Dr SW Choi,

The University of Hong Kong,

Department of Anaesthesiology,

Queen Mary Hospital,

102, Pokfulam Road,

HKSAR

Validation of Surgical Outcome Risk Tool (SORT) on Hepatectomy Surgery in a Hong Kong Chinese Population – Should we just be looking at the Receiver Operating Characteristic Curve?

G.T.C.WONG[1], W.C ANG[2], T.C.L. WONG[3], S.W CHOI[4]

1 Clinical Associate Professor, 2 Medical Student, 3 Clinical Assistant Professor, 4 Assistant Research Officer, The University of Hong Kong, Department of Anaesthesiology

Correspondence to: Dr S.W CHOI, E-Mail: htswchoi@hku.hk

G.T.C Wong: gordon@hku.hk

W.C ANG: ang2020@connect.hku.hk

T.C.L. Wong: wongtcl@hku.hk

Article type: Letter to the Chief Editor

Keywords: Surgical risk assessment, receiver operating characteristic, validation, hemihepatectomy

Hepatic resection is currently accepted as the mainstay of curative therapy for localised hepatocellular carcinoma or metastatic disease. Evaluation of the patients' general fitness for this type of major surgery is critical since post operative complications is an important cause of postoperative mortality. The Surgical Outcome Risk Tool (SORT) [1] is a simple to use risk assessment scoring system that is available for the UK population to assess risk of 30 day mortality. Unlike the Physiological and Operative Severity Score (POSSUM) [2] the SORT does not require extensive laboratory data for its calculation. However, the SORT has not been validated in populations outside the UK. Due to the high prevalence of hepatitis B infection, the age-standardised incidence rate of hepatocellular carcinoma in Hong Kong is over ten times that of the UK [3]. It is against this background that we set out to validate the SORT in a population of patients undergoing hemihepatectomy in a tertiary hospital in Hong Kong.

After obtaining IRB approval, anonymized data from patients who underwent hemihepatectomy from 2010-2016 were retrieved from a database administered by the Department of Surgery, Queen Mary Hospital. No imputation methods were used, and patient data with any of the six required SORT variables missing were excluded from the study. Four hundred and seventy five patients, with a 30-day mortality rate of 0.016% (a total of eight patients), were included in the analysis (Table 1). The SORT considers the following variables: severity, ASA-PS classification, urgency, risk of surgery, presence of cancer, and age. Hemihepatectomy is categorized under high-risk surgery and all the patients were classified under the same severity and urgency variables, (i.e. Xmajor/Complex and Expedited, respectively). All cases were cancer patients, predominantly hepatocellular carcinoma. Table 1 shows the patient characteristics according to the six SORT variables.

Together with the actual 30-day mortality data, the discriminatory accuracy of SORT was obtained by calculating the area under the receiver operating characteristic curve (AUROC) using Statistical Package for Social Sciences (SPSS V.22).

The mean SORT risk percentage for hemihepatectomy was 3.94% (±4.24), ranging from 1.25% to 23.06% (Figure 1 and Table 2). With an AUROC of 0.822 (p=0.000), the SORT score should predict risk of death in our population well (Figure 1) but, the predicted number of deaths at 30 days is 17, whereas the actual number of deaths in our population was 8, indicating that the SORT over-estimated number of deaths by about 43%. The co-existence of a high AUROC curve and an over-estimation of 30-day mortality of over 40% seem to be at odds. So what exactly does the AUROC curve tell us, and can we rely on this to give us information about how good a risk assessment tool is?

AUROC is the usual metric used in validation of scores for risk assessment and it measures discrimination [4]. Discrimination, in practical terms, means if you were to randomly pick out two patients from the group, the one who died at 30 days (the "case") should have the higher SORT score. Perfect discrimination (AUROC = 1) is achieved if the score is higher in the case compared with the non-case 100 % of the time. Of course, the score has no discriminative powers if it is correct only in 50% of the time (AUROC =0.5), ie, no better than a coin toss. An AUROC curve of 0.82 as seen in the current study means that that if you pick any two patients out from the group, there is an 82% chance the case had a higher risk score than the non-case (the survivor).

The AUROC also gives you some information about how well the model can rank order cases and non-cases, but ranking order overall is not always useful. Using SORT as an example, a model can have perfect discrimination even if every case had a SORT score of say 2% and every non-case of 1%. Along the same lines, a model may also have perfect discriminant capability even if it assigned every case with a SORT score of anything from 3 to 30%, and every non-case with a score of 1% or less. In this way, the model has assigned someone with a score of 5% the same clinical importance as someone with a score of say 23%. Clearly using these numbers may not be meaningful when trying to explain risk of mortality to the patient. As the AUROC is an aggregate, for anything less than perfect discrimination one cannot tell whether the model performs better with high or low SORT scores. Therefore looking at the AUROC curve alone does not tell us whether the discrimination capacity is large enough to permit useful distinctions between those patients at high risk, and those at low risk of death post-surgery.

It is suggested here, that when comparing, or validating models of risk assessment, tests of model fit, such as likelihood ratio test and Hosmer-Lemeshow test should also be considered [5]. Whilst these tests have been criticized, they nevertheless provide a useful measure to identify accuracy in risk assessment models according to deciles of risk. In the setting of prospective risk prediction, the proportion of patients classified correctly (similar to table 2), rather than the AUROC curve, would seem to have more clinical relevance, and this data should be presented in any publication validating risk assessment tools.

Figure legends

Figure 1. Bar chart showing distribution of risk scores among 475 patients undergoing hemihepatectomy

Figure 2. Area under Receiver Operating Characteristics (AUROC) for Surgical Outcome Risk Tool (SORT) predicting the 30-day mortality of hemihepatectomy patients. AUROC is 0.822 (95% CI, 0.728 to 0.916).

Disclosures: SWC is statistical advisor at *Anaesthesia***.** None of the other authors have any conflict of interests.

Table 1. Descriptive Patient Characteristics according to the six SORT variables

	No. of patients
All patients	475
Mortality	8
Age	
<65	308
≥65 and <80	156
≥80	11
ASA-PS grade	
I	22
II	300
III	144
IV	9
V	0
Urgency of surgery	
Elective	0
Expedited	475
Urgent	0
Immediate	0
Severity	
Minor	0
Intermediate	0
Major	0
Xmajor/complex	475
Cancer	475

Cholangio CA	41
Liver metastases	85
HCC	337
Mixed CA	7
Necrotic Tumour	2
RPC + Cholangio CA	3

ASA-PA, American Society of Anesthesiologists Physical Status; Xmajor, extra major; Cholangio CA, Cholangiocarcinoma; HCC, Hepatocellular carcinoma; Mixed Ca, Mixed Hepatocholangiocarcinoma; RPC, Recurrent Pyogenic Cholangitis.

Table 2. Predicted and observed frequencies of mortality at 30 days post-surgery

SORT Score	No. of patients (%)	No. of predicted deaths at 30 days post-surgery	No. of observed deaths at 30 days post-surgery
1.25	235 (49.47)	3	0
2.68	82 (17.26)	2	1
4.93	70 (14.74)	3	2
5.85	5 (1.05)	0	1
10.14	68 (14.32)	7	2
12.11	3 (0.63)	0	1
20.29	6 (1.26)	1	1
23.06	6 (1.26)	1	0
	Total	17	8

References

1. Protopapa KL, Simpson JC, Smith NCE, Moonesinghe SR. Development and validation of the Surgical Outcome Risk Tool (SORT). *The British Journal of Surgery* 2014; **101:** 1774-83.
2. Brooks MJ, Sutton R, Sarin S. Comparison of Surgical Risk Score, POSSUM and p-POSSUM in higher-risk surgical patients. *Br J Surg* 2005; **92:** 1288-92.
3. Nordenstedt H, White DL, El-Serag HB. The changing pattern of epidemiology in hepatocellular carcinoma. *Digestive and Liver Disease* 2010; **42:** S206-S14.
4. Thompson ML, Zucchini W. On the statistical analysis of ROC curves. *Statistics in Medicine* 1989; **8:** 1277-90.
5. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ : British Medical Journal* 2015; **351**.

Figure 1.

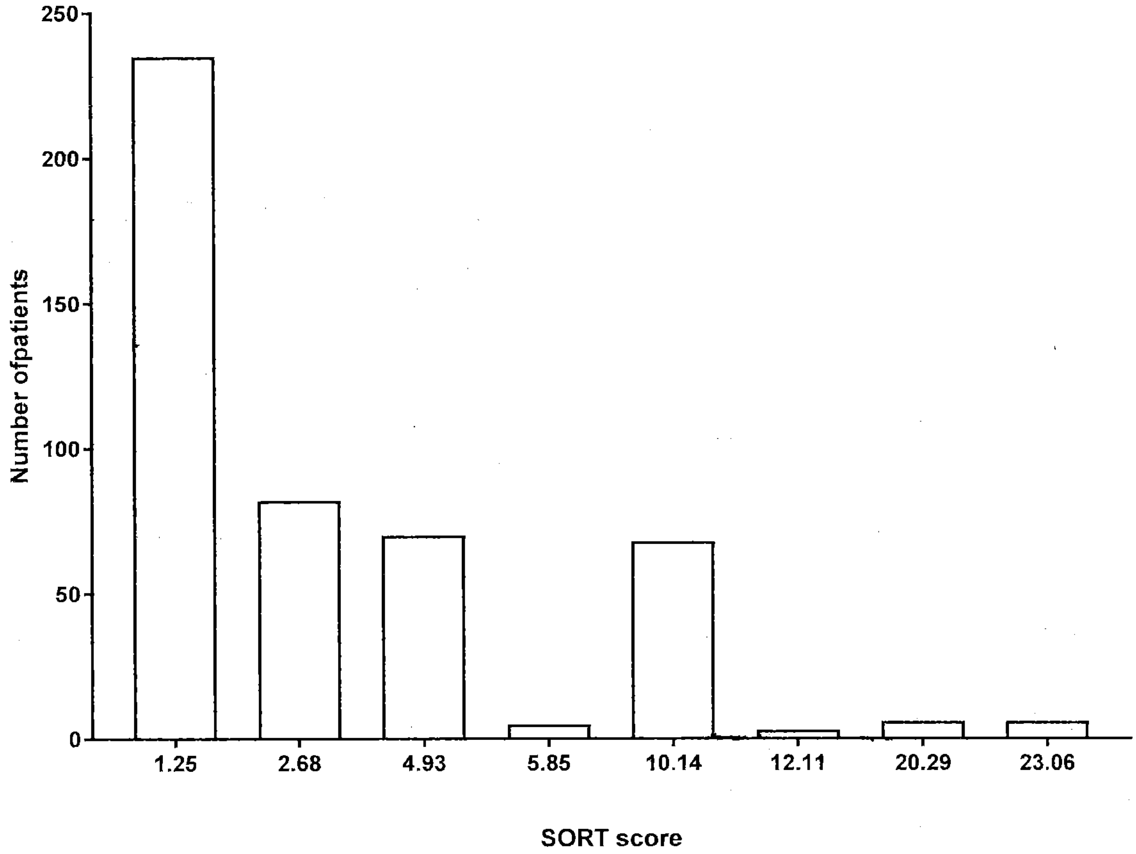
250
200
150
100
50
0
Number ofpatients
1.25
2.68
4.93
5.85
10.14
12.11
20.29
23.06
SORT score

Figure 2.