In fact, increasing evidence points toward additional clinically relevant problems with commonly used sedative drugs. Benzodiazepines, among the most common drugs in our arsenal, contribute to the development of delirium after ICU sedation.[6] Delirium is associated with increased hospital length of stay and with increased mortality.[7] Propofol, common in adult ICUs despite the above mentioned concerns, is not recommended for long-term sedation in children or in higher infusion rates for adults because of the risk of propofol infusion syndrome.[8] Moreover, long-term use of propofol may contribute to withdrawal.[9]

Several studies of volatile anesthetics for sedation purposes in humans—with clinically relevant endpoints—have shown promising results. Rapid pulmonary excretion and limited metabolism of all the modern agents are intrinsically attractive characteristics. Wake-up times are shorter and more predictable than with intravenous sedatives, as is time to cooperation.[10,11] There may be beneficial cardiac effects of volatile anesthetic sedation.[12] The memory panorama from the ICU stay, an important patient-related outcome,[13] also appears to be favorable compared with that of midazolam.[14]

Simply put, we need more evidence and knowledge about the advantages and risks of the sedative drugs that we use, be they benzodiazepines and propofol or volatile anesthetics. We advocate for additional evaluation of volatile anesthetics as a promising option for long-term sedation in ventilator-dependent ICU patients. In any case, we can not afford to idly administer routine cocktails of sedatives unaware of the risks we may be taking. Every patient deserves a carefully considered sedation strategy.

Peter V. Sackey, M.D., Ph.D.,* Peter J. Radell, M.D., Ph.D. *Karolinska University Hospital, Stockholm, Sweden. peter.sackey@karolinska.se

References

1. Sackey PV, Eriksson LI, Martling CR, Radell PJ: Tailored sedation to the individual needs of the intensive care unit patient. ANESTHESIOLOGY 2010; 113:1439–46
2. Jevtovic-Todorovic V, Benshoff N, Olney JW: Ketamine potentiates cerebrocortical damage induced by the common anaesthetic agent nitrous oxide in adult rats. Br J Pharmacol 2000; 130:1692–8
3. Young C, Jevtovic-Todorovic V, Qin YQ, Tenkova T, Wang H, Labruyere J, Olney JW: Potential of ketamine and midazolam, individually or in combination, to induce apoptotic neurodegeneration in the infant mouse brain. Br J Pharmacol 2005; 146:189–97
4. Bercker S, Bert B, Bittigau P, Felderhoff-Müser U, Bührer C, Ikonomidou C, Weise M, Kaisers UX, Kerner T: Neurodegeneration in newborn rats following propofol and sevoflurane anesthesia. Neurotox Res 2009; 16:140–7
5. Wilder RT, Flick RP, Sprung J, Katusic SK, Barbaresi WJ, Mickelson C, Gleich SJ, Schroeder DR, Weaver AL, Warner DO: Early exposure to anesthesia and learning disabilities in a population-based birth cohort. ANESTHESIOLOGY 2009; 110:796–804
6. Pandharipande P, Shintani A, Peterson J, Pun BT, Wilkinson GR, Dittus RS, Bernard GR, Ely EW: Lorazepam is an independent risk factor for transitioning to delirium in intensive care unit patients. ANESTHESIOLOGY 2006; 104:21–6
7. Ely EW, Shintani A, Truman B, Speroff T, Gordon SM, Harrell FE Jr, Inouye SK, Bernard GR, Dittus RS: Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. JAMA 2004; 291:1753–62
8. Vasile B, Rasulo F, Candiani A, Latronico N: The pathophysiology of propofol infusion syndrome: A simple name for a complex syndrome. Intensive Care Med 2003; 29:1417–25
9. Liatsi D, Tsapas B, Pampori S, Tsagourias M, Pneumatikos I, Matamis D: Respiratory, metabolic and hemodynamic effects of clonidine in ventilated patients presenting with withdrawal syndrome. Intensive Care Med 2009; 35:275–81
10. Röhm KD, Wolf MW, Schöllhorn T, Schellhaass A, Boldt J, Piper SN: Short-term sevoflurane sedation using the Anaesthetic Conserving Device after cardiothoracic surgery. Intensive Care Med 2008; 34:1683–9
11. Meiser A, Sirtl C, Bellgardt M, Lohmann S, Garthoff A, Kaiser J, Hügler P, Laubenthal HJ: Desflurane compared with propofol for postoperative sedation in the intensive care unit. Br J Anaesth 2003; 90:273–80
12. Hellström J, Öwall A, Bergström J, Sackey PV: Cardiac outcome after sevoflurane *versus* propofol sedation following coronary bypass surgery: A pilot study. Acta Anaesthesiol Scand 2011; 55:460–7
13. Wunsch H, Kress JP: A new era for sedation in ICU patients. JAMA 2009; 301:542–4
14. Sackey PV, Martling CR, Carlswärd C, Sundin O, Radell PJ: Short- and long-term follow-up of intensive care unit patients after sedation with isoflurane and midazolam: A pilot study. Crit Care Med 2008; 36:801–6

(Accepted for publication March 28, 2011.)

Incomplete Validation of Risk Stratification Indices

To the Editor:

Using 2001–2006 Medicare hospital data (Medicare Provider Analysis and Review [MEDPAR]) in approximately 17 million patients aged 65 yr and older, Sessler and colleagues[1] have proposed four risk stratification indices (RSIs) for mortality and duration of hospital stay. With a complex, stepwise hierarchical-selection algorithm, the authors[1] chose a parsimonious set of statistically significant predictors from the approximately 20,500 International Classification of Diseases, Ninth Revision, diagnostic and procedure codes. For example, in-hospital mortality was modeled on 184 predictor codes with odds ratios varying from 0.131 to 57.821. Using a split sample design, these RSIs were internally validated on MEDPAR data for another 17 million patients and were externally validated on 100,000 patient records from the Cleveland Clinic (Ohio; Perioperative Health Documentation System). Working in the parameter space (β coefficients), validation of the RSIs was demonstrated on the development, validation, and external datasets by the c (concordance) statistic,[2] which revealed very good discrimination in all datasets.

However, the performance of these RSIs has not been adequately justified. To do so requires calculation of the

prediction probability for each patient by exponentiation of the RSI (inverse logit; $P_i = 1/[1 + e^{-RSI}]$); 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.[2] 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.[3]

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.[4]

Sessler *et al.*[1] 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.[1] 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.[2] 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.[3] 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