

prediction probability for each patient by exponentiation of the RSI (inverse logit;  $P_i = 1/[1 + e^{-RSI_i}]$ );  $P_i$  ranges from 0 to 1 (open interval). For each patient, prediction probability  $P_i$  is compared with the observed dichotomous outcome  $Y_i = 0$  (dead) or  $Y_i = 1$  (alive). Overall performance of RSI is measured by the distance of the predicted outcome ( $P_i$ ) from the actual outcome ( $Y_i$ ); a good model of risk will have a short average distance. The accepted measures for overall performance in the validation datasets are the Brier score and the Nagelkerke  $R^2$  statistic.<sup>2</sup> Overall performance can be partitioned into two characteristics: discrimination and calibration. Statistical software tools for estimation of overall performance, discrimination, and calibration are readily available.

The c statistic is a measure of discrimination; it is a rank order statistic for predictions *versus* actual outcomes and is equivalent to the area under the receiver operating characteristic curve. As rank order statistics are invariant under monotonic transformations, the c statistic of RSI is identical to the c statistic of  $P_i$ . Perfect discrimination corresponds with a c statistic of 1 and is achieved if the  $P_i$  or RSI scores for all patients dying are higher than those for all patients not dying, with no overlap. A c statistic value of 0.5 indicates an RSI without discrimination (*i.e.*, no better than flipping a coin). While a good risk model will have high discrimination, by itself the c statistic is not optimal in assessing or comparing risk models.<sup>3</sup>

The third aspect of performance measures is calibration (*i.e.*, the agreement between observed outcomes and predictions). For example, if an RSI score has a predicted probability of 20% for in-hospital mortality, then approximately 20% of inpatients with that RSI score are expected to die. The calibration of prediction probability can be assessed by regression plots of  $Y_i$  *versus*  $P_i$ , with patients grouped by deciles; there is also a specialized binary regression method.<sup>4</sup>

Sessler *et al.*<sup>1</sup> should be congratulated for their statistical models of risk that may, in the future, allow comparisons of outcomes of health care across institutions. I hope that they will provide supplementary analyses to demonstrate that, besides good discrimination, their RSIs also have good calibration and overall performance.

**Nathan L. Pace, M.D., M.Stat.,** University of Utah, Salt Lake City, Utah. n.l.pace@utah.edu

## References

1. Sessler DI, Sigl JC, Manberg PJ, Kelley SD, Schubert A, Chamoun NG: Broadly applicable risk stratification system for predicting duration of hospitalization and mortality. *ANESTHESIOLOGY* 2010; 113:1026–37
2. 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
3. Cook NR: Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; 115:928–35
4. Cox DR: Two further applications of a model for binary regression. *Biometrika* 1958; 45:562–5

(Accepted for publication March 30, 2011.)

## In Reply:

In a recent report, we describe risk stratification indices (RSIs) for mortality and duration of hospital stay.<sup>1</sup> We reported the C statistic along with graphical receiver operating characteristic curves to assess the performance of these predictive models on a prospective validation data set. Pace correctly points out that the C statistic is a measure of model discrimination and that a complete validation also requires an assessment of calibration.

The RSI for in-hospital mortality is derived using a logistic model and therefore the C statistic is an appropriate metric of discrimination. The RSIs of 30-day mortality, 1-yr mortality, and 30-day discharge, however, are derived using Cox proportional hazards modeling. For these, a more appropriate measure of discrimination is Harrell's C (concordance) index, which is defined as the proportion of all usable data samples in which the predictions and the outcomes are concordant.<sup>2</sup> Although the C statistic is defined for dichotomous outcomes, the C index is more broadly applicable, being appropriate for censored time-to-event response variables as well as continuous and ordinal outcomes.

We calculated the C index for each of these three RSIs on the Cleveland Clinic validation data set using a bootstrap methodology to estimate the 95% confidence intervals. The C indices (table 1) were nearly identical to the previously reported C statistics—although with somewhat wider confidence intervals—thus revealing good discrimination across all four RSI models.

As suggested by Pace, we assessed calibrations of the RSI models on the Cleveland Clinic validation data set by means of calibration graphs, which are graphical representations of the Hosmer-Lemeshow goodness-of-fit test.<sup>3</sup> These were constructed by grouping patients into approximately equal-size bins of equivalent RSI values. The number of bins was chosen to achieve as even a distribution of patients among bins as possible, given the existence of ties. The mean RSI within each bin was then plotted against the mortality rate or mean length-of-stay within that bin.

The graphs indicate good calibration across the four RSI models (fig. 1), with mortality and extended-stay events most prevalent in the higher predicted-risk groups. (There is no expectation of linearity in these plots; goodness of calibration is indicated by monotonic left-to-right increases.) The low event rate for the in-hospital mortality endpoint results in very few events in the lower predicted risk groups; this gives rise to the "hockey stick" appearance of the graph. As these

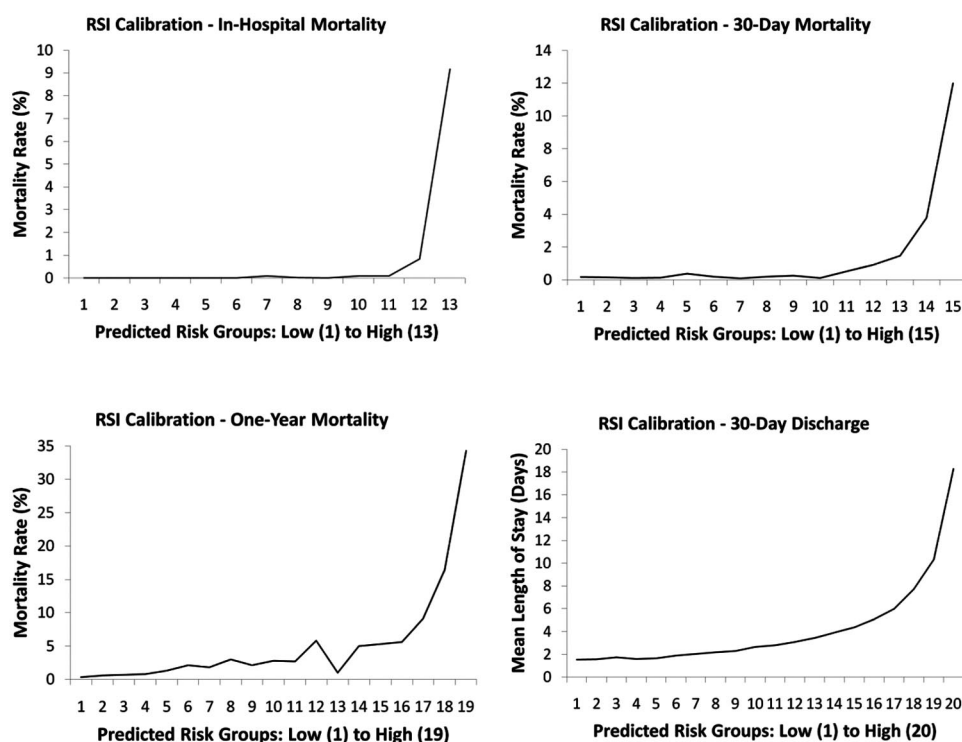
**Table 1.** Prospective Assessment of RSI Model Discrimination

Cleveland Clinic Validation Data Set				
—	In-hospital Mortality (C-statistic, 95% CI)	30-day Mortality (C-Index, 95% CI)	1-yr Mortality (C-Index, 95% CI)	30-day Discharge (C-index, 95% CI)
Demographics	0.684 (0.670–0.698)	0.705 (0.592–0.822)	0.681 (0.645–0.718)	0.472 (0.460–0.483)
CCI	0.654* (0.640–0.669)	0.759 (0.656–0.857)	0.761* (0.725–0.797)	0.431* (0.420–0.440)
CCI + demographics	0.711 (0.697–0.724)	0.803 (0.710–0.895)	0.792* (0.760–0.822)	0.444* (0.432–0.454)
RSI	0.977*†‡ (0.975–0.980)	0.847 (0.737–0.948)	0.829*† (0.800–0.856)	0.813*†‡ (0.805–0.819)
RSI + demographics	0.979*†‡ (0.977–0.981)	0.873 (0.792–0.954)	0.853*†‡ (0.826–0.878)	0.794*†‡§ (0.786–0.802)

Demographics are age, sex, and race.

\*  $P < 0.05$  compared with demographics alone. †  $P < 0.05$  compared with Charlson comorbidity index (CCI). ‡  $P < 0.05$  compared with CCI + demographics. §  $P < 0.05$  compared with RSI.

RSI = risk stratification indices.



**Fig. 1.** Calibration curves for 30-day discharge and for in-hospital, 30-day, and 1-yr mortality.

events accumulate in the 30-day and 1-yr mortality graphs, more events occur in the lower predicted risk groups, and the calibration graphs become smoother. The continuous end-point in the 30-day discharge graph yields a smooth calibration relationship.

In summary, our risk stratification models exhibit both good discrimination and calibration. The indices can thus be used to adjust for differences in baseline and procedural risk, permitting fair outcome comparisons among hospitals and practices. We have put the system in the public domain to facilitate use; details, including model coefficients, SPSS programs, and sample files, are available at the Web site.\*

**Jeffrey C. Sigl, Ph.D., Daniel I. Sessler, M.D.,† Scott D. Kelley, M.D., Nassib G. Chamoun, M.S.** †The Cleveland Clinic, Cleveland, Ohio. ds@or.org

## References

1. Sessler DI, Sigl JC, Manberg PJ, Kelley SD, Schubert A, Chamoun NG: Broadly applicable risk stratification system for predicting duration of hospitalization and mortality. *ANESTHESIOLOGY* 2010; 113: 1026–37
2. Harrell FE Jr, Lee KL, Mark DB: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15:361–87
3. 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

\* [www.ClevelandClinic.org/RSI](http://www.ClevelandClinic.org/RSI). Accessed March 28, 2011.

(Accepted for publication March 30, 2011.)