

Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Kim WR, Biggins SW, Kremers WK, et al. Hyponatremia and mortality among patients on the liver-transplant waiting list. *N Engl J Med* 2008;359:1018-26.

On-line Appendix

The Hosmer-Lemeshow statistic is a calibration measure that typically partitions the data into 10 equal sets (deciles) according to the model's predicted probabilities and then compares the observed number of events in each decile with the expected number of events defined as the sum of the predicted probabilities in each decile. For validation, the statistic is applied to the model validation data set broken into deciles defined by the data set used in developing the model. Thus, the Hosmer-Lemeshow statistic is a measure of the discrepancy between the observed and predicted. A better calibrated model would have a smaller discrepancy between the observed and predicted and thus a smaller Hosmer-Lemeshow statistic.

For survival data, the observations with incomplete follow-up (censored) require a modification of the usual formula. May and Hosmer relate the original methods to the observed/expected cumulative hazard of a Cox model, and show how to compute a revised statistic using standard proportional hazards software.¹ D'Agostino et al show a relation to the expected survival methods of Ederer, and derive a test that is closely related to the one-sample log-rank.^{2,3} In practice, the two methods give very similar answers, as they did on this data set; the May and Hosmer method is easier to apply since it uses standard software.

Figure 3 of the manuscript shows the observed and expected probabilities of death for the 2005 and 2006 data, with the patients broken into deciles by the MELDNa score and expected values based on either the MELD or MELDNa. There are three observations to be made. First, comparison of the observed and predicted by MELDNa between the two figures (blue and red bars) shows that the predicted tracks the observed fairly closely for both data sets. The discrepancy is a bit larger for the validation data set, which is to be expected. Second, in the 2006 data, MELDNa tracks the observed events better than MELD in virtually all deciles. Third, while the fit is very good there is a small systematic departure in the upper two categories: the scores somewhat underestimate risk in group 9 and overestimate it in group 10.

The following table represents the Hosmer-Lemeshow statistics corresponding to the second figure, namely prediction by MELD and MELDNa for the 2006 in comparison to the observed. These analyses clearly show that MELDNa is better calibrated than MELD.

Expected death by	Hosmer-Lemeshow	P
MELD	51.3	0.000
MELDNa	17.8	0.023

The integrated discrimination improvement (IDI) has been proposed by Pencina et al as a measure of discrimination, ability of the model to rank subjects correctly according to their outcome.⁴ In comparing the discrimination of a new model to that of an old model, a given subject's rank may go up or down. Comparison of the concordance statistics, equivalent to the area under the receiver-operating characteristic curve, captures the proportion of subjects who moves (up or down) in the correct direction, without regard to the magnitude of the change.

IDI is defined as $(IS_{new} - IS_{old}) - (IP_{new} - IP_{old})$, where IS is the integral of sensitivity over all possible cut-off values from the (0, 1) interval and IP the corresponding integral of 'one minus specificity'. The subscript 'new' in the expression refers to the model with a new variable and subscript 'old' to the model without it. Thus, it quantifies jointly the overall improvement in sensitivity and specificity.

Mathematically, IDI can be calculated as the standard error of paired differences of new and old model-based predicted probabilities across all event subjects, \widehat{se}_{events} . Denoting the corresponding estimator for non-events by $\widehat{se}_{nonevents}$ and assuming independence between events and non-events and their predicted probabilities, a simple asymptotic test for the null hypothesis of IDI=0 can be obtained by:

$$z = \frac{IDI}{\sqrt{(\widehat{se}_{events})^2 + (\widehat{se}_{nonevents})^2}}$$

In our 2006 data, there were 477 who died within 90 days of registration, whereas the remaining 6694 survived. The paired differences between probability of death predicted by MELDNa and that by MELD among the dead had a mean of 0.012 (range: -0.108 – 0.203). Among those who survived, the mean difference between prediction by MELDNa and that by MELD was 0.001. Based on these numbers, IDI is calculated as 0.011 ($p < 0.001$).

References

1. May S, Hosmer DW. A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model. . *Lifetime Data Analysis* 1998;4:109-20.
2. D'Agostino RB, Nam B-H. Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures. In: Balakrishnan N, Rao CR, eds. *Advances in Survival Analysis*. 1st ed. Amsterdam: Elsevier; 2004:1-25.
3. Ederer F, Axtell L, Cutler S. *The relative survival rate: a statistical methodology.*; 1961.
4. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in medicine* 2008;in press.