibration capabilities in the same way as in the Finnish external validation of NORMIT 1 performed by Raj et al.[22]. The discrimination of a survival prediction model refers to its ability to distinguish between survivors and non-survivors. Discrimination was assessed for each model by calculating the area under the receiver operating characteristic (ROC) curve (AUC)[34] with 95% confidence interval (CI). Random

guess produces an AUC of 0.5, whereas 1.0 represents perfect model performance. AUCs ≥0.90 are considered excellent, AUCs ≥0.80 good, and AUCs <0.70 poor[35]. The discrimination capabilities of two models were not considered to be significantly different if their 95% CIs overlapped.

The calibration of a model refers to the agreement between predicted and observed outcomes[34]. To assess calibration, the GiViTI (Gruppo Italiano per la Valutazione degli Interventi in Terapia Intensiva) calibration belt was utilised, using its R-package (©Nattino & Finazzi)[36, 37]. The GiViTI test is specifically designed to visually demonstrate the relationship between observed and predicted outcomes by fitting a polynomial function between the two and calculating the 80% and 95% CIs, respectively[36]. Statistically significant deviations occur when the diagonal bisector line is not contained within the 95% CI. Thus, by using the GiViTI calibration belt, it was possible to identify specific risk intervals with over- and underprediction of survival by the model[37]. Wider CIs are seen with a higher degree of uncertainty, mainly caused by a lower number of patients at the specific risk stratum.

Data were primarily analysed with SPSS (Statistical Package for the Social Sciences, Version 23.0.0, SPSS, Inc., Chicago, IL).

Ethics

The study was approved by the Regional Ethical Review Board in Stockholm, Sweden (reference number: 2017/2024-31/5).

RESULTS

During the 3-year period, 27953 patients were registered in SweTrau. After exclusion of secondary admissions (n=1449, 5.2%), 26504 patients were included in the study (Figure 1). The proportion of missing data was 18.7% (n=4950) in the NT population and 2.6% (n=103) in the TC subpopulation. After exclusion of patients with missing data, the NT population consisted of 21554 patients, and the TC subpopulation consisted of 3972 patients. The number of patients with NISS >15 was 3133 in the NT population and 1094 in the TC subpopulation. The characteristics of the two study populations are presented in Table 1. In comparison with the NT population, patients in the TC subpopulation were more severely injured, were more physiologically deranged on admission, and had a higher comorbidity and mortality.

Median Ps values for survivors and non-survivors in the NT and TC populations calculated by NORMIT 1 and 2 are shown in Table 2. Median Ps for non-survivors was higher with NORMIT 2 than with NORMIT 1 in both populations. Both NORMIT 1 and 2 produced higher median Ps values for non-survivors in the NT population than in the TC subpopulation.

Both models displayed excellent discrimination in all populations and subgroups when evaluated with AUC (Figure S1, supporting information), with no significant differences between the groups (Table 3). However, both models displayed a poor calibration in the NT population (Figure 2) with lower observed than predicted survival, in particular for NORMIT 2 which overestimated survival through the Ps interval of 0.23-0.98 (Figure 2, upper right panel). Similar relationships were demonstrated in the severely injured (NISS >15) subgroup of the NT population, where NORMIT 2 overestimated survival through the entire Ps interval (Figure 2, lower right panel). In contrast, in the TC subpopulation NORMIT 1 underestimated survival in the Ps interval 0.38-1.00 (Figure 3 upper left panel), while

NORMIT 2 showed good calibration as the 95% CI calibration belt never crossed the diagonal bisector line, i.e., there was no under- or overestimation of survival (Figure 3, upper right panel). A nearly identical picture was observed in the TC subpopulation with NISS >15 (Figure 3, lower panels).

DISCUSSION

Key findings

External validation of the NORMIT 1 and 2 prediction models showed that they had excellent survival prediction abilities in both the NT and TC populations. Calibration was poor for both models in the NT population with overestimated survival, particularly for NORMIT 2. In the TC subpopulation, NORMIT 1 showed poor calibration due to underestimated survival, whereas the updated NORMIT 2 showed good calibration including in the subgroup of severely injured patients.

Model accuracy – discrimination and calibration

The high AUCs for NORMIT 1 and 2 in the current study demonstrated an excellent ability to separate survivors and non-survivors[38] in both study populations, and also in their subgroup with severe injuries. The AUC statistics are popular in development of diagnostic tests but may not necessarily detect small differences in discriminative ability between two models[39]. A large majority of the trauma patients did not have life-threatening injuries, thus yielding a high Ps and therefore easy to predict as survivors. A better indicator of high discriminating ability is therefore less overlap in Ps values between trauma survivors and non-survivors[34]. In both populations, NORMIT 1 and 2, yielded higher median Ps values among trauma survivors than among non-survivors. The median Ps value in non-survivors was higher in the NT population than the TC subpopulation (0.60 vs. 0.35 with NORMIT 1 and 0.66 vs. 0.49 with NORMIT 2). In other words, patients who died in the NT population had a clearly higher probability of surviving than those who died in the TC subpopulation. The actual difference between trauma centres and non-trauma centres in the current study is probably even larger, since the TC subpopulation also constituted 18% of the NT population (Figure 1 and Table 1).

The observed differences need to be commented. Firstly, the TC subpopulation is expected to have generally lower Ps values because it consists of more severely injured patients who are also older and have more comorbidity (Table 1). Secondly, the resources and trauma competences available at a designated trauma centre should be associated with better performance, leading to more lives saved and thus to lower Ps in non-survivors. The observed median Ps values in the TC subpopulation in the current study are comparable to those previously found in the Norwegian OUH population, where median Ps values were 0.32 (NORMIT 1 with AIS 98) and 0.41 (NORMIT 2 with AIS 08)[23]. Both OUH and KUH are designated trauma centres with comparable trauma populations, and assumedly with similar trauma processes and quality of care which should yield similar outcomes.

Furthermore, the Ps values for non-survivors were lower for NORMIT 1 compared to NORMIT 2 in the NT population (0.60 vs. 0.66) and even more pronounced in the TC subpopulation (0.35 vs 0.49). This observation might suggest that the NORMIT 1 model showed better discrimination than NORMIT 2. An alternate interpretation is that Ps values estimated with NORMIT 2 can be expected to be higher because the NORMIT 2 model was derived from a more recent OUH trauma population (2005-2009) with improved treatment and therefore lower risk-adjusted mortality compared to the original NORMIT 1 population (2000-2006). A study of risk-adjusted survival as a function of time at OUH during 2001–2011 has confirmed this assumption, demonstrating improved survival in patients with critical neurotrauma after late 2004[40].

More important in this setting is to accurately assess the model calibration, i.e., the agreement between survival predictions and observed outcomes over the full range of probabilities[38]. Skaga et al. stated[23] that the mildly and the very severely injured patients are easier to predict as survivors and non-survivors respectively, and therefore a well-calibrated prediction model is distinguished by high performance in the mid-bands of Ps strata. In the

current study, the GiViTI calibration belts displayed a variety of deviations dependent of population and injury severity. In the NT population, both models performed poorly (generally too optimistic) but NORMIT 1 overestimated survival in both injury severity groups to a lesser extent than NORMIT 2. Contrary to this, the observed survival rates in both injury severity groups in the TC subpopulation were equal to those predicted by NORMIT 2, while NORMIT 1 underestimated survival in the higher Ps intervals.

Differences in case-mix may have contributed to the the different model performance between the NT and TC populations. However, this is exactly what the NORMIT model was designed to adjust for. The variation in outcome between the two trauma populations may more probably be caused by differences in care processes, systems and quality of care. Designated trauma centres receive higher volumes of critically injured patients and are expected to perform better than non-designated trauma hospitals, in particular amongst severely injured patients. Consequently, risk-adjusted survival can be expected to be lower in the NT population which consisted of patients admitted to all Swedish hospitals with an emergency unit.

The poor performance of the NORMIT model in a national Swedish setting might also be due to differences between the Swedish national trauma system and the local system in Southeast Norway where the NORMIT model was derived (i.e. selection differences). In the Southeast Norway trauma system, there is a single regional Level 1 trauma centre with cooperating local trauma receiving hospitals and an extensive emergency medical system including anaesthesiologist-manned rapid response cars and helicopters delivering advanced emergency care at the site of injury and during patient transport[23], that supplement ground ambulances[41, 42]. This may also contribute to explain the good performance at KUH, which in many regards is similar to the original model derivation and validation system[10].

Different methods of measuring model performance

In development of the NORMIT 1 model, calibration was first explored through calibration plots, and second by using a Hosmer–Lemeshow (H-L) goodness-of-fit test[43]. In the NORMIT 2 model, calibration was explored as in NORMIT 1 through calibration plots but the overall model performance was evaluated with the scaled Brier score[44]. In the current study, the GiViTI calibration belt was used. Even though the GiViTI belt and H-L test have been found to generate similar results[37, 45, 46], the different methods of exploring model performance might, although less likely, have contributed to the different results observed.

Different coding

Differences in injury severity coding might also have contributed to the poor calibration ability of NORMIT 1 and 2 in the NT population. In SweTrau, NISS is coded according to AIS 08, while both NORMIT models are based on AIS 98. Different AIS versions are not always comparable[34] and it has been suggested that AIS 08 generates lower ISS and NISS than AIS 98[47]. Seemingly lower injury severity would lead to higher estimated Ps, i.e., overestimated survival for a given injury. This could however not explain the results from the TC subpopulation with underestimation of survival by NORMIT 1 and good model performance by NORMIT 2, which in fact contradicts the above reasoning. Further, in the original NORMIT 2 study the model showed even better performance when it was "stressed" with AIS 08[23].

Missing data

The pattern of missing data in trauma registries is rarely at random, and studies from US trauma populations have demonstrated that patients excluded due to missing RTS values had

worse prognosis than patients with complete data and that such differential exclusion could bias the conclusions drawn[7, 14, 48, 49]. Large numbers of patients with missing data may therefore bias study results. In the original NORMIT study by Jones et al. missing data was <1%[15], in the Finnish external validation study by Raj et al. 7.1%[22], and in the recent NORMIT 2 update study by Skaga et al. 0.26%[23]. In the current study, missing data was 18.7% in the NT population and 2.7% in the TC subpopulation. There is no established cut-off in the literature regarding an acceptable percentage of missing data for valid statistical inferences, but a missing rate of greater than 10% has been suggested to interfere with statistical analysis[50]. Similar to other studies, missing RTS values and ASA-PS classification were the most common reason for exclusion in the present study and this may have affected the analysis regarding NORMIT's accuracy, particularly in the NT population. Therefore, corrective actions should be taken to minimize the amount of missing data in SweTrau, in particular in the national population, in order to increase its reliability for future research.

In summary**,** the study suggests that NORMIT 2 is well suited to predict survival in a Swedish trauma centre population for both less and severely injured patients, but both NORMIT models perform poorly in a more heterogeneous national trauma population. The reasons for these discrepancies are not clear, but differences in quality of care between dedicated trauma centres and the national population may contribute.

ACKNOWLEDGMENTS

The authors would like to thank the trauma registrars across the country for their hard work registering trauma patients in the SweTrau.

REFERENCES

1. Demetriades D, Chan L, Velmanos GV, Sava J, Preston C, Gruzinski G *et al*. TRISS methodology: an inappropriate tool for comparing outcomes between trauma centers. *J Am Coll Surg* 2001; 193: 250-254.

2. Shafi S, Nathens AB, Cryer HG, Hemmila MR, Pasquale MD, Clark DE *et al*. The trauma quality improvement program of the American College of Surgeons Committee on Trauma. *J Am Coll Surg* 2009; 209: 521-530.

3. Roques F, Nashef S, Michel P, Gauducheau E, De Vincentiis C, Baudet E *et al*. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. *Eur J Cardiothoratic Surg* 1999; 15: 816-823.

4. Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. *J Trauma* 1987; 27: 370-378.

5. 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 Acute Care Surg* 2006; 61: 701-710.

6. Huber-Wagner S, Lefering R, Qvick L-M, Körner M, Kay MV, Pfeifer K-J *et al*. Effect of whole-body CT during trauma resuscitation on survival: a retrospective, multicentre study. *Lancet* 2009; 373:1455-1461.

7. Schluter PJ, Nathens A, Neal ML, Goble S, Cameron CM, Davey TM *et al*. Trauma and injury severity score (TRISS) coefficients 2009 revision. *J Trauma Acute Care Surg* 2010; 68: 761-770.

8. 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.

9. Champion HR, Copes WS, Sacco WJ, Lawnick MM, Keast SL, Bain Jr L *et al*. The Major Trauma Outcome Study: establishing national norms for trauma care. *J Trauma* 1990; 30: 1356-1365.

10. Ghorbani P, Ringdal KG, Hestnes M, Skaga NO, Eken T, Ekbom A *et al*. Comparison of risk-adjusted survival in two Scandinavian Level-I trauma centres. *Scand J Trauma Resusc Emerg Med* 2016; 24: 66.

11. Cayten C, Stahl W, Murphy J, Agarwal N, Byrne D. Limitations of the TRISS method for interhospital comparisons: a multihospital study. *J Trauma* 1991; 31: 471-482.

12. Demetriades D, Chan L, Velmahos G, Berne T, Cornwell E, Belzberg H *et al*. TRISS methodology in trauma: the need for alternatives. *Br J Surg* 1998; 85: 379-384.

13. Osler T, Baker SP, Long W. A modification of the injury severity score that both improves accuracy and simplifies scoring. *J Trauma Acute Care Surg* 1997; 43: 922-926.

14. 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 Acute Care Surg* 2006; 60: 538-547.

15. Jones J, Skaga N, Søvik S, Lossius H, Eken T. Norwegian survival prediction model in trauma: modelling effects of anatomic injury, acute physiology, age, and co-morbidity. *Acta Anaesthesiol Scand* 2014; 58: 303-315.

16. Lefering R. Development and validation of the revised injury severity classification score for severely injured patients. *Eur J Trauma Emerg Surg* 2009; 35: 437-447.

17. 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 Acute Care Surg* 2004; 56: 1312-1320.

18. 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 Acute Care Surg* 2003; 55: 1083-1088.

19. Cayten CG, Stahl WM, Agarwal N, Murphy JG. Analyses of preventable deaths by mechanism of injury among 13,500 trauma admissions. *Ann Surg* 1991; 214: 510-520.

20. American Society of Anesthesiologists. *ASA Physical Status Classification System*. http://www.asahq.org/resources/clinical-information/asa-physical-status-classification-system [accessed 1 September 2018].

21. Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the Trauma Score. *J Trauma* 1989; 29: 623-629.

22. Raj R, Brinck T, Skrifvars M, 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.

23. Skaga NO, Eken T, Søvik S. Validating performance of TRISS, TARN and NORMIT survival prediction models in a Norwegian trauma population. *Acta Anaesthesiol Scand* 2018; 62: 253-266.

24. The National Swedish Trauma Registry – SweTrau. http://rcsyd.se/swetrau/ [accessed 2 September 2018].

25. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015; 13: 1.

26. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ *et al*. STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiology* 2007; 18: 805-835.

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

28. Brattström O, Larsson E, Granath F, Riddez L, Bell M, Oldner A. Time dependent influence of host factors on outcome after trauma. *Eur J Epidemiol* 2012; 27: 233-241.

29. Ghorbani P, Falkén M, Riddez L, Sundelöf M, Oldner A, Strömmer L. Clinical review is essential to evaluate 30-day mortality after trauma. *Scand J Trauma Resusc Emerg Med* 2014; 22: 18.

30. Ghorbani P, Strömmer L. Analysis of preventable deaths and errors in trauma care in a Scandinavian trauma level-I centre. *Acta Anaesthesiol Scand* 2018; 00: 1-8.

31. The Abbreviated Injury Scale 2005 revision – Update 2008. Des Plains, IL: Association for the Advancement of Automotive Medicine, 2008.

32. Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. *Lancet* 1974; 304: 81-84.

33. The Abbreviated Injury Scale 1990 revision – Update 98. Des Plaines, IL: Association for the Advancement of Automotive Medicine, 1998.

34. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass)* 2010; 21: 128-138.

35. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression: John Wiley & Sons; 2013.

36. Finazzi S, Poole D, Luciani D, Cogo PE, Bertolini G. Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PLoS One* 2011; 6: e16110.

37. Nattino G, Finazzi S, Bertolini G. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. *Stat Med* 2014; 33: 2390-2407.

38. 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: 661-672.

39. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; 115: 928-935.

40. 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: 1722-1730.

41. Langhelle A, Lossius HM, Silfvast T, Björnsson HM, Lippert FK, Ersson A *et al*. International EMS Systems: the Nordic countries. *Resuscitation* 2004; 61: 9-21.

42. Krüger AJ, Skogvoll E, Castrén M, Kurola J, Lossius HM. Scandinavian pre-hospital physician-manned Emergency Medical Services—Same concept across borders? *Resuscitation* 2010; 81: 427-433.

43. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997; 16: 965-980.

44. Steyerberg E. Clinical prediction models. Berlin, Germany: Springer, 2009.

45. Raj R, Siironen J, Kivisaari R, Hernesniemi J, Skrifvars MB. Predicting outcome after traumatic brain injury: development of prognostic scores based on the IMPACT and the APACHE II. *J Neurotrauma* 2014; 31:1721-1732.

46. Raj R, Skrifvars MB, Bendel S, Selander T, Kivisaari R, Siironen J *et al*. Predicting six-month mortality of patients with traumatic brain injury: usefulness of common intensive care severity scores. *Crit Care* 2014; 18: R60.

47. Tohira H, Jacobs I, Matsuoka T, Ishikawa K. Impact of the version of the abbreviated injury scale on injury severity characterization and quality assessment of trauma care. *J Trauma Acute Care Surg* 2011; 71: 56-62.

48. Clark DE. Comparing institutional trauma survival to a standard: current limitations and suggested alternatives. *J Trauma Acute Care Surg* 1999; 47: 92-98.

49. Kirkham JJ. A comparison of hospital performance with non-ignorable missing covariates: an application to trauma care data. *Stat Med* 2008; 27: 5725-5744.

50. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res* 1999; 8: 3-15.

FIGURE LEGENDS

Figure 1. Study populations.
ASA-PS: American Society of Anesthesiologists Physical Status classification system
NISS: New Injury Severity Score
RTS: Revised Trauma Score
SBP: Systolic blood pressure
RR: Respiratory rate
GCS: Glasgow Coma Scale.

Figure 2. GiViTI calibration belt for NORMIT 1 (left panels) and NORMIT 2 (right panels) in the national trauma (NT) population (upper panels) and in the same population but with NISS >15 (lower panels). The bisector (red reference line) represents agreement between predicted and observed survival rate. The calibration belt (grey area) depicts the estimated relationship between the model predictions and the probabilities of the true response, with 80% (light grey) and 95% (dark grey) confidence levels. The bottom-right table reports the ranges of the predicted probabilities where the calibration belt deviates significantly from the bisector, i.e., where observed survival is significantly different from what the model predicts.
NORMIT: Norwegian survival prediction model in trauma.
NISS: New Injury Severity Score.

Figure 3. GiViTI calibration belt for NORMIT 1 (left panels) and NORMIT 2 (right panels) in the trauma centre (TC) subpopulation (upper panels) and in the same subpopulation but with NISS >15 (lower panels). See Figure 2 for details.
NORMIT: Norwegian survival prediction model in trauma.
NISS: New Injury Severity Score.

SUPPORTING INFORMATION

Additional supporting information is found in the online version of this article:

Figure S1. Receiver operating characteristic (ROC) curves for NORMIT 1 (black line) and NORMIT 2 (green line). A, National trauma (NT) population; B, Trauma centre (TC) subpopulation. The red reference (diagonal) line represents the performance of a model that is no better than a random guess, i.e., AUC = 0.5. See Table 3 for numeric values regarding AUC.
NORMIT: Norwegian survival prediction model in trauma
AUC: Area under the receiver operating characteristic (ROC) curve.

Table 1. Patient baseline characteristics for the national trauma (NT) population and the trauma centre (TC) subpopulation.

	National trauma (NT) population (n=21554)	**Trauma centre (TC) subpopulation (n=3972)**	**P-value**
Age (years)	35 (21-56)	41 (26-59)	<0.001
Male gender	13531 (62.8)	2689 (67.7)	0.015
Penetrating injury	1498 (7.0)	457 (11.5)	<0.001
ISS	2 (1-9)	5 (1-12)	<0.001
NISS	3 (1-9)	6 (3-17)	<0.001
30-day mortality	652 (3.0)	172 (4.3)	<0.001
Pre-injury ASA-PS			
1	15595 (72.4)	2534 (63.8)	<0.001
2	4169 (19.3)	980 (24.7)	
3	1688 (7.8)	425 (10.7)	
4	102 (0.5)	33 (0.8)	
RR RTS Category			
4	20435 (94.8)	3716 (93.6)	<0.001
3	837 (3.9)	156 (3.9)	
2	110 (0.5)	47 (1.2)	
1	26 (0.1)	9 (0.2)	
0	146 (0.7)	44 (1.1)	
SBP RTS Category			
4	21139 (98.1)	3854 (97.0)	<0.001
3	175 (0.8)	37 (0.9)	
2	99 (0.5)	35 (0.9)	
1	34 (0.1)	9 (0.2)	
0	107 (0.5)	37 (0.9)	
GCS RTS Category			
4	20264 (94.0)	3489 (87.8)	<0.001
3	547 (2.5)	179 (4.5)	
2	268 (1.3)	115 (2.9)	
1	128 (0.6)	60 (1.5)	
0	347 (1.6)	129 (3.3)	
T-RTS	12 (12-12)	12 (12-12)	<0.001

Categorical data are presented as numbers and proportions (%) and continuous data as medians and quartiles.
ISS: Injury Severity Score
NISS: New Injury Severity Score
ASA-PS: American Society of Anesthesiologists Physical Status
RR: Respiratory rate
RTS: Revised Trauma Score
SBP: Systolic blood pressure
GCS: Glasgow Coma Scale
T-RTS: Triage-RTS.

Table 2. Predicted survival in survivors and non-survivors in the national trauma (NT) population and the trauma centre (TC) subpopulation.

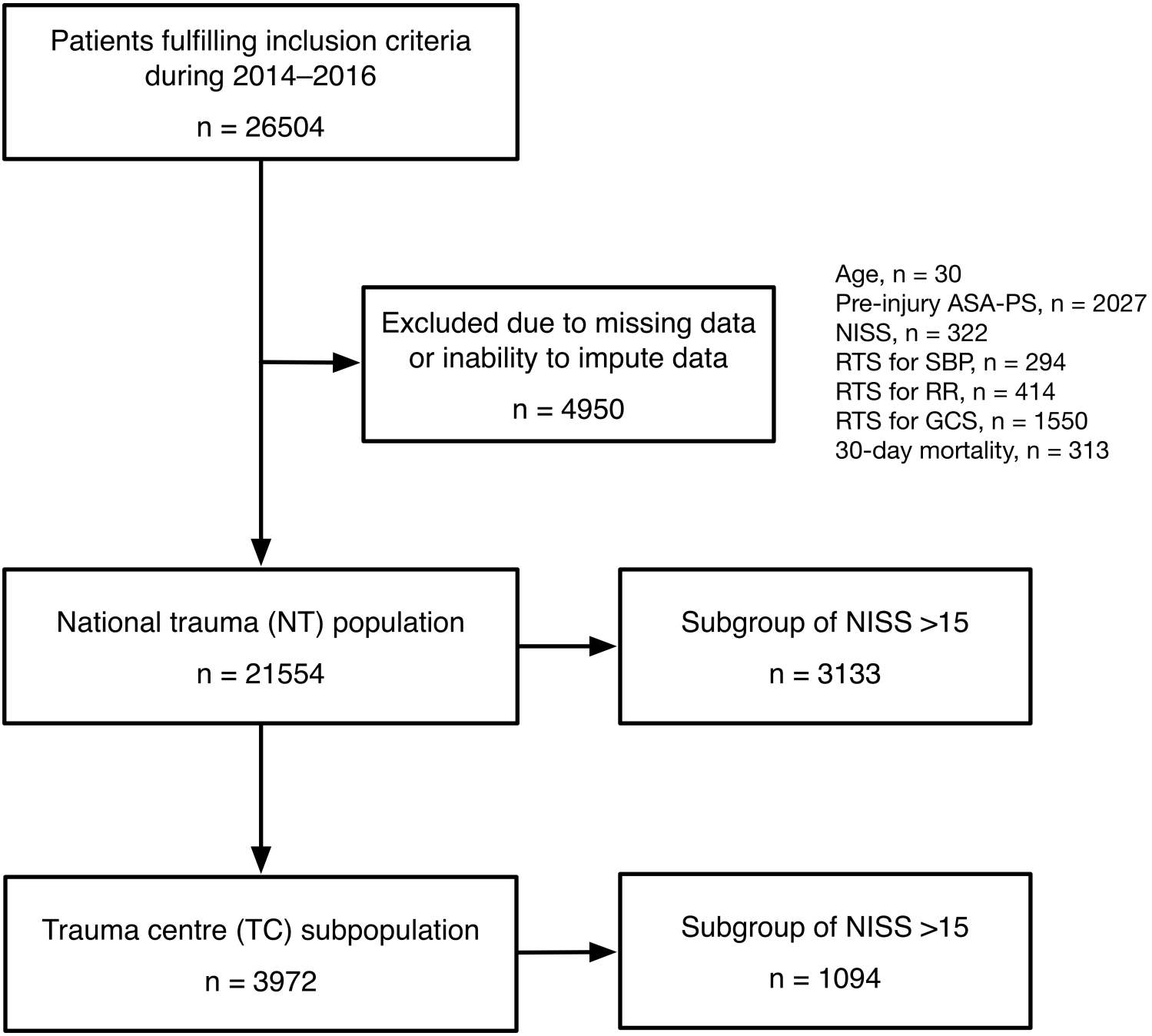
	National trauma (NT) population (n=21554)		**Trauma centre (TC) subpopulation (n=3972)**		
	Survivors (n=20902)	**Non-survivors (n=652)**	**Survivors n=3800**	**Non-survivors n=172**	**P-value***
NORMIT 1 Ps	0.9994 (0.9961-0.9997)	0.6004 (0.2450-0.8689)	0.9983 (0.9889-0.9996)	0.3464 (0.0838-0.6983)	<0.001
NORMIT 2 Ps	0.9984 (0.9954-0.9991)	0.6647 (0.3139-0.9051)	0.9973 (0.9902-0.9989)	0.4881 (0.1245-0.7671)	<0.001
P-value†		<0.001		<0.001	

Data are presented as medians and quartiles.
*P-value measured between non-survivors in the two populations in each model
†P-value measured between non-survivors in the two models in each population
NORMIT: Norwegian survival prediction model in trauma
Ps: Probability of survival.

Table 3. Area under the receiver operating characteristic (ROC) curve (AUC) for NORMIT 1 and 2 in the national trauma (NT) and trauma centre (TC) populations and their subgroups (NISS >15).

	NORMIT 1	**NORMIT 2**
National trauma (NT) population	0.968 (0.962-0.974)	0.971 (0.965-0.977)
NISS >15 subgroup	0.933 (0.922-0.945)	0.937 (0.926-0.948)
Trauma centre (TC) subpopulation	0.974 (0.965-0.983)	0.976 (0.967-0.985)
NISS >15 subgroup	0.964 (0.952-0.976)	0.965 (0.954-0.977)

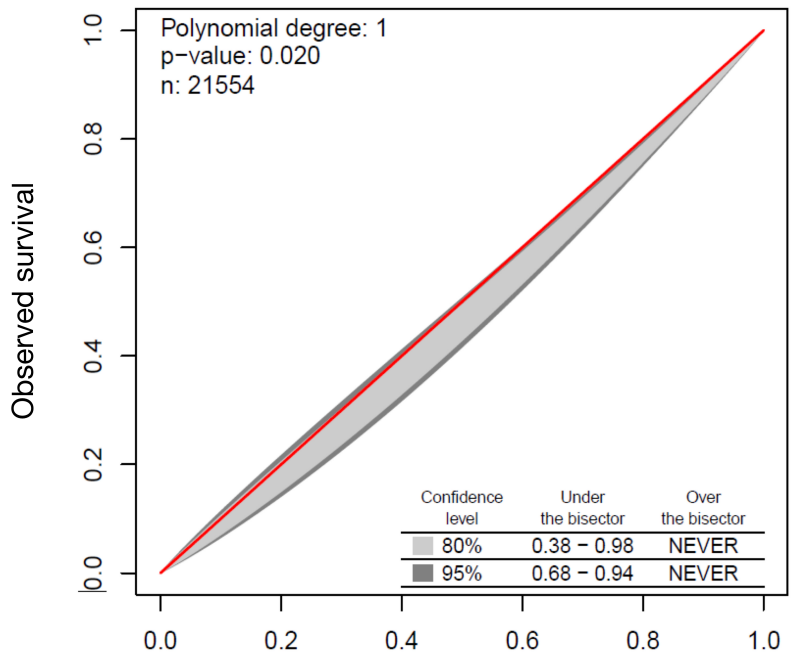
Data are presented with 95% confidence intervals (CI). The discrimination capabilities of the two models were not considered to be significantly different if their 95% CIs overlapped.
NORMIT: Norwegian survival prediction model in trauma
NISS: New Injury Severity Score.

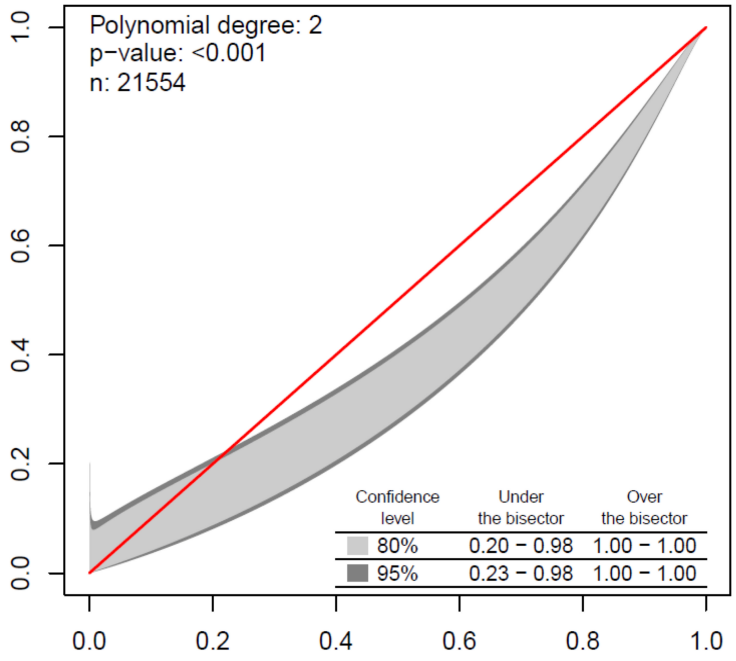

Patients fulfilling inclusion criteria during 2014–2016
n = 26504
Excluded due to missing data or inability to impute data
n = 4950
Age, n = 30
Pre-injury ASA-PS, n = 2027
NISS, n = 322
RTS for SBP, n = 294
RTS for RR, n = 414
RTS for GCS, n = 1550
30-day mortality, n = 313
National trauma (NT) population
n = 21554
Subgroup of NISS >15
n = 3133
Trauma centre (TC) subpopulation
n = 3972
Subgroup of NISS >15
n = 1094

NORMIT 1

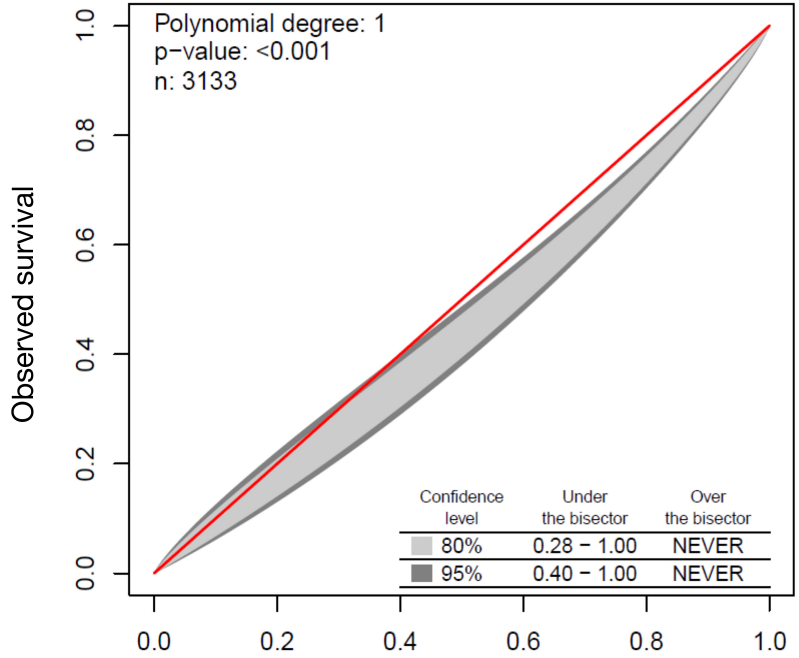
NORMIT 2

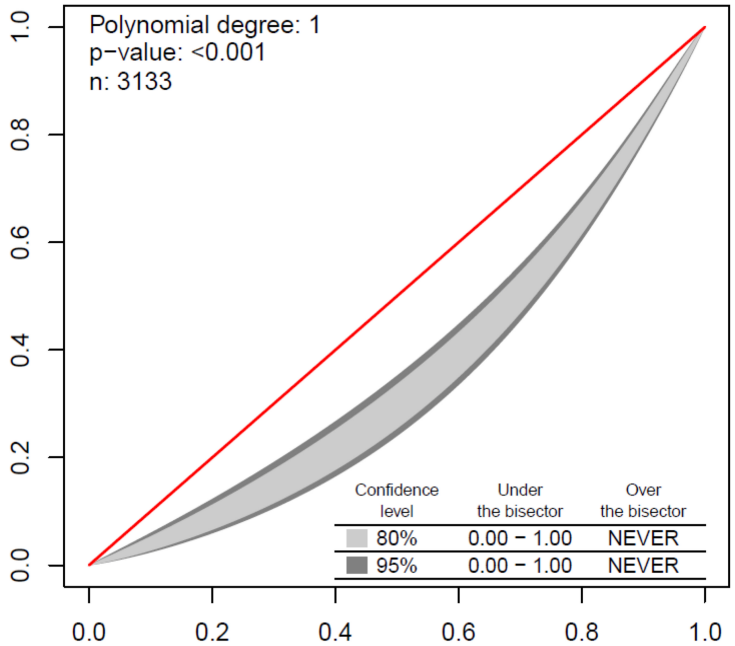
NT population, total





NT population, NISS >15

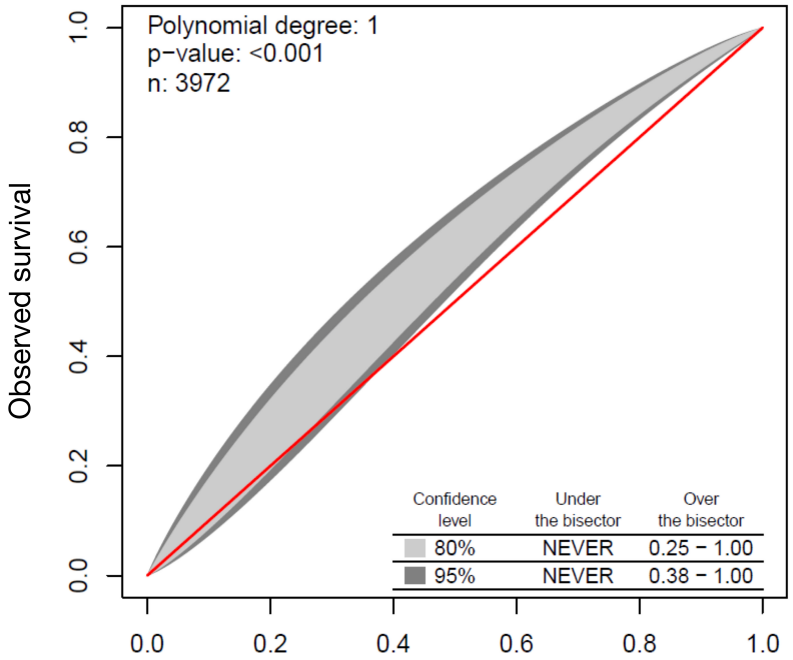


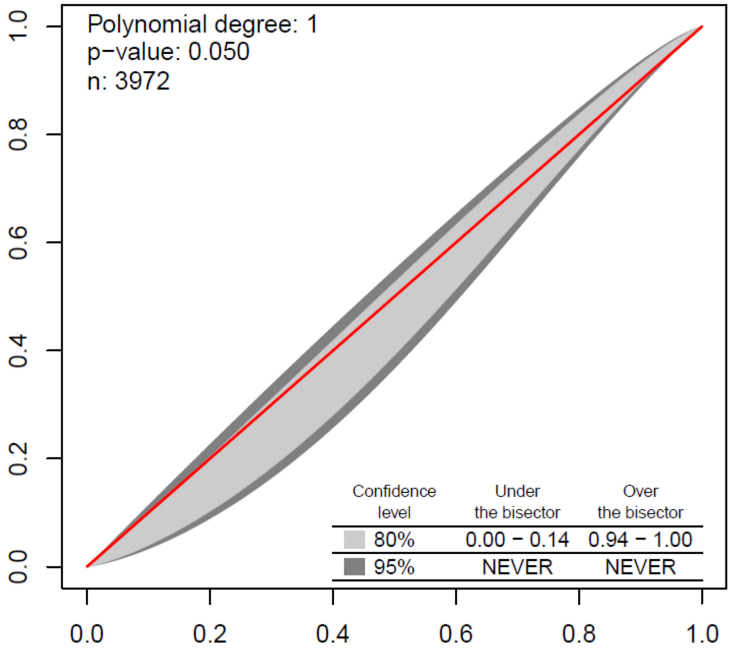


NORMIT 1

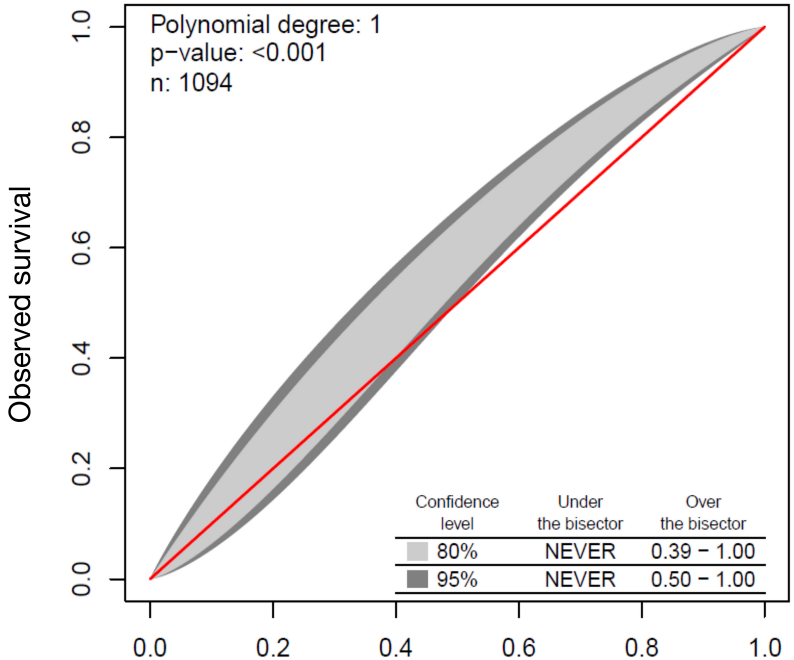
NORMIT 2

TC subpopulation, total





TC subpopulation, NISS >15



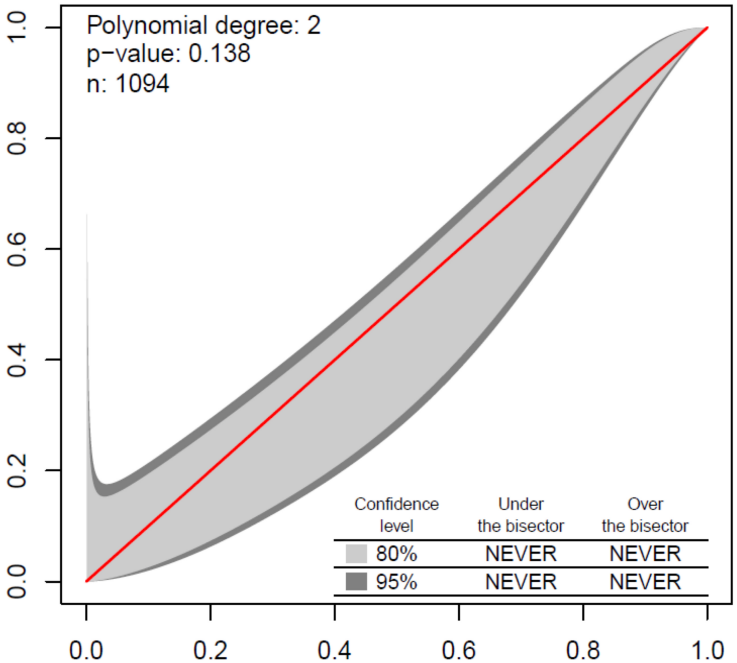
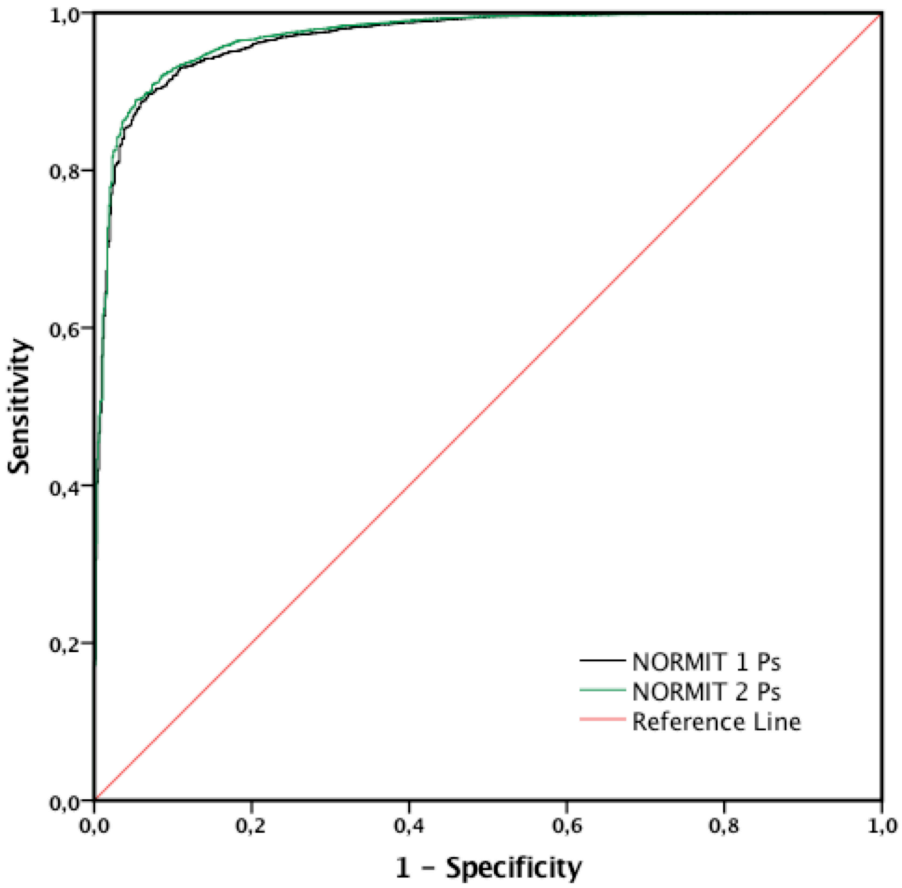
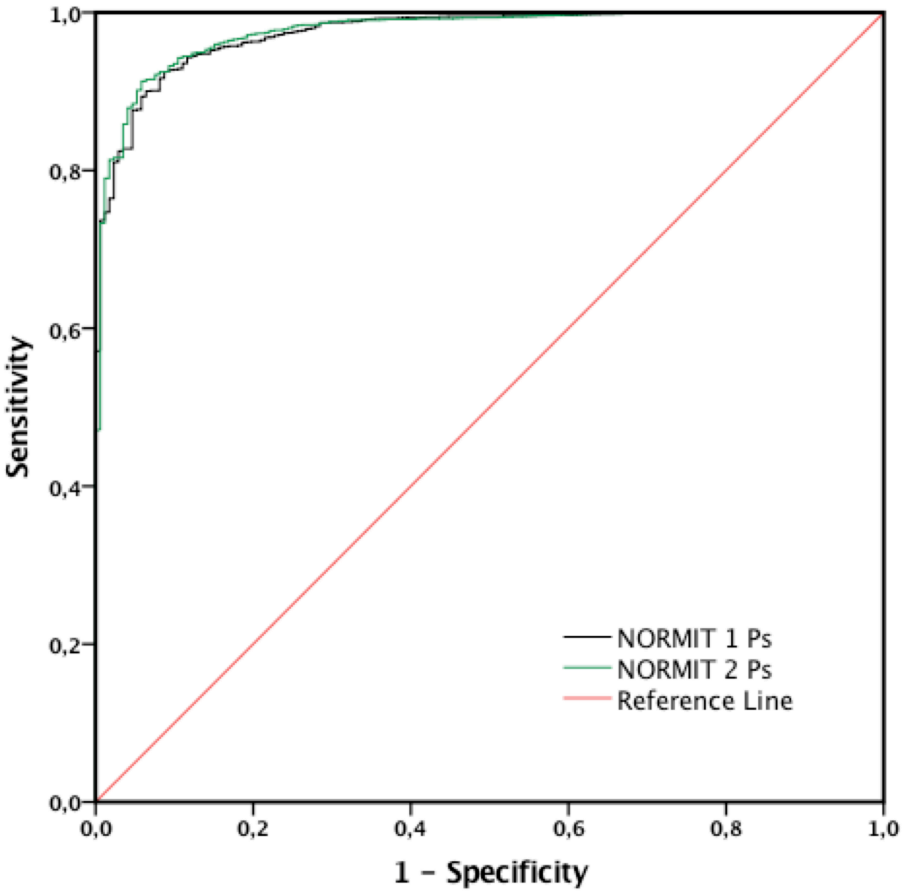


Figure S1

A


1,0
0,8
0,6
0,4
0,2
0,0
Sensitivity
0,0
0,2
0,4
0,6
0,8
1,0
1 – Specificity
NORMIT 1 Ps
NORMIT 2 Ps
Reference Line

B


1,0
0,8
0,6
0,4
0,2
0,0
Sensitivity
0,0
0,2
0,4
0,6
0,8
1,0
1 – Specificity
NORMIT 1 Ps
NORMIT 2 Ps
Reference Line