Reliability and accuracy of the South African Triage Scale when used by nurses in the emergency department of Timergara Hospital, Pakistan

M K Dalwai,[1,2] MB ChB; **M Twomey,**[2] BSc, PhD; **J Maikere,**[1] MB ChB, FCS, PhD; **S Said,**[1] BSc (Nursing); **M Wakeel,**[3] MB ChB; **J-P Jemmy,**[4] MB ChB; **P Valles,**[4] MB ChB; **K Tayler-Smith,**[4] **L Wallis,**[2] MB ChB, FCEM, PhD; **R Zachariah,**[4] MB ChB

[1] *Médecins Sans Frontières, Pakistan*

[2] *Division of Emergency Medicine, Faculty of Health Sciences, University of Cape Town, South Africa*

[3] *Ministry of Health, Islamabad, Pakistan*

[4] *Médecins Sans Frontières, Medical Department (Operational Research), Operational Centre, Brussels, Belgium, MSF-Luxembourg (LuxOR), Luxembourg*

Corresponding author: M K Dalwai (mkdalwai@gmail.com)

Background. Triage is one of the core requirements for the provision of effective emergency care and has been shown to reduce patient mortality. However, in low- and middle-income countries this strategy is underused, under-resourced and poorly researched.

Objective. To assess the inter- and intra-rater reliability and accuracy of nurse triage ratings when using the South African Triage Scale (SATS) in an emergency department (ED) in Timergara, Pakistan.

Methods. Fifteen ED nurses assigned triage ratings to a set of 42 reference vignettes (written case reports of ED patients) under classroom conditions. Inter-rater reliability was assessed by comparing these triage ratings; intra-rater reliability was assessed by asking the nurses to re-triage 10 random vignettes from the original set of 42 vignettes and comparing these duplicate ratings. Accuracy of the nurse ratings was measured against the reference standard.

Results. Inter-rater reliability was substantial (intraclass correlation coefficient 0.77; 95% confidence interval (CI) 0.69 - 0.85). The intra-rater agreement was also high with 87% exact agreement (95% CI 67 - 100) and 100% agreement allowing for a one-level discrepancy in triage ratings. Overall, the SATS had high specificity (97%) and moderate sensitivity (70%). Across all acuity levels the proportion of over-triage did not exceed the acceptable threshold of 30 - 50%. Under-triage was acceptable for all except emergency cases (66%).

Conclusion. ED nurses in Pakistan can reliably use the SATS to assign triage acuity ratings. While the tool is accurate for 'very urgent' and 'routine' cases, importantly, it may under-triage 'emergency' cases requiring immediate attention. Approaches that will improve accuracy and validity are discussed.

S Afr Med J 2014;104(5):372-375. DOI:10.7196/SAMJ.7604

With the increase in urbanisation and violent conflicts, together with the growing burden of chronic non-communicable diseases in many low- and middle-income countries (LMICs), there is an increased burden on emergency healthcare services.[1] In many LMICs, one of the main challenges facing emergency services is the capacity to deal with high patient loads.[2] The process of 'triage' is one way of addressing this challenge, since it optimises the allocation and use of existing resources.

Triage is the process of sorting critically ill patients who need immediate lifesaving interventions from patients who need medical attention but can safely wait to be seen.[3] Triage aims to determine a patient's 'acuity level' – i.e. how urgently they require medical attention. Triage is recognised as being one of the core requirements for the provision of effective emergency care and has been shown to reduce patient mortality.[4] However, in LMICs this strategy is underused, under-resourced and poorly researched.

The South African Triage Scale (SATS) was developed in 2004 for pre- and in-hospital emergency units throughout South Africa (SA).[5] It was specifically designed to be used by nursing assistants and as such was intended to serve as a coping measure to address medical staff shortages and limited resources – challenges that are commonplace in SA, as in other LMICs.[6]

In 2011, Médecins Sans Frontières (MSF), an international medical humanitarian organisation, implemented the SATS in Timergara Hospital (TH) in the rural district of Lower Dir in the province of Khyber Pakhtunkhwa (KPK), Pakistan. MSF had been working at this hospital alongside the Pakistan Ministry of Health to improve emergency healthcare for the population. Against a backdrop of limited resources, overstretched staff and the absence of a standardised triage system, MSF implemented the SATS with good preliminary results.[7]

The SATS has been assessed extensively in SA and implemented in several LMIC settings.[8,9] However, a more formal assessment of the SATS in a LMIC setting outside sub-Saharan Africa has not yet been undertaken.

The two most common measures for assessing a triage scale are reliability and validity. Reliability is the extent to which the triage scale yields the same result on repeated assessments of the same patient. Inter-rater reliability determines whether there is variability between different staff rating the same patient, while intra-rater reliability assesses the variability for one member of staff re-triaging the same patient. Validity has been defined as indicating how closely an acuity rating assigned using the triage scale is to the true acuity of that patient.[10] However, limitations exist when trying to validate triage scales in any setting, owing to lack of a gold standard. As

such, validity has been assessed using surrogate markers such as hospital admission, discharge and resource utilisation.[11] In LMICs, however, the use of these surrogate markers is difficult owing to poor record keeping, varying levels of clinical skills and limited resources. Previous studies in LMICs have instead attempted to assess the validity of a triage scale by comparing the triage ratings assigned by emergency department (ED) staff for a series of simulated cases against those obtained from an expert panel based on the panel's expert opinion.[12] For the purpose of this study, we will refer to this methodology and use a set of 42 reference vignettes as a reference standard against which accuracy is measured.[13]

This study therefore aimed to determine the reliability (inter- and intra-rater) and accuracy of the adult version of the SATS when used by ED nurses in TH, Pakistan.

Methods

Study design
This was a cross-sectional study using a set of 42 reference vignettes (short, written, clinical case reports of ED patients) as a proxy for live ED cases.

Setting
TH is situated in the predominantly rural district of Lower Dir in the KPK province of Pakistan. It is the only district hospital in Lower Dir, serving an estimated population of 1.8 million. The ED has an estimated annual caseload of ~48 000 patients, comprising both adults and children. The caseload is largely made up of medical emergencies (typically respiratory infection, cardiac disease and gastrointestinal illness) and trauma (most often road traffic accidents).

SATS and its use in the TH ED
The SATS uses a physiologically based composite scoring system, the Triage Early Warning Score, together with a list of discriminators, with which to triage patients into one of five colour-coded groups according to their degree of urgency for medical attention. The colour categories are as follows: (i) red, 'emergency' (to be seen immediately); (ii) orange, 'very urgent' (to be seen within 10 min); (iii) yellow, 'urgent' (to be seen within 60 min); (iv) green, 'routine' (to be seen within 240 min, i.e. minor injuries/ illness); and (v) black, 'dead'.

The SATS was introduced in the TH ED in June 2011. All ED staff received a 1-hour structured training course, which was carried out by the expatriate ED doctor. It involved explaining patient flow in the ED together with each step of the triage algorithm and the composite physiological score where each vital sign is not seen in isolation but rather as a composite part of an early warning score. Each discriminator was explained using common local ED examples.

Using the SATS, triage was routinely undertaken by two triage nurses during each work shift. Once triaged, 'red' and 'orange' patients were seen by the MSF team (a national doctor, three nurses and an expatriate doctor) in the resuscitation room, while 'yellow' and 'green' patients were seen by the national casualty medical officers in a room adjacent to the ED. At the time of the study, 23 nurses were on the ED rota and carrying out triage.

Study population
The study included all nurses at TH who fulfilled the following inclusion criteria: (i) those who had received training in the SATS and had at least 1 month's experience performing patient triage using this tool; and (ii) those who agreed to participate in the study. As the study attempted to recruit all nurses fulfilling the above criteria, it was not necessary to calculate the required sample size.

Data collection
Under classroom conditions, nurses participating in the study were required to assign one of four priority categories to the set of 42 reference vignettes according to the SATS acuity levels of 'emergency', 'very urgent', 'urgent' and 'routine'. The vignettes had been collected and validated in a previous study and were based on real ED cases from a secondary hospital in SA.[13] The type and spectrum of patient presentations captured in these vignettes closely mirrored the sort of cases presenting at the TH ED. The vignettes included information on patient gender, age, presenting complaint, mode of arrival to the ED, and vital signs. Some vignettes also included information from additional investigations such as blood glucose test and haemoglobin, as done at the time of triage. For the purpose of this study, the vignettes were translated from English into Urdu, the national language of Pakistan. This was carried out by a professional translator and ratified by a local bilingual doctor to ensure the correct medical terminology.

Reliability
Inter-rater reliability was measured by comparing the different nurse triage ratings for the 42 vignettes, while intra-rater reliability was measured by asking the nurses to re-triage 10 random vignettes from the original set of 42 vignettes and comparing these duplicate ratings.

Accuracy
The accuracy of nurse triage ratings for the 42 vignettes was measured by comparing their ratings with the acuity ratings assigned to the same set of vignettes by an international expert panel. The panel of 18 experts, made up of emergency medicine physicians and emergency nurses from developing and developed countries, were chosen from countries where triage scales were already established and validated or being established and validated. They had already independently reviewed the vignettes used in the current study, and via a modified Delphi technique, obtained consensus on 'true' acuity level for each vignette. They assigned an acuity level based on their expert opinion rather than through the application of the SATS. The acuity levels that they assigned had to fall into one of four categories to mirror the SATS categories of 'emergency', 'very urgent', 'urgent' and 'routine'.

Data analysis
In accordance with the Guidelines for Reporting Reliability and Agreement Studies (GRRAS), inter-rater reliability was assessed using the unweighted, linearly weighted and quadratically weighted κ (QWK) statistic, as well as the intraclass correlation coefficient (ICC).[14] The QWK is commonly used when reporting on reliability studies because it takes into account the degree of disagreement. A weighted κ uses maximum weights at two opposite ends of the scale and is therefore identical to the ICC.[10] Whereas the unweighted and linear weighted κ is not commonly used in triage literature, it has been reported in this case to follow the GRRAS for easy comparisons between other studies.[14] Point estimate values for QWK and ICC were graded using the Landis and Koch classification system as follows: 0.0 - 0.20 – slight agreement; 0.21 - 0.40 – fair agreement; 0.41 - 0.60 – moderate agreement; 0.61 - 0.80 – substantial agreement; and 0.81 - 1.00 – almost perfect agreement.[10] Intra-rater reliability was assessed by calculating the percentage of exact agreement and also the percentage of agreement allowing for one level of discrepancy in the triage ratings.

The accuracy of the nurse triage ratings was assessed by calculating the sensitivity, specificity, and associated over-/under-triage relative to the experts' triage ratings. Over- and under-triage were interpreted using an accepted range for average under-triage of not more than 5 - 10% and an

associated average over-triage rate of 30 - 50%; these are the ranges considered acceptable by the American College of Surgeons Committee on Trauma.[7] Data were analysed using STATA (version 9.2).[15]

Ethics approval

Ethics approval was obtained from the MSF Ethics Review Board, Geneva, Switzerland, and the Human Research Ethics Committee, University of Cape Town, as well as the Pakistan Bioethics Review Board. Informed consent was obtained from all nurses participating in the study.

Results

Characteristics of the study population

Of a total of 23 nurses carrying out triage, 20 met the study inclusion criteria and were invited to participate in the study. Fifteen of these nurses agreed to participate, while five declined due to scheduling conflicts and transport issues. The convenience sample therefore represented 75% of all eligible triage nurses.

Reliability of nurse triage ratings

A total of 780 ratings were obtained for analysis, consisting of 15 nurses assigning ratings for 42 vignettes (n=630) and the same 15 nurses assigning ratings for the 10 duplicate vignettes (n=150). Table 1 summarises the different reliability measures calculated to assess inter- and intra-rater reliability. Inter-rater reliability, as measured by the ICC and QWK, was substantial. Similarly, the level of exact intra-rater agreement among the nurses in our study was almost perfect (87%; 95% confidence interval (CI) 67 - 100), and there was 100% agreement when allowing for a one-level discrepancy in triage ratings.

Accuracy of nurse triage ratings

Table 2 summarises the accuracy of the nurse acuity ratings using the SATS, compared with the expert panel ratings of the vignettes. Overall, the SATS demonstrated a high level of specificity (97%) and a moderate level of sensitivity (70%). Broken down by acuity level, the SATS showed the highest sensitivity (93%) for 'very urgent' cases. However, the

level of sensitivity for 'emergency' cases was exceptionally low (34%). Across all acuity levels, over-triage rates did not exceed the acceptable threshold of 30 - 50%. Similarly, for 'very urgent', 'urgent' and 'routine' cases, under-triage rates were below the acceptable threshold (5 - 10%). However, for emergency cases, the rate of under-triage was exceptionally high (66%), although almost all of these mis-triaged cases were only under-triaged by one acuity level, being rated as 'very urgent'.

Discussion

This is the first study to assess the reliability and accuracy of nurse triage ratings using the SATS in a resource-poor Asian setting.[7] Nurse ratings using this triage scale demonstrated good inter- and intra-rater reliability and acceptable accuracy for 'very urgent' and 'routine' cases. However, nearly two-thirds of 'emergency' cases were under-triaged as 'very urgent', which warrants attention.

Supported by study findings from Botswana and SA,[6,8] our study demonstrates that after minimal formal training, the SATS can be applied reliably by nursing staff in an ED in Pakistan. However, there are concerns about the accuracy of these ratings. In our study, the degree of accuracy of the nurse triage ratings using the SATS was acceptable for 'very urgent' and 'routine' cases, but not for 'urgent' and 'emergency' cases. In particular, a high proportion of emergency cases were under-triaged, which mirrors the findings from a study in SA evaluating the validity of the SATS.[13] The under-triage of 'emergency' cases may be reflected inaccurately on account of several study biases which we discuss below. Alternatively, it may be that this is really the case. If so, this could either be because nursing staff are applying the SATS inaccurately, or because the SATS is poorly constructed to accurately identify true emergency cases. We suspect

---

Table 1. Different measures calculated to assess inter- and intra-rater reliability of ED nurse triage ratings using the SATS at Timergara Hospital, Pakistan

Reliability measure	Point estimate (95% CI)	Level of agreement*
Inter-rater reliability		
Intra-class correlation coefficient	0.77 (0.69 - 0.85)	Substantial
κ statistic		
Unweighted	0.55 (0.51 - 0.60)	Moderate
Linearly weighted	0.65 (0.61 - 0.71)	Substantial
Quadratically weighted	0.77 (0.69 - 0.84)	Substantial
Intra-rater reliability		
Exact agreement, % (95% CI)	87 (67 - 100)	-
Agreement with one SATS category discrepancy, %	100	-

ED = emergency department; SATS = South African Triage Scale; CI = confidence interval.
*According to the Landis and Koch criteria.[10]

Table 2. Comparison of TH ED nurse ratings using the SATS with the expert panel's ratings of the vignettes

Expert panel triage category	Vignettes, n	Triage ratings, n	Nurse ratings, % (N=630)				SATS performance v. expert panel (reference standard)			
			Emergency	Very urgent	Urgent	Routine	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Over-triage, % (95% CI)	Under-triage, % (95% CI)
Emergency	9	135	34*	0.4	1	0	34 (30 - 38)	99 (99 - 100)	0	66 (62 - 70)
Very urgent	17	255	64	93*	34	6	93 (91 - 95)	97 (95 - 98)	0.4 (0 - 1)	7 (5 - 9)
Urgent	10	150	2	4	59*	17	59 (55 - 63)	94 (93 - 96)	35 (33 - 37)	7 (5 - 9)
Routine	6	90	0	3	7	78*	78 (75 - 81)	97 (96 - 98)	22 (19 - 25)	0
Mean							70 (66 - 74)	97 (92 - 100)	15 (4 - 25)	22 (9 - 34)

TH = Timergara Hospital; ED = emergency department; SATS = South African Triage Scale; CI = confidence interval.
* Nurse ratings matching the expert panel's rating (reference standard) across each acuity level.

that staff inaccuracy is not to blame, as regular audits of the SATS in Pakistan together with the findings from a previous study have shown a high level of staff accuracy.[13]

If the construct of the SATS itself is responsible for the under-triage of 'emergency' cases, this needs further investigation. The clinical implications of under-triage of 'emergency' cases in our setting are negligible as almost all of the 'under-triaged' emergency cases were rated as 'very urgent', and in the context of TH all 'emergency' and 'very urgent' patients are seen by the same cadre of healthcare workers in the same area and within the same timeframe. Although we do not have data to substantiate this, a 10-minute delay linked to misclassification of 'emergency' to 'very urgent' cases is unlikely to have clinical implications. Nonetheless, in a setting where there are clear distinctions between the ways in which 'emergency' and 'very urgent' patients are managed, under-triage in this way needs to be avoided, as it may be associated with poorer outcomes (i.e. a higher risk of mortality, worsening morbidity and additional medical complications). This makes the case for ensuring that any assessment of the SATS is context specific.

Study limitations
A number of study limitations and various methodological issues related to assessing the validity of a triage tool have been brought to our attention by this study.

First, while there is no universally accepted time period recommended between assessments for inter- and intra-rater reliability, 2 - 14 days has been suggested.[10] Owing to ED staff time constraints, we conducted the intra-rater assessments immediately after the inter-rater assessment; this may have led to a recall bias in the response ratings.

Second, although the vignettes were paper based, in the absence of non-verbal patient cues and contextual information, raters' triage decisions may have been affected. That said, a previous study comparing the use of paper-based cases with live ED patients as a way of assessing the inter-rater reliability of a triage tool showed an acceptable level of agreement between the two methods.[16] The main benefits of using paper-based vignettes over real ED cases in LMIC settings is that they provide a cost-effective, time-saving, non-invasive and culturally acceptable way of undertaking this type of study.

Third, the written vignettes were based on ED cases seen in SA, not in the TH ED in Pakistan. In the study by Twomey *et al.*,[13] a set of vignettes ratified by a modified Delphi technique are proposed as a set of reference standard vignettes. Using these vignettes in Pakistan was deemed appropriate due to the following: (*i*) SA and Pakistan are both LMIC settings; (*ii*) the two settings have similar rates of trauma (66 trauma presentations per 1 000 patients in SA and 41/1 000 in Pakistan);[17,18] and (*iii*) the reference vignettes depict similar case presentations. However, the epidemiological pattern of disease is different. In future studies like this, it would seem important to develop specific reference vignettes based on ED cases seen in the actual study setting. This would ensure the use of a better reference standard of comparison adapted to the study context.

Fourth, when comparing nurse acuity ratings using the SATS to acuity ratings assigned by the expert panel, we cannot be sure whether an identified discrepancy between the two was: (*i*) because the nursing staff were not applying the SATS accurately; or (*ii*) because the SATS had poor construct validity – in other words did not measure what it purports to. As indicated earlier, we suspect that staff inaccuracy did not account for many of the observed discrepancies in this study. However, in future studies assessing the

validity of a triage tool, it would be more appropriate to compare the ratings by several SATS experts (using the SATS) to the expert panel ratings (reference standard). This would help to control for the issue of staff error.

Finally, as in other studies, our reference standard was an expert panel that assigned acuity ratings to a series of paper vignettes according to their expert opinion. Almost all of these experts were based in high-income rather than LMIC settings and as such their opinion of 'true' patient acuity level may not have fully reflected the reality as in LMIC settings like Pakistan – they may have tended to over-rate patient acuity, especially at the higher end of the triage spectrum. In conjunction with this, it has been reported that nurses tend to under-rate patient acuity when using paper-based vignettes over live cases.[16] In our particular study, these two factors may have contributed to the under-triage of emergency cases that was reported.

Conclusion
Our study shows that the SATS can be used reliably by nurses in an ED in Pakistan. Our results suggest that the SATS is accurate for very urgent and routine cases but, importantly, may 'under-triage' 'emergency' cases. Although this is unlikely to influence patient outcomes in TH, there may be serious implications in other settings and it therefore merits specific investigation and correction.

References
1. Van Rooyen M, Venugopal R, Greenough PG. International humanitarian assistance: Where do emergency physicians belong? Emerg Med Clin North Am 2005;23(1):115-131. [http://dx.doi.org/10.1016/j.emc.2004.09.006]
2. Hodkinson PW, Wallis LA. Emergency medicine in the developing world: A Delphi study. Acad Emerg Med 2010;17(7):765-774. [http://dx.doi.org/10.1111/j.1553-2712.2010.00791.x]
3. Horne S, Vassallo J, Read J, Ball S. UK triage – an improved tool for an evolving threat. Injury 2013;44(1):23-28. [http://dx.doi.org/10.1016/j.injury.2011.10.005]
4. Robison JA, Ahmad ZP, Nosek CA, et al. Decreased pediatric hospital mortality after an intervention to improve emergency care in Lilongwe, Malawi. Pediatrics 2012;130(3):e676-e682. [http://dx.doi.org/10.1542/peds.2012-0026]
5. Emergency Medicine Society of South Africa. The South African Triage Scale (SATS). http://emssa.org.za/sats/ (accessed 11 March 2014).
6. Twomey M, Wallis LA, Thompson ML, Myers JE. The South African Triage Scale (adult version) provides valid acuity ratings when used by doctors and enrolled nursing assistants. African Journal of Emergency Medicine 2012;2(1):3-12. [http://dx.doi.org/10.1016/j.afjem.2011.08.014]
7. Dalwai M, Tayler-Smith K. Implementation of a triage score system in an emergency room in Timergara, Pakistan. Public Health Action 2013;3(1):43-45. [http://dx.doi.org/10.5588/pha.12.0083]
8. Twomey M, Mullan PC, Torrey SB, Wallis L, Kestler A. The Princess Marina Hospital accident and emergency triage scale provides highly reliable triage acuity ratings. Emerg Med J 2012;29(8):650-653. [http://dx.doi.org/10.1136/emermed-2011-200503]
9. Harrison H-L, Raghunath N, Twomey M. Emergency triage, assessment and treatment at a district hospital in Malawi. Emerg Med J 2012;29(11):924-925. [http://dx.doi.org/10.1136/emermed-2011-200472]
10. Streiner D, Norman G. Health Measurement Scales. A Practical Guide to Their Development and Use. 4th ed. New York: Oxford University Press, 2008.
11. Twomey M, Wallis LA, Myers JE. Limitations in validating emergency department triage scales. Emerg Med J 2007;24(7):477-479. [http://dx.doi.org/10.1016/emj.2007.046383]
12. Olofsson P, Gellerstedt M, Carlström ED. Manchester Triage in Sweden – interrater reliability and accuracy. Int Emerg Nurs 2009;17(3):143-148. [http://dx.doi.org/10.1016/j.ienj.2008.11.008]
13. Twomey M, Wallis L, Myers J. Evaluating the construct of triage acuity against a set of reference vignettes developed via modified Delphi method. Emerg Med J 2013. [http://dx.doi.org/10.1136/emermed-2013-202352]
14. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 2011;64(1):96-106. [http://dx.doi.org/10.1016/j.jclinepi.2010.03.002]
15. StataCorp. 2005. Stata Statistical Software: Release 9. College Station, TX: StataCorp LP.
16. Worster A, Sardo A, Eva K, Fernandes CMB, Upadhye S. Triage tool inter-rater reliability: A comparison of live versus paper case scenarios. J Emerg Nurs 2007;33(4):319-323. [http://dx.doi.org/10.1016/j.jen.2006.12.016]
17. Wallis LA, Twomey M. Workload and casemix in Cape Town emergency departments. S Afr Med J 2007;97(12):1276-1280.
18. Nasrullah M, Xiang H. The epidemic of injuries in Pakistan – a neglected problem. J Pak Med Assoc 2008;58(8):420-421.