# Validation and Calibration of the Risk Stratification Index

George F. Chamoun, Linyan Li, M.S., Nassib G. Chamoun, M.S., Vikas Saini, M.D., Daniel I. Sessler, M.D.

## ABSTRACT

**Background:** The Risk Stratification Index was developed from 35 million Medicare hospitalizations from 2001 to 2006 but has yet to be externally validated on an independent large national data set, nor has it been calibrated. Finally, the Medicare Analysis and Provider Review file now allows 25 rather than 9 diagnostic codes and 25 rather than 6 procedure codes and includes present-on-admission flags. The authors sought to validate the index on new data, test the impact of present-on-admission codes, test the impact of the expansion to 25 diagnostic and procedure codes, and calibrate the model.

**Methods:** The authors applied the original index coefficients to 39,753,036 records from the 2007–2012 Medicare Analysis data set and calibrated the model. The authors compared their results with 25 diagnostic and 25 procedure codes, with results after restricting the model to the first 9 diagnostic and 6 procedure codes and to codes present on admission.

**Results:** The original coefficients applied to the 2007–2012 data set yielded C statistics of 0.83 for 1-yr mortality, 0.84 for 30-day mortality, 0.94 for in-hospital mortality, and 0.86 for median length of stay—values nearly identical to those originally reported. Calibration equations performed well against observed outcomes. The 2007–2012 model discriminated similarly when codes were restricted to nine diagnostic and six procedure codes. Present-on-admission models were about 10% less predictive for in-hospital mortality and hospital length of stay but were comparably predictive for 30-day and 1-yr mortality.

**Conclusions:** Risk stratification performance was largely unchanged by additional diagnostic and procedure codes and only slightly worsened by restricting analysis to codes present on admission. The Risk Stratification Index, after calibration, thus provides excellent discrimination and calibration for important health services outcomes and thus appears to be a good basis for making hospital comparisons. **(ANESTHESIOLOGY 2017; 126:623-30)**

PAYERS have an interest in comparing hospital performance to determine which facilities provide the most cost-effective care. Patients are similarly interested in determining which hospitals provide the best outcomes. However, patient populations vary considerably in baseline risk, and procedures differ among hospitals. Fair comparisons thus require accurate risk stratification. Many risk stratification systems require clinical data, which may be difficult to obtain and apply only to specific populations. Relatively few systems apply broadly and are based on only generally available administrative data. Commonly used risk-adjustment systems include the Charlson Comorbidity Index,[1] the Elixhauser Comorbidity Index,[2] the Hierarchical Condition Category,[3] and the American Society of Anesthesiologists physical status score.[4] In the past decade, many new risk-adjustment models have been developed. The Procedural Index for Mortality Risk,[5] the Risk Quantification Index,[6] the Preoperative Score to Predict Postoperative Mortality,[7] and the Surgical Mortality and Probability Model[8] are among a few proposed as potentially robust alternatives to the commonly used methodologies. However, few have been fully validated.

The Risk Stratification Index (RSI, 2010) was developed from the national Medicare Analysis and Provider Review (MEDPAR) data set of 35 million inpatient medical and surgical hospitalizations in patients 65 yr or older from 2001 to 2006.[9] The model used logistic and Cox proportional hazards regression to assign risk scores to International Classification of Diseases, Ninth Revision (ICD-9) codes associated with a hospital stay and was internally validated in a split sample of the population. From a statistical perspective, a model validated on a split random sample with more than 17 million cases is sufficient to demonstrate predictive

### What We Already Know about This Topic

- The Risk Stratification Index, a multivariable risk-adjustment model, was developed from 35 million Medicare hospitalizations from 2001 to 2006 but has yet to be externally validated on an independent large national data set.
- Additionally, the Medicare Analysis and Provider Review file now allows 25 rather than 9 diagnostic codes and 25 rather than 6 procedure codes, and includes present-on-admission flags.
- This study sought to validate the index on new data, test the impact of present-on-admission codes, test the impact of the expansion to 25 diagnostic and 25 procedure codes, and calibrate the model.

### What This Article Tells Us That Is New

- Risk stratification performance was largely unchanged by additional diagnostic and procedure codes and only slightly worsened by restricting analysis to codes present on admission. The Risk Stratification Index, after calibration, thus provides excellent discrimination and calibration for important health services outcomes and metrics.

Submitted for publication August 19, 2016. Accepted for publication January 4, 2017. From the Lown Institute, Boston, Massachusetts (G.F.C., N.G.C., V.S.); Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (L.L.); and Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, Cleveland, Ohio (D.I.S.).

 Anesthesiology 2017; 126:623-30



 10.1097/ALN.0000000000001560

stability.[10,11] Comparable RSI discrimination has since been shown in three relatively small single-center noncardiac surgical populations that were not restricted to Medicare patients.[9,12,13] The method was also validated using the California Inpatient Database from 2004 to 2009, a large population of patients older than 18 yr.[14] However, RSI has yet to be externally validated on an independent large national data set comparable to that of its original developmental data set.

Discrimination, as expressed by the C statistic for example, describes a model's ability to rank order patients or populations with respect to the outcome of interest. However, a model with excellent discrimination may be highly nonlinear for outcome. A consequence is that patients or populations with a given absolute difference in model values may have various actual differences in outcome. Calibration is the process by which predicted outcomes are aligned with observed outcomes. The original RSI models were not calibrated, which restricted real-world application.

Since the original study, coding practices have changed.[15] For example, the 2001–2006 MEDPAR data set was restricted to nine diagnostic and six procedure codes. Consequently, the original RSI development was restricted to a total of 15 codes. But since 2010, Medicare allowed up to 25 diagnostic codes and 25 procedure codes per admission. Including all available codes would be consistent with Medicare's Principles for Risk Adjustment Model Development, which recommends that risk-adjustment models be resistant to variation in code count.[3] Still, whether RSI performs comparably with up to 25 potential diagnostic and 25 potential procedure codes per patient remains unknown.

Furthermore, until 2008, Medicare participants were not required to identify conditions as being present on admission (POA).[16] Diagnostic codes before 2008 thus represented a combination of POA conditions that reflected patients' baseline status and conditions acquired due to hospital-acquired complications or diagnostic oversights. The problem with including diagnostic codes associated with hospital-acquired complications is that the complications are attributed to baseline illness, thus improving apparent performance. An analysis of California inpatients suggests a small, but nontrivial, effect of restricting analysis to POA codes on discrimination for in-hospital mortality. However, it remains unknown to what extent POA codes influence discrimination for 30-day and 1-yr mortality or for hospital length of stay.

We thus sought to validate RSI in a completely new national population to evaluate the effects of using all codes available in the MEDPAR file—a potential 25 *versus* 9 diagnostic codes and a potential 25 *versus* 6 procedure codes—on discrimination and to evaluate the importance of restricting codes to POA codes. We also sought to calibrate the models to provide absolute accuracy as well as discrimination.

## Materials and Methods

### Data

Stay records of 39,753,036 patients from the 2007–2012 MEDPAR files served as our validation data set. The MEDPAR file contains records from fee-for-service inpatient hospitalizations. Each line corresponds to a hospital stay and contains up to 25 diagnostic and 25 procedure ICD-9 codes. We selected this database, as it was the same one used for development of the original RSI. Beneficiaries under age 65 yr and without at least one procedure (surgical or medical) were excluded. Figures 1 to 3 detail the preliminary exclusion criteria.

We included up to all 50 potential codes in our initial calculation of RSI. Thereafter, we restricted the RSI computation to the first nine diagnostic and the first six procedure codes, as in our original developmental data set.

We limited the restricted code analysis to the years 2010 to 2012, as uniform compliance did not begin until 2010. The 2007–2012 MEDPAR file also has indicator fields for POA diagnostic codes. Thus, to test the model prospectively, our final computation of RSI included only diagnostic codes labeled as POA and only procedure codes in the same Clinical Classifications Software[17] (CCS) category as the primary procedure. We used years 2009 to 2012 for POA analysis as this measure was not uniformly available until 2009 (figs. 1 to 3). This left us with three data sets, one with all codes and all years, one with years 2010 to 2012 and 25-code compliance, and one with years 2008 to 2012 and POA coding compliance. We refer to the restricted data sets (figs. 2 and 3) as the compliance data sets.

### RSI Computation

We built the model as specified in the original RSI description (eqn. 1). Briefly, the approximate risk for each ICD-9 risk class *ceteris paribus* from each record in the MEDPAR file was summed into a risk score for that specific hospital stay. This technique is mathematically expressed in equation 1. The ICD-9 RSI classes for each endpoint can be found on http://my.clevelandclinic.org/services/anesthesiology/outcomes-research/risk-stratification-index, under the All Covariates Excel File. The endpoints we validated were 30-day mortality, 1-yr mortality, in-hospital mortality, and length of stay.

Rather than rederive the covariate means for each of the coefficients included in the model, we modified the model's

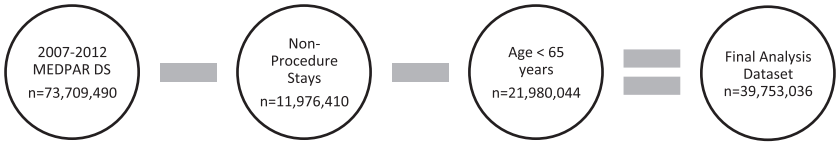


**Fig. 1.** Final analysis data set—preliminary exclusion criteria. The Medicare Analysis and Provider Review (MEDPAR) data set (DS) was used in our analyses.


**Fig. 2.** Twenty-five-code-compliance data set—exclusions. Beginning in 2010, all hospitals were required to expand their diagnostic coding fields from 9 diagnostic and 6 procedure fields to 25 diagnostic and 25 procedure fields. The Medicare Analysis and Provider Review data set was used in our analyses.


**Fig. 3.** Present-on-admission (POA) compliance data set—exclusions. In 2008, POA coding compliance was required for all hospitals. The Medicare Analysis and Provider Review data set was used in our analyses.

offset formula to more simply account for population differences between data sets. Specifically, the original equation was

$$\mathrm{RSI}_{\mathrm{Endpoint}\,i} = \sum_{j=1}^{\mathrm{NCodes}} x_{ij}(\beta_{ij} - \mu_{ij}), \tag{1}$$

where Endpoint = RSI (*i.e.*, 1-yr mortality), $i$ = an individual in our population, $j$ = the list of model covariates (variables), $x$ = a binary variable indicating the presence of $j$ in $i$, $\beta$ = the estimated coefficient of $j$, and $\mu$ = the covariate mean of $j$.

We modified the equation to

$$\mathrm{RSI}_{\mathrm{Endpoint}\,i} = \sum_{j=1}^{\mathrm{NCodes}} \beta_{ij} x_{ij} - \frac{\sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{NCodes}} \beta_{ij} x_{ij}}{\mathrm{N}}. \tag{2}$$

The covariate means are expressed as a population mean, instead of on an individual level, as in equation 1.

This modification is analogous to equation 1 in the sense that it centers the distribution of the betas on 0, while saving tedious and redundant computation. The equation was used to compute risk scores for each hospital stay in the MEDPAR population. We then repeated the procedure, restricting patients to no more than the first nine diagnostic and the first six procedure codes, as used in development of the original RSI model. Finally, we repeated the procedure using only diagnostic codes classified as POA and only procedure codes in the same CCS category as the principal procedure. We thus assumed that secondary procedures within a CCS category were expected based on the primary procedure; other procedures were assumed to reflect a complication.

### Calibration

To provide accuracy as well as discrimination, we developed a new calibration method that was applicable for both logistic regression and Cox regression models (appendix). Using our 2007–2012 MEDPAR data set, we selected a 20% random sample for developing calibration models. As RSI was developed using both Cox and logistic models, our calibration procedures vary slightly between each of the four RSIs (30-day, 1-yr, in-hospital mortality, and length-of-stay measures). For the logistic models, we computed the observed over expected (O/E) plots with respect to the estimated log-odds ratio or risk score:

$$\log \frac{Pr(x)}{1 - Pr(x)} = \beta_o + \beta_1 x_1 + \cdots + \beta_n x_n = \mathrm{RiskScore}. \tag{3}$$

For the Cox models, we used the natural log of the hazard ratio to build our O/E plots:

$$\log \frac{h_i(t)}{h_o(t)} = \beta_1 x_{i1} + \cdots + \beta_n x_{in} = \mathrm{RiskScore}. \tag{4}$$

We found that a logistic hazard function with an additional scalar parameter linearized the in-hospital, 30-day, and 1-yr mortality data well. Due to the slightly unconventional shape of the length-of-stay function, using the odds ratio to estimate the hazard function proved unstable. We therefore fit a cumulative distribution function to a plot of the estimated log-odds length of stay and the corresponding mortality. We found that a scaled extreme-value distribution with an additional offset term best fit the length-of-stay curve. Additional details regarding the development and derivation of the calibration technique are outlined in the appendix.

## Results

Population characteristics were broadly similar in the 2001–2006, 2007–2012, and compliance data sets (figs. 1 to 3). However, beneficiaries in our current validation data set were more than 3 yr older than those in the original RSI developmental data set (table 1).

After calculating RSI for all beneficiaries and for all diagnostic and procedure fields in our data set, we found discrimination was similar in the 2007–2012 and 2001–2006 populations across all endpoints in comparison (table 2). The C statistics for 1-yr and 30-day mortality are both within a half percent of the original scores. For in-hospital mortality, there was a small improvement from the original validation. Discrimination for the length-of-stay model improved 6%.

The calibration methods we developed resulted in the plots shown in figures 4 through 7. The $R^2$ values indicate successful alignment of RSI predictions with observed outcomes using the entire 2007–2012 MEDPAR data set. The calibration models are summarized and characteristic functions are provided in table 3.

Compared to the average of 7.1 ± 2.3 diagnostic codes in the 2001–2006 MEDPAR data set, the current data set had an average of 9.7 ± 4.4 codes, probably reflecting the increase to 25 diagnostic and 25 procedure code slots available. In contrast, the number of procedure codes was virtually identical—even when restricting analysis to the 2010–2012 25-code-compliance data set. Predicted values

**Table 1.** Characterization of Data Sets

| | Original RSI Validation Data Set | 2007–2012 MEDPAR Final Analysis Data Set |
|---|---|---|
| | (N = 17,589,824) | (N = 37,753,036) |
| Age, yr | 74.1 ± 10.2 | 77.11 ± 8.174 |
| Female, % | 54.4 | 55.3 |
| White, black, other, % | 82.5, 12.2, 5.4 | 84.3, 10.4, 5.3 |
| Number of diagnostic codes | 7.1 ± 2.3 | 9.65 ± 4.425 |
| Number of procedure codes | 2.6 ± 1.7 | 2.71 ± 2 |
| Median length of stay (IQR) | 5 (3,8) | 4 (3,8) |
| In-hospital mortality, % | 5.3 | 4.9 |
| 30-day mortality (postdischarge), % | 5 | 4.9 |
| 1-yr mortality (postdischarge), % | 19.3 | 22 |

IQR = interquartile range; MEDPAR = Medicare Analysis and Provider Review; RSI = Risk Stratification Index.

**Table 2.** C Statistics

| | Original RSI Data Set | 2007–2012 Final Analysis Data Set* | 2010–2012 RSI Restricted (9 and 6) Codes† | 2010–2012 25-Code-Compliance Data Set† | 2009–2012 Present-on-Admission Codes‡ |
|---|---|---|---|---|---|
| C Statistic (95% CI) | (N = 17,589,824) | (N = 39,753,036) | (N = 20,311,252) | (N = 20,311,252) | (N = 30,223,443) |
| Length of stay (median) | 0.792 (0.776, 0.827) | 0.851 (0.851, 0.852) | 0.845 (0.845, 0.845) | 0.844 (0.844, 0.844) | 0.741 (0.741, 0.741) |
| 1-yr mortality | 0.833 (0.832, 0.834) | 0.832 (0.832, 0.832) | 0.830 (0.829, 0.830) | 0.833 (0.833, 0.833) | 0.808 (0.808, 0.808) |
| 30-day mortality | 0.859 (0.858, 0.860) | 0.837 (0.837, 0.837) | 0.840 (0.839, 0.840) | 0.837 (0.836, 0.837) | 0.810 (0.810, 0.811) |
| In-hospital mortality | 0.946 (0.945, 0.948) | 0.936 (0.936, 0.936) | 0.938 (0.937, 0.938) | 0.941 (0.941, 0.941) | 0.855 (0.854, 0.855) |

*All codes were used on the entire data set. †In 2010 began the compliance of coding increase to 25 diagnostic and procedure codes. ‡In 2009 began the compliance of present-on-admission indicators.
RSI = Risk Stratification Index.

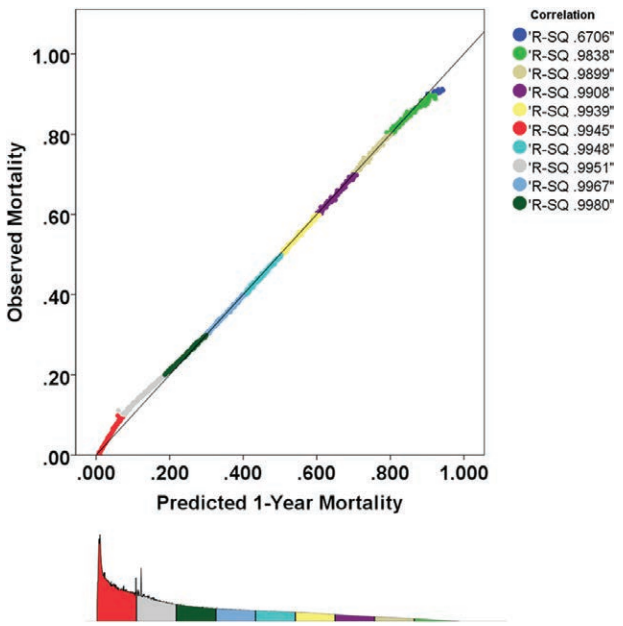


**Fig. 4.** One-year mortality calibration and population density. $R^2$ of 0.67 when predicted mortality was between 0.9 and 1, 0.98 when predicted mortality was between 0.7 and 0.9, and greater than 0.99 for all other predicted mortalities.

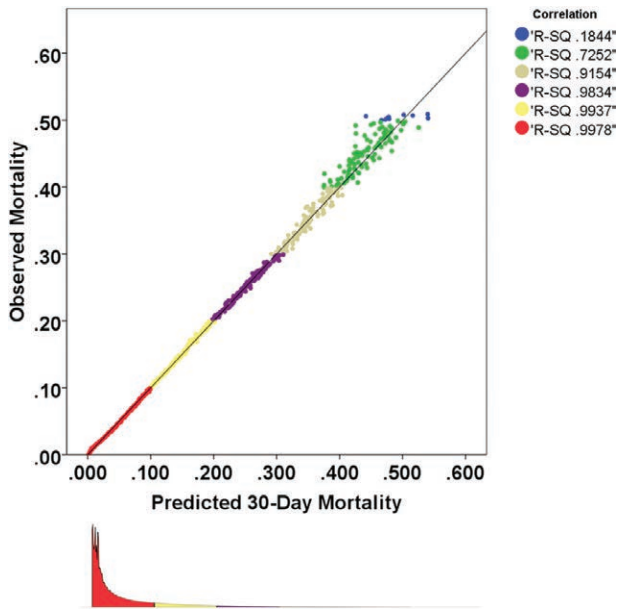


**Fig. 5.** Thirty-day mortality calibration and population density. $R^2$ of 0.18 when predicted mortality was between 0.5 and 0.6, 0.72 when predicted mortality was between 0.4 and 0.5, 0.91 when predicted mortality was between 0.3 and 0.4, 0.98 when predicted mortality was between 0.2 and 0.3, and greater than 0.99 for all other predicted mortalities.

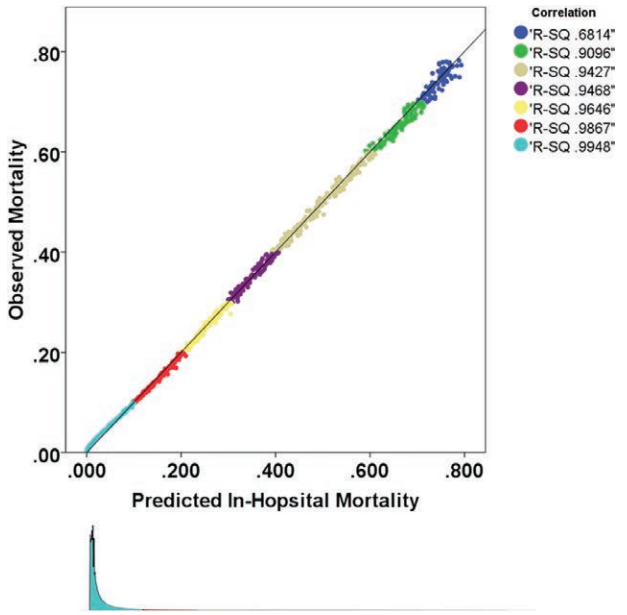


**Fig. 6.** In-hospital mortality calibration and population density. $R^2$ of 0.68 when predicted mortality was between 0.7 and 0.8, 0.90 when predicted mortality was between 0.6 and 0.7, 0.94 when predicted mortality was between 0.3 and 0.6, 0.97 when predicted mortality was between 0.2 and 0.3, and greater than 0.99 for all other predicted mortalities.

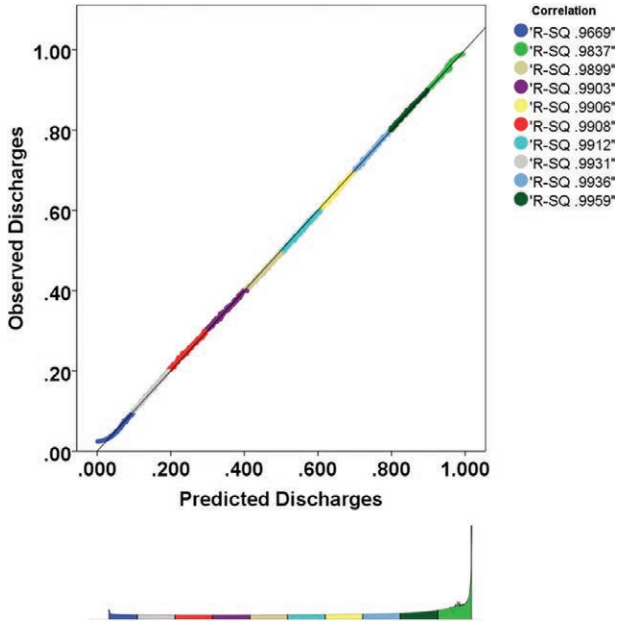


**Fig. 7.** Median length-of-stay calibration and population density. $R^2$ of 0.97 when predicted discharges were between 0.1 and 0.2, 0.98 when predicted discharges were between 0.9 and 1, and greater than 0.99 for all other predicted discharges.

and associated C statistics with all codes and with those in an analysis restricted to 9 diagnostic and 6 procedure codes differed only trivially from those produced in the model with 25 diagnostic and 25 procedure codes (table 2).

When restricting our analysis to only codes labeled as POA, all models declined in accuracy. For 30-day and 1-yr mortality, the decline averaged only 2%. But predictive accuracy for in-hospital mortality and length of stay each declined about 10%.

## Discussion

Comparative effectiveness research of clinical practices requires risk adjustment to reduce the variation in reported outcomes that may be due to variations in the underlying morbidity of sample populations. Given the broad interest in assessing the performance of healthcare systems, there continues to be an unmet need for a robust risk-adjustment methodology. Of models that attempt to address this issue, the Charlson Comorbidity Index, the Elixhauser Comorbidity Index, the Hierarchical Condition Category,[3] and the American Society of Anesthesiologists physical status score[4] are most commonly used. The RSI, developed and reported in 2010, was found to be robust, apparently broadly applicable, and had the advantage of only requiring administrative data.

Using coefficients derived previously from the 2001–2006 MEDPAR data set, we found that discrimination of RSI in a new national population covering the period from 2007 to 2012 was excellent. In both cases, RSI had high C statistics, ranging from 0.79 to 0.95 for in-hospital, 30-day, and 1-yr mortality, as well as for hospital length of stay. That results were similar in a large national data set covering a completely different time period provides strong validation for the original RSI models. Our validation is also consistent with limited previous analyses showing that RSI was highly predictive for patients at the Cleveland Clinic (Cleveland, Ohio),[9] Massachusetts General Hospital (Boston, Massachusetts),[12] and Duke University (Durham, North Carolina).[13] It is also consistent with a POA code analysis of California inpatients.[14] RSI thus appears to be broadly applicable to a wide variety of patients over the entire adult age range.

Calibration derived from a 20% sample resulted in highly linear correlations between predicted and observed mortality across all deciles in the remaining 80% of the cases. Our calibrated models can thus be used to accurately predict mortality and length of stay across the entire range of risk found in the population.

The 2007–2012 MEDPAR data set allowed up to 25 diagnostic and 25 procedure codes, whereas the older data set was restricted to 9 diagnostic and 6 procedure codes. As might be expected, the newer data set included slightly more diagnostic codes per admission ($n \approx 10$) than the original data set ($n \approx 7$), whereas the number of procedure codes was unchanged. In theory, allowing more codes permits clinicians to identify more baseline illness, which would be expected to enhance discrimination. Interestingly though, the additional codes made little difference with respect to model discrimination. All C statistics were within an absolute 0.1% of the 25-code models, with the exception of the length-of-stay model that improved by an absolute 1%. Presumably,

**Table 3.** Calibration Equations*

| RSI Endpoint | Function Name | Functional Form | Parameter Estimates |
|---|---|---|---|
| In hospital | Logit hazard | $-\log\left[1-\left(1+\exp\left(-\frac{x-\mu}{\beta}\right)\right)^{-1}\right]*c$ | $\mu = 3.74 \beta = 1.211 c = 0.362$ |
| 30 day | Logit hazard | $-\log\left[1-\left(1+\exp\left(-\frac{x-\mu}{\beta}\right)\right)^{-1}\right]*c$ | $\mu = 1.16\ \beta = 0.728\ c = 0.138$ |
| 1 yr | Logit hazard | $-\log\left[1-\left(1+\exp\left(-\frac{x-\mu}{\beta}\right)\right)^{-1}\right]*c$ | $\mu = 1.62\ \beta = 0.806 c = 1.75$ |
| LOS | Extreme value | $d+\left[1-\exp\left[-e^{\frac{x-\mu}{\beta}}\right]\right]*c$ | $\mu = 0.164\ \beta = 0.8393\ c = 0.9654$ $d = 0.0239$ |

*Refer to the technical appendix for more details on use of these equations for calibration purposes.

$\beta$ = SD parameter; $\mu$ = mean parameter; $c$ = scalar parameter; $d$ = offset parameter (special LOS case); LOS = length of stay; RSI = Risk Stratification Index; $x$ = unmodified RSI RiskScore.

discrimination changed little because more important codes are still listed first, and they contribute most to the RSI estimate. Additional diagnostic codes thus contributed relatively little information about patients' baseline risk.

A limitation of using all available codes in claims data is that some diagnoses may be associated with hospital-acquired complications. Hospital-acquired codes may therefore reflect diagnostic intensity rather than underlying risk, making hospital acuity appear worse than it is—essentially giving hospitals credit for diagnostic labeling efforts and for complications. By restricting the codes used in the model to only those labeled as POA, we were able to limit this bias. In-hospital mortality and length-of-stay discrimination were each reduced by about 10% when restricting to POA codes only. This result is unsurprising as one would expect hospital-acquired complications to both increase length of stay and correspond to highly morbid patient populations, including those most likely to die in hospital. In contrast, there was a small decline in discrimination for the 30-day and 1-yr mortality indices when they were restricted to POA codes only, which is consistent with longer term mortality being due to underlying medical issues, especially cancer, rather than hospital-acquired complications.

## Conclusion

Applying RSI model coefficients from the original analysis to our 2007–2012 validation data set yielded C statistics nearly identical to those originally reported. Calibration resulted in highly linear relationships between predicted and observed outcomes over the entire range of risk for the entire population. The recent expansion to 25 codes did not significantly change model performance since discrimination was similar when the model was restricted to 9 diagnostic and 6 procedure codes. POA models were comparably predictive for 30-day and 1-yr mortality but were about 10% less predictive for in-hospital mortality and hospital length of stay, indicating that codes added during an admission unsurprisingly carried information relevant to in-hospital outcomes.

RSI is a stable and robust claims-based risk-adjustment methodology. The models performed well in an entirely new national sample, providing excellent discrimination across various outcomes that are important for health services research including in-hospital, 30-day, 1-yr mortality, and length of stay. The models may also be useful for risk adjustment in retrospective cohort and case–control analyses, as well as for quality metrics. RSI models have the characteristics required for the development and validation of broadly applicable hospital ranking systems.

## Research Support

Supported by the Lown Institute, Boston, Massachusetts.

## Competing Interests

The authors declare no competing interests.

## Correspondence

Address correspondence to Dr. Sessler: Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, 9500 Euclid Ave—P77, Cleveland, Ohio 44195. DS@OR.org. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. ANESTHESIOLOGY's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

## Appendix: Comparison of Exponential Hazard Ratios *versus* Actual Hazard Ratios

Our goal in calibrating our models was to identify simple mathematical functions that, when applied to our risk scores, produced a linear relationship between predicted and observed mortality and length of stay. Calibrating Cox and logistic regressions requires aligning two hazard functions. The first function is a theoretical hazard function, characterized by the form of the estimated regression equations. The second function is characterized by the plot of the risk scores and the observed hazard. The logistic odds ratio regression equation is

$$\frac{Pr(x)}{1-Pr(x)} = \exp(\beta_o + \beta_1 x_1 + \cdots + \beta_n x_n). \quad \text{(A1)}$$

The Cox hazard ratio regression equation is

$$\frac{h_i(t)}{h_o(t)} = \exp(\beta_1 x_{i1} + \cdots + \beta_n x_{in}). \quad \text{(A2)}$$

As we are able to compute the observed mortality rate with respect to each log-odds or log-hazard prediction in our data set, we needed to find a function that would align our regression equations with the actual mortality. First, we divided our Cox and logistic RiskScores (eqns. 3 and 4) into bins using increments of 0.01 and eliminated all bins with fewer than 500 corresponding events. This left us with 858 data bins for 30-day mortality, 818 data bins for 1-yr mortality, 643 data bins for length of stay, and 1,225 data bins for in-hospital mortality. We then transformed the corresponding mortality rates for each bin into an odds ratio using the formula

$$\text{Observed odds} = \frac{\text{Observed mortality}}{1-\text{Observed mortality}} \quad \text{(A3)}$$

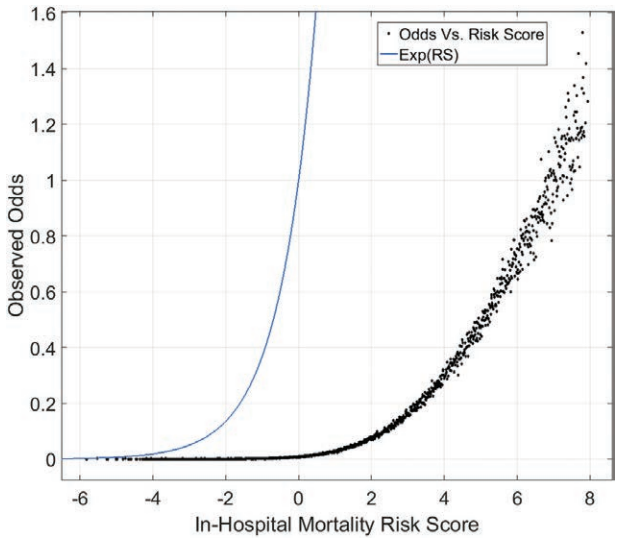


**Fig. A1.** Hazard misalignment.

and plotted observed odds against our regression-estimated log-odds bins.

If the models were initially perfectly calibrated, exponentiation of our RiskScores would yield hazard or odds ratios consistent with our observations. However, it was clear that the graphs were misaligned after plotting the exponential risk scores side by side with the actual odds ratio (see fig. A1).

We empirically evaluated various continuous functions and selected ones that produced high $R^2$ values, indicating linearity of predicted and observed outcomes. Using the transformations below, we fit the logistic distribution to the observed hazard or odds of our target population. We also added a scalar, $c$, to scale our model in proportion to the population odds ratio.

To derive the calibration function, we let $T$ denote survival time. The failure-time probability density function $f_T(t)$ has cumulative density $F_T(t)$, where

$$F_T(t) = Pr(T \le t) = \int_0^t f_T(t)dt \quad \text{(A4)}$$

The probability an individual survives is simply the complement of $F_T(t)$, known as the survivor function, denoted by $S_T(t)$:

$$S_T(t) = Pr(T > t) = 1 - F_T(t). \quad \text{(A5)}$$

The relationship between the survivor function and the hazard function $h_T(t)$ is

$$\begin{aligned} h_T(t) &= \frac{f_T(t)}{S_T(t)} \\ &= \frac{f_T(t)}{1-F_T(t)} \\ &= -\frac{d}{dt}\log[1-F_T(t)] \\ &= -\frac{d}{dt}\log[S_T(t)]. \end{aligned} \quad \text{(A6)}$$

Thus, our cumulative hazard function $H_T(t)$ is

$$H_T(t) = -\log[S_T(t)]. \quad \text{(A7)}$$

The cumulative density of the logistic distribution is given by

$$Pr(T \le t) = \frac{1}{1+e^{-\frac{t-\mu}{\beta}}}. \quad \text{(A8)}$$

Using the cumulative hazard function definition above, substituting RiskScore for $t$ in our failure distribution and changing the lower bound of our integral to $-\infty$ to accommodate the range of the RiskScore, we find the logistic hazard function shown below with the scalar parameter $c$,

$$-\log(1-Pr(T \le t)) * c = -\log\left(1 - \frac{1}{1+e^{-\frac{t-\mu}{\beta}}}\right) * c. \quad \text{(A9)}$$

Estimation of the parameters can be performed in any curve-fitting tool using a nonlinear least-squares model. We used a restricted step method to estimate parameters because the Levenberg–Marquardt algorithm will not converge. Finally, we reconverted the calibrated odds ratios into mortality with the equation

$$\text{Predicted mortality} = \frac{\text{Predicted odds}}{1 + \text{Predicted odds}}. \qquad \text{(A10)}$$

The result was a linear relationship between observed and predicted mortality rates.

## References


1. Charlson ME, Pompei P, Ales KL, MacKenzie CR: A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. J Chronic Dis 1987; 40:373–83
2. Elixhauser A, Steiner C, Harris DR, Coffey RM: Comorbidity measures for use with administrative data. Med Care 1998; 36:8–27
3. Gregory C. Pope M, John Kautter P, Melvin J. Ingber P, Sara Freeman M, Rishi Sekar B, Cordon Newhart M: Evaluation of the CMS-HCC risk adjustment model. Durham, RTI International, 2011
4. Fitz-Henry J: The ASA classification and peri-operative risk. Ann R Coll Surg Engl 2011; 93:185–7
5. van Walraven C, Wong J, Bennett C, Forster AJ: The Procedural Index for Mortality Risk (PIMR): An index calculated using administrative data to quantify the independent influence of procedures on risk of hospital death. BMC Health Serv Res 2011; 11:258
6. Dalton JE, Kurz A, Turan A, Mascha EJ, Sessler DI, Saager L: Development and validation of a risk quantification index for 30-day postoperative mortality and morbidity in noncardiac surgical patients. Anesthesiology 2011; 114:1336–44
7. Le Manach Y, Collins G, Rodseth R, Le Bihan-Benjamin C, Biccard B, Riou B, Devereaux PJ, Landais P: Preoperative Score to Predict Postoperative Mortality (POSPOM): Derivation and validation. Anesthesiology 2016; 124:570–9
8. Glance LG, Lustik SJ, Hannan EL, Osler TM, Mukamel DB, Qian F, Dick AW: The Surgical Mortality Probability Model: Derivation and validation of a simple risk prediction rule for noncardiac surgery. Ann Surg 2012; 255:696–702
9. 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
10. Amari S, Murata N, Muller KR, Finke M, Yang HH: Asymptotic statistical theory of overtraining and cross-validation. IEEE Trans Neural Netw 1997; 8:11
11. Refsgaard JC, Henriksen HJ: Modelling guidelines—Terminology and guiding principles. Adv Water Resour 2004; 27:71–82
12. Sigakis MJ, Bittner EA, Wanderer JP: Validation of a risk stratification index and risk quantification index for predicting patient outcomes: In-hospital mortality, 30-day mortality, 1-year mortality, and length-of-stay. Anesthesiology 2013; 119:525–40
13. Wahl KM, Moretti E, White W, Hale B, Gan T: Validation of a Risk-Stratification Index for Predicting 1-Year Mortality. Durham, Duke University Medical Center, 2011
14. Dalton JE, Glance LG, Mascha EJ, Ehrlinger J, Chamoun N, Sessler DI: Impact of present-on-admission indicators on risk-adjusted hospital mortality measurement. Anesthesiology 2013; 118:1298–306
15. Song Y, Skinner J, Bynum J, Sutherland J, Wennberg JE, Fisher ES: Regional variations in diagnostic practices. N Engl J Med 2010; 363:45–53
16. Department of Health and Human Services, Centers for Medicare and Medicaid Services: Hospital-acquired Conditions (Present on Admission Indicator) Reporting Provision. Available at: https://www.cms.gov. Accessed February 3, 2017
17. Elixhauser A, Steiner C, Palmer L: Clinical Classifications Software (CCS) 2015, Rockville, Maryland, U.S. Agency for Healthcare Research and Quality, 2015