Expert Consensus vs Empirical Estimation of Injury Severity

Effect on Quality Measurement in Trauma

Laurent G. Glance, MD; Turner M. Osler, MD; Dana B. Mukamel, PhD; Wayne Meredith, MD; Andrew W. Dick, PhD

Objective: To determine the extent to which the Injury Severity Score (ISS) and Trauma Mortality Probability Model (T-MPM), a new trauma injury score based on empirical injury severity estimates, agree on hospital quality.

Design, Setting, and Patients: Retrospective cohort study based on 66,214 patients in 68 hospitals. Four risk-adjustment models based on either ISS or T-MPM were constructed, with or without physiologic information.

Main Outcome Measures: Hospital quality was measured using the ratio of the observed-to-expected mortality rates. Pairwise comparisons of hospital quality based on ISS vs T-MPM were performed using the intraclass correlation coefficient and the $\kappa$ statistic.

Results: There was almost perfect agreement for the ratios of the observed to expected mortality rates based on the T-MPM vs the ISS when physiologic information was included in the model (intraclass correlation coefficient, 0.93). There was substantial agreement on which hospitals were identified as high-, intermediate-, and low-quality hospitals ($\kappa = 0.79$). Excluding physiologic information decreased the level of agreement between the T-MPM and the ISS (intraclass correlation coefficient, 0.88 and $\kappa = 0.58$).

Conclusions: The choice of expert-based or empirical Abbreviated Injury Score severity scores for individual injuries does not seem to have a significant effect on hospital quality measurement when physiologic information is included in the prediction model. This finding should help to convince all stakeholders that the quality of trauma care can be accurately measured and has face validity.

Arch Surg. 2009;144(4):326-332
izes all patient injuries into a single anatomic severity score based on the AIS. The ISS is then calculated by taking the sum of squares of the AIS of the 3 body regions with the highest AIS. The most important limitation of the ISS is that the AIS for each of the 1400 injuries is based on expert opinion rather than on actual outcomes of patients with these injuries.

In a manner analogous to AIS codes, ICD-9-CM codes can be used to describe patient injury. For each ICD-9-CM code, the proportion of survivors is used to estimate the survival risk ratio for each injury. Unlike AIS, survival risk ratios are calculated using actual data. The main criticism of survival risk ratios is that many patients sustain more than 1 injury. Therefore, survival risk ratios for individual ICD-9-CM codes are contaminated with information from other injuries. Overall injury severity is based on the product of a patient’s survival risk ratios. Most trauma surgeons believe that the AIS lexicon more accurately describes patient injuries because the AIS lexicon was designed specifically for traumatic injuries, whereas ICD-9-CM codes were designed to be used for billing purposes.

The Agency for Healthcare Research and Quality has recently funded our investigations, in collaboration with the American College of Surgeons Committee on Trauma, to assess the effect of nonpublic report cards on trauma outcomes by randomizing hospitals to receive or not to receive feedback on their risk-adjusted mortality rates for trauma patients. To that end, we have developed the Trauma Mortality Probability Model (T-MPM). This model is based on AIS coding. However, unlike the ISS and other AIS-based models, AIS in the T-MPM is empirically estimated using regression modeling, as opposed to being based on expert consensus. It has previously been shown that the T-MPM has superior discrimination and is better calibrated than the ISS.4

Before adopting a new injury model, it is important to consider whether a new model (ie, the T-MPM) leads to differences in hospital ranking. Virtually every study examining the effect of risk adjustment on quality measurement has shown that hospital quality ranking depends on the choice of the risk-adjustment model.5 Because there is no gold standard that we can use to measure trauma center quality, it is impossible to know which of these 2 risk-adjustment models (the ISS vs the T-MPM) most accurately measures quality. A priori, because the T-MPM exhibits better model fit than the ISS,4 we believe that hospital quality measures based on the T-MPM should be a less biased measure of true quality. In this study, our goal is to determine the extent to which the ISS and the T-MPM agree on hospital quality. Although we expect that these 2 scoring systems will not agree on the quality ratings for many of the hospitals, the finding that hospital quality does not depend substantially on the choice of the injury model would provide strong support for the validity of hospital quality measurement in trauma.

**METHODS**

**PATIENT POPULATION**

This analysis was conducted using data from the National Trauma Data Bank (NTDB). The American College of Surgeons created the NTDB to serve as the “principal national repository for trauma center registry data.”6 Data elements in the NTDB include patient demographics, hospital demographics, AIS codes, mechanism of injury (based on ICD-9-CM codes), encrypted hospital identifiers, physiologic values, and outcomes. This analysis was based on patients admitted in 2005. Only hospitals that had annual patient volumes of at least 250 in 2005 and assigned valid AIS codes to at least 95% of trauma patients were included in this study. Hospitals missing the Glasgow Coma Scale (GCS) motor component on more than 20% of the patients were excluded. Patients younger than 1 year, with nontraumatic diagnoses or missing information on age or sex, or who were dead on arrival or who were transferred to another facility were excluded from the analysis. The final data set included 66,214 patients in 68 hospitals.

**TRAUMA MORTALITY PREDICTION MODEL**

The development and validation of the T-MPM have been described by Osler et al.4 Because of the many possible injuries (1407 possible individual AIS codes) and the fact that more than 50% of the injuries occurred fewer than 100 times in the entire NTDB (and 14% occurred <10 times), a simple prediction model in which each injury is coded as a binary predictor leads to imprecise coefficient estimates for many injuries. The current standard, the ISS, collapses AIS-based injury severities for the 3 most severe injuries into a single scalar measure of injury severity.7 The ISS is based on individual AIS injury severities assigned by a panel of experts.

In contrast, the T-MPM is based on empirical estimates of injury severity derived using the NTDB. Empirical estimates of injury severity for each AIS code were initially estimated using probit regression in which each injury is coded as a binary predictor. Because some of the injuries are sparsely populated, a second regression model collapsed the 1407 possible AIS codes into 49 regional severity codes (based on the 9 AIS body regions and the 6 expert-based possible AIS injury severities [not every body region had injuries in all 6 severity strata]). The AIS injury severities were then calculated by taking a weighted sum of the coefficients estimated using these 2 separate models. The final model, the T-MPM, uses the 5 most severe injuries (coded using empirical estimates of injury severity) as predictors of mortality.

**AUGMENTED INJURY MODELS**

We constructed the following 4 probit regression models to predict inhospital mortality: (1) the T-MPM augmented with age, sex, and mechanism of injury; (2) the T-MPM augmented with age, sex, mechanism of injury, and the motor component of the GCS; (3) the ISS augmented with age, sex, and mechanism of injury; and (4) the ISS augmented with age, sex, mechanism of injury, and the motor component of the GCS.

A priori, we limited the physiologic information in our model to the motor component of the GCS because the motor component contributes most of the information in the GCS3 and is less sensitive to initial therapy than the verbal and eye components of the GCS. We have assumed that all patients can be assigned a GCS motor component; patients who are pharmacologically paralyzed are assumed to have been assessed before the administration of neuromuscular blockers. Blood pressure and respiratory rate were not included as predictor variables because it is impossible to distinguish a respiratory rate or blood pressure of zero from a missing value in the NTDB. The decision to exclude respiratory rate from the prediction model was also made to avoid excluding intubated patients from the risk-adjustment model. Multiple imputation was used to impute missing values of the motor component of the GCS using the STATA (StataCorp LP, College-

©2009 American Medical Association. All rights reserved.
estimators13 were used because the outcomes of patients treated to impute missing data and results in hospital quality measures that are almost identical to those based on a data set without missing values.13 The method of fractional polynomials was used to determine the optimal transformation for age.12 Robust variance estimators13 were used because the outcomes of patients treated at the same hospital may be correlated. Model discrimination was evaluated using the C statistic, and calibration was evaluated using the Hosmer-Lemeshow statistic.14

IDENTIFICATION OF HOSPITAL QUALITY OUTLIERS

The expected mortality rate for each hospital was calculated using each of the 4 models. Hospital quality was quantified using the ratio of the observed to expected mortality rates (O/E ratio). Hospitals whose O/E ratio was significantly less than 1 were classified as high-quality outliers, whereas hospitals whose O/E ratio was significantly greater than 1 were classified as low-quality outliers. The 95% confidence interval around the O/E ratio was constructed using the normal approximation of the binomial distribution.15

STATISTICAL ANALYSIS

We performed 2 different hospital-level analyses to examine the level of agreement between the risk-adjusted measures of hospital quality based on the augmented T-MPM vs the augmented ISS. We assessed the level of agreement for (1) O/E ratios using the intraclass correlation coefficient and (2) categorical measures of hospital quality using the \( \kappa \) statistic. All statistical analyses were performed using commercially available software (STATA SE/MP version 10.0, StataCorp LP).

RESULTS

DATABASE DESCRIPTION

The analysis was based on 66,214 patients in 68 hospitals. Sixty-five percent of the patients were male, and the median patient age was 37 years. Forty-three percent of the patients sustained blunt trauma, 29% were in motor vehicle crashes, 11% had low falls, and the remainder of trauma injuries were caused by gunshot wounds, pedestrian accidents (ie, automobile and vehicle crashes involving a pedestrian), or stab wounds. The cohort mortality rate was 4.22%. Approximately 5% of the patients were missing the motor component of the GCS score. The hospital rate of missing data for the motor component of the GCS varied between 0% and 20%.

MODEL PERFORMANCE

Each of the models exhibited excellent discrimination (Table 1). For the sake of comparison, we have also included the C statistic and the Hosmer-Lemeshow statistic for the injury models without any additional risk factors. The calibration curves are shown in Figure 1.

DATA

Figure 2 shows the level of agreement for the O/E ratios based on the augmented T-MPM vs the O/E ratios based on the augmented ISS. The intraclass correlation coefficient of 0.93 indicates almost perfect agreement between the ISS and the T-MPM when the GCS motor component was included in the model. Excluding the GCS motor component as a predictor reduced the intraclass correlation coefficient to 0.88.

There was substantial agreement between the augmented T-MPM and the augmented ISS on which hospitals were identified as high, intermediate, and low quality when the GCS motor component was included in the prediction models (\( \kappa = 0.79 \) (Table 2). Omitting the GCS motor component in the prediction models caused the \( \kappa \) statistic to decrease to 0.38, indicating moderate agreement between the augmented T-MPM and the augmented ISS (Table 3).

The caterpillar graph, which shows the O/E ratios for each of the hospitals as a function of the T-MPM vs the ISS, also illustrates the high level of agreement between these 2 different scoring systems on hospital quality assessment (results for injury model augmented with age, sex, injury mechanism, and GCS motor component are shown). These results are shown in Figure 3.

COMMENT

With the publication of the seminal Institute of Medicine report Crossing the Quality Chasm,16 there is increasing recognition that medical care is neither always safe nor always effective. Bridging the “quality chasm” requires the development of robust outcome measures to guide quality improvement by identifying and then learning from high-quality hospitals.17 Long before quality measurement became the mantra of health care reform, trauma surgeons championed the development of trauma outcome registries and conducted research on injury scoring. After more than 20 years of research, no injury scoring system has yet replaced the venerable ISS. We have proposed a new injury scoring system, the T-MPM, which replaces the consensus-based AIS used in the ISS with empirical AIS. When injury alone is used to predict mortality, it has been shown that the T-MPM significantly outperforms the ISS.4 However, the addition of demographic variables, mechanism of injury, and physiologic informa-

### Table 1. Model Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>C Statistic</th>
<th>Hosmer-Lemeshow Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISS</td>
<td>0.968</td>
<td>215.5</td>
</tr>
<tr>
<td>ISS with age, sex, injury mechanism</td>
<td>0.905</td>
<td>47.0</td>
</tr>
<tr>
<td>ISS with age, sex, injury mechanism, GCS motor component</td>
<td>0.956</td>
<td>15.7</td>
</tr>
<tr>
<td>T-MPM</td>
<td>0.990</td>
<td>59.6</td>
</tr>
<tr>
<td>T-MPM with age, sex, injury mechanism</td>
<td>0.927</td>
<td>14.9</td>
</tr>
<tr>
<td>T-MPM with age, sex, injury mechanism, GCS motor component</td>
<td>0.962</td>
<td>25.2</td>
</tr>
</tbody>
</table>

Abbreviations: GCS, Glasgow Coma Scale; ISS, Injury Severity Score; T-MPM, Trauma Mortality Probability Model.
tion causes both models to exhibit similar levels of statistical performance. In the present study, we compared hospital quality based on the augmented ISS vs the augmented T-MPM and found that they exhibited almost perfect agreement on the quality of trauma centers when physiologic information is included in the model. This finding is especially surprising given that the ISS and the T-MPM are based on 2 different approaches to injury severity modeling.

Our study has important implications for the measurement of trauma center quality. In the recent Institute of Medicine report Performance Measurement: Accelerating Improvement, Birkmeyer argues that, although hospital “performance measures will never be perfect,”18(p192) health care leaders will need to “make decisions about when imperfect measures are good enough to act upon.”18(p192) In her seminal article on risk adjustment, Iezonni5 showed that different risk-adjustment models often led to different conclusions regarding hospital quality and that the main risk of risk adjustment is that different quality yardsticks will not always agree on hospital quality. How good does risk adjustment need to be before we can evaluate hospital qual-

Figure 1. Calibration curves for the Trauma Mortality Probability Model (T-MPM) and the Injury Severity Score (ISS) as a function of additional risk predictors. A, Injury model without additional risk factors. B, Injury model with age, sex, and injury mechanism. C, Injury model with age, sex, injury mechanism, and Glasgow Coma Scale (GCS) motor component. The 95% confidence intervals are based on the binomial distribution.
ity using risk-adjusted outcomes? We propose to set the bar for trauma to a level where hospital quality becomes insensitive to the choice of risk-adjustment model. In the case of trauma, we believe that our findings that the augmented ISS and the augmented T-MPM agree on hospital quality suggest that injury scoring based on empirical injury severities, in combination with patient demographic and physiologic information, is sufficiently robust to serve as the basis for evaluating trauma center quality.

Previous comparisons of trauma scoring systems have focused on the statistical performance of competing trauma scoring systems, as opposed to assessing the effect of the choice of risk adjustment on hospital quality assessment. To our knowledge, an earlier study examining whether the Trauma Injury Severity Score methods and a severity characterization of trauma ranked hospitals differently is the only study of its type in the trauma literature. In that study using the original model coefficients for each of these prediction models, a substantial level of disagreement was found between the Trauma Injury Severity Score and a severity characterization on the identity of high- and low-quality trauma centers. The fact that both models were poorly calibrated in the hospital cohort used in that study may have been an important factor in these findings.

The most feasible explanation for our results in the present study is that the ISS and the T-MPM have almost equivalent model fit when age, sex, mechanism of injury, and physiologic information are added to the injury models. The addition of the motor component of the GCS markedly improved the appearance of the calibration curve for the ISS to the extent that the calibration curves for the augmented T-MPM and the augmented ISS were virtually indistinguishable. Coupled with the almost perfect discrimination of both models, it is not unreasonable for both prediction models to exhibit such a high level of agreement on hospital quality assessment. It is likely that the motor component of the GCS is a sufficiently powerful predictor that adding it to the final models blurred the previously significant differences in model performance between the T-MPM and the ISS.

Our finding that excluding physiologic information in the prediction model reduces the extent of agreement between the T-MPM and the ISS on hospital quality also has important implications for benchmarking. When physiologic information is excluded in the prediction model, the discrimination and calibration of the T-MPM are significantly better than those of the ISS. We strongly discourage the use of the ISS for benchmarking hospital performance when physiologic information is unavailable because without this information the T-MPM significantly outperforms the ISS.

This study has several potentially significant limitations. First, the NTDB is not population based but instead represents a convenience sample of self-selected hospitals. Furthermore, our study is based on a small cohort of hospitals within the NTDB that coded more than 95% of their patients using AIS codes and were missing GCS.

Table 2. Hospital Quality Outliers in the Injury Models Augmented With Age, Sex, Injury Mechanism, and Glasgow Coma Scale Motor Component

<table>
<thead>
<tr>
<th>Augmented ISS</th>
<th>Augmented T-MPM Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>High</td>
</tr>
<tr>
<td>High</td>
<td>4</td>
</tr>
<tr>
<td>Intermediate</td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviations: ISS, Injury Severity Score; T-MPM, Trauma Mortality Probability Model.

Table 3. Hospital Quality Outliers in the Injury Models Augmented With Age, Sex, and Injury Mechanism

<table>
<thead>
<tr>
<th>Augmented ISS</th>
<th>Augmented T-MPM Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>High</td>
</tr>
<tr>
<td>High</td>
<td>6</td>
</tr>
<tr>
<td>Intermediate</td>
<td>4</td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviations: ISS, Injury Severity Score; T-MPM, Trauma Mortality Probability Model.
data on less than 5% of their patients. Therefore, our study is not necessarily generalizable outside of the study data set and needs to be replicated. Second, we used multiple imputation to impute missing GCS data. Multiple imputation assumes that patients have missing data conditional on their observed risk factors. If the missing data also depend on unmeasured risk factors, then multiple imputation may be biased. However, this assumption (whether the missing data are also a function of unobserved factors) is not testable, and multiple imputation may still be preferable to other approaches designed for situations where the assumption underlying multiple imputation is not entirely valid.26,27 Third, we were unable to include other potentially important predictors such as blood pressure and comorbidities because of coding issues and missing data in the NTDB.

In conclusion, trauma center quality can be assessed using injury scoring augmented with demographic information, mechanism of injury, and physiologic information. The choice of expert-based or empirical AIS for individual injuries does not seem to have a significant effect on hospital quality measurement when physiologic information is included in the prediction model. This finding should help convince all stakeholders that the quality of trauma care can be accurately measured and has face validity.

Accepted for Publication: March 15, 2008.

Correspondence: Laurent G. Glance, MD, Department of Anesthesiology, University of Rochester School of Medicine, 601 Elmwood Ave, Box 604, Rochester, NY 14642 (Laurent_Glance@urmc.rochester.edu).

Figure 3. Hospital quality as a function of the augmented Trauma Mortality Probability Model (T-MPM) vs the augmented Injury Severity Score (ISS) (augmented with age, sex, injury mechanism, and Glasgow Coma Scale motor component). Vertical bars represent 95% confidence interval around the point estimate of the hospital ratio of the observed to expected mortality rates (O/E ratio).

Author Contributions: Study concept and design: Glance, Osler, Mukamel, Meredith, and Dick. Analysis and interpretation of data: Glance, Osler, and Mukamel. Drafting of the manuscript: Glance and Mukamel. Critical revision of the manuscript for important intellectual content: Glance, Osler, Mukamel, Meredith, and Dick. Statistical analysis: Glance, Osler, Mukamel, and Dick. Obtained funding: Glance and Meredith. Administrative, technical, and material support: Glance.

Financial Disclosure: None reported.

Funding/Support: This study was supported by grant R01 HS 16737 from the Agency for Healthcare and Quality Research.

Disclaimer: The views presented in this article are those of the authors and may not reflect those of the Agency for Healthcare and Quality Research or of the American College of Surgeons Committee on Trauma.

REFERENCES

Glanz et al have produced an important article that outlines a new and useful trauma quality method, the T-PMPI, that compares favorably with the ISS in terms of predicting mortality. The T-PMPI is based on the ICD-9-CM-9 diagnosis codes, which are used for the hospital administrative data set as part of the hospital billing process. Conversely, the ISS, currently used by many trauma centers, is derived from the AIS, requiring a more labor-intensive coding and trauma database process. The authors further demonstrated that the ISS and T-PMPI could be improved (augmented ISS and T-PMPI) by inclusion of 3 patient-specific factors to the calculation (age, sex, and the motor component of the GCS), resulting in almost identical conclusions regarding trauma mortality.

Why are these important findings? America spent $72.5 billion in 2005 on trauma care, ranking second only to expenditures for heart conditions. Pay for performance and value-based purchasing programs now exist as keystones of many federal and commercial insurance health care programs to reduce costs and to promote quality. These programs will surely expand to grade trauma outcomes at lower incremental costs.

However, there is more work to be done. The ISS and the T-PMPI are mortality measures, and nationally only 4.4% of trauma patients reaching a hospital die. Urgently needed for trauma quality improvement is a mechanism that will predict morbidity and resource consumption (cost) for most patients surviving trauma. Because trauma survivors account for the bulk of health care costs, having a mechanism to predict resource consumption for a given level of trauma will be essential for cost reduction of trauma care, while improving quality. This study provides a great platform to begin that search for a morbidity-based outcomes cost metric.

Charles D. Mabry, MD

Correspondence: Dr Mabry, Department of Surgery, University of Arkansas for Medical Sciences, 1801 W 40th, Ste 7B, Pine Bluff, AR 71603 (mabrycharlesd@uams.edu).

Financial Disclosure: None reported.