Published in final edited form as:

J Am Coll Surg. 2010 August ; 211(2): 224–231. doi:10.1016/j.jamcollsurg.2010.03.033.

Predicting Risk-Adjusted Mortality for Trauma Patients: Logistic versus Multilevel Logistic Models

David E Clark, MD, FACS^{(1),(2)}, Edward L. Hannan, PhD⁽³⁾, and Chuntao Wu, MD, PhD^{(3),(4)}

¹⁾Maine Medical Center, Portland ME

²⁾Harvard Injury Control Research Center, Boston MA

³⁾State University of New York, Albany NY

⁴⁾Penn State Hershey College of Medicine, Hershey PA

Abstract

Background—Theoretical advantages of random-intercept multilevel (ML) logistic regression (LR) modeling over standard LR include separating variability due to patient-level and hospital-level predictors, “shrinkage” of estimates for lower-volume hospitals toward the overall mean, and fewer hospitals falsely identified as outliers.

Study Design—We used Nationwide Inpatient Sample (NIS) data from 2002–2004 to construct LR models of hospital mortality after admission with a principal ICD-9-CM injury diagnosis (ICD-9-CM 800–904, 910–929, 940–957, 959). After considering various predictors, we used patient-level indicator variables for age groups, sex, maximum AIS for the head region (3,4,5), maximum AIS for other body regions (3,4,5), and mechanisms (fall, gunshot, motor vehicle). Using standard LR and MLLR, we compared predictions based upon 2002, 2003, and 2004 data to actual mortality observed in the same hospitals in 2004, 2005, and 2006 NIS respectively.

Results—Patient-level fixed effects were similar for the two methods in all years, with mortality associated most strongly with AIS=5 head injury, other AIS=5 injury, or higher age groups. ML models identified fewer hospitals as outliers. Differences between actual and predicted mortality were significantly smaller with MLLR models compared to standard LR models.

Conclusions—ML models may have advantages for the measurement and explanation of interhospital differences in trauma patient outcomes.

INTRODUCTION

Statistical models for predicting hospital performance are increasingly of interest to health planners, regulators, and patients.⁽¹⁾ Multilevel (ML) models (also called “hierarchical models”) have been widely accepted for decades among statisticians,^(2–3) and are increasingly popular in health services research.^(4–9) Theoretical advantages of ML models

© 2010 American College of Surgeons. Published by Elsevier Inc. All rights reserved.

Correspondence address: David E. Clark, MD, 887 Congress Street, Suite 210, Portland ME 04102, Phone 207-774-2381, Fax 207-774-0459, clarkd@mmc.org.

Presented in part at the International Surgical Week, Adelaide, Australia, September 2009, and at the American Public Health Association, Philadelphia, PA, November 2009.

Disclosure Information: Nothing to disclose.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

include: 1) reflecting the natural data structures more correctly than alternative models, 2) accounting for variance structures correctly (such as correlations between patients within hospitals, i.e. “clustering”), 3) acknowledging that some ‘treatment effects’ (such as the hospitals in the model) may be only a sample of effects from the population of interest (all similar hospitals), and 4) producing statistically optimal “shrunk” predictors that account for different amounts of information available on different ‘treatments’.

The major drawback to the use of ML modeling has been computational complexity, which is now being overcome as a result of technological and theoretical progress. However, the ML approach remains unfamiliar to many clinicians and it may be argued whether its theoretical advantages justify the additional complexity. (10-11) This study was therefore conducted to evaluate whether predictions of trauma patient outcomes using ML modeling of administrative data were more accurate than predictions using standard logistic regression models. The process of collecting and preparing comparable and reliable data may take several years, so it seemed appropriate to judge the accuracy of these statistical models by their ability to predict outcomes two years later.

METHODS

Data management

Nationwide Inpatient Sample (NIS) data for 2002-2006 were obtained from the Healthcare Cost and Utilization Project (HCUP) of the Agency for Healthcare Research and Quality (AHRQ), in compliance with its Data Use Agreements. The proposed study was also reviewed and approved by an Institutional Review Board at Maine Medical Center. Stata (Version 10, StataCorp, College Station TX) and SAS (Version 9.1, Cary NC) were used for data management and statistical modeling, and MLwiN (Version 2.02, Bristol UK) was also used for multilevel modeling.

Relevant variables were extracted from the 2002 NIS for patients admitted with a principal International Classification of Diseases, Ninth Revision (ICD-9) diagnosis of injury (Codes 800-959, excluding 905-909, 930-939, or 958).

Hospitals were considered ineligible for this study if they had admitted fewer than 25 cases of injury as defined above, had transferred more than 10% of their cases to another acute care hospital, or were missing discharge vital status (dead or alive at the time of discharge) in more than 1% of their cases. From the 2004 NIS we identified a “comparable” subgroup of hospitals that met the above eligibility criteria in 2004 as well as in 2002. Data from these comparable hospitals in the 2002 NIS constituted the “prediction sample”, and data from these comparable hospitals in the 2004 NIS constituted the “validation sample”.

Similarly, a “prediction sample” was obtained from the 2003 NIS, with “comparable” hospitals identified in a “validation sample” obtained from the 2005 NIS. Likewise, a “prediction sample” was obtained from the 2004 NIS, with “comparable” hospitals identified in a “validation sample” obtained from the 2006 NIS. Because of changes in the sampled hospitals each year, the 2004 “validation sample” for which predictions were made using 2002 data was not the same as the 2004 “prediction sample” used to make predictions for 2006.

Individual ICD-9 Clinical Modification (ICD-9-CM) injury diagnoses were each mapped to a body region (head, face, chest, abdomen, extremities, general) and Abbreviated Injury Score (AIS)(12) using an algorithm described elsewhere.(13) AIS was not assigned for burn diagnoses. Cases were excluded from analysis if no injury diagnosis could be mapped to a body region and AIS severity. Cases were also excluded if the Admission Type (ATYPE)

variable indicated an elective or neonatal admission, if the Uniform Disposition (DISPUNIFORM) variable indicated transfer to another acute-care hospital, or if discharge vital status was missing. ICD-9-CM External Cause of Injury Codes (E-Codes) were classified into general injury mechanisms as recommended by the Centers for Disease Control and Prevention.(14)

Because of the small number of cases individually excluded due to missing data, we did not attempt to impute outcome data for cases without a valid AIS or those resulting in transfer to another acute care hospital. Because we were not attempting to estimate population-based statistics, we did not weight the data by hospital; for various other theoretical reasons, weights are generally not advised for multilevel modeling of NIS data.(15)

Preliminary data exploration suggested that categorizing cases by their maximal AIS score (AISmax) in the head region and their AISmax in any other body region would be effective. (16-17) For this purpose, the few cases with AISmax of 6 were grouped with those having AISmax of 5.

Statistical modeling

Logistic regression (LR) models predicting hospital mortality for cases of blunt trauma were created using these measures of injury severity along with indicator variables to designate age groups, sex, mechanism of injury (motor vehicle versus other blunt), and whether the patient had been transferred from another hospital. In order to improve numerical stability, and allow for more descriptive interpretation, the values of all covariates were entered into the models as their deviations from the overall mean value for the corresponding sample (grand-mean centering).(2) For each patient i in hospital j , LR models estimate the log odds (logit) of the probability of hospital death,

$$y_{ij} = \beta_0 + \sum_{k=1}^K \beta_k (X_{ijk} - X_{..k}) \quad (\text{Equation 1})$$

where β_0 is the intercept; β_k is the regression coefficient associated with covariate k , X_{ijk} is the value of covariate X_k for the i th patient in hospital j ; $X_{..k}$ is the overall mean value for covariate X_k ; and K is the number of covariates. The probability of hospital death (p_{ij}) can then be obtained using the antilogit function, that is

$$p_{ij} = \frac{1}{1 + \exp(-y_{ij})} \quad (\text{Equation 2})$$

For the standard LR model, the observed mortality (OM) for hospital j can be compared to the expected mortality (EM) calculated as the sum of p_{ij} for all n_j cases in that hospital. A confidence interval (CI) can be constructed using the variance of this sum, namely

$$v_{1j} = \sum_{i=1}^{n_j} p_{ij} (1 - p_{ij}) \quad (\text{Equation 3})$$

as was done in the Major Trauma Outcome Study,(18) or approximating this variance by a Poisson distribution,(19) as was done in this study. Another formulation of standard LR would create an indicator variable X_j (equal to 1 for patients in hospital j and 0 otherwise) and substitute

$$\beta_0 = \beta_{00} + \beta_j X_j \quad (\text{Equation 4})$$

into Equation 1; if the reference database is very large compared to any individual hospital, the intercept β_{00} for patients other than those in hospital j will be little different from β_0 , and the fixed coefficients β_k will also be little different, but the coefficient β_j and its confidence interval can be used as a measure of effect for hospital j .(6, 20) After applying any of these methods, hospitals with effects significantly different from the reference group may be defined as outliers.

Using the same variables described above, random-intercept multilevel (ML) LR models were created using MLwiN (and replicated using the “xtmelogit” command in Stata). The first level of this model is similar to standard LR, estimating the probability of hospital death as

$$y_{ij} = \beta_{0j} + \sum_{k=1}^K \beta_k (X_{ijk} - X_{..k}), \quad (\text{Equation 5})$$

but in this case, the variable β_{0j} is determined by a second level of the model

$$\beta_{0j} = \gamma_{00} + u_{0j}, \quad (\text{Equation 6})$$

where β_{00} is the mean intercept across hospitals and u_{0j} is the random effect of hospital j on the mean, assumed to be normally distributed with mean of zero and a variance v_2 estimated from the data. This combination of Equation 6 with Equation 5 is thus similar to the combination of Equation 4 with Equation 1 using standard LR. For the random-intercept ML model, the first-level variance is estimated the same as in Equation 3, but the ratio of v_{1j} to v_2 is used to derive a “reliability factor” λ_j , tending toward 0 when n_j is small and tending toward 1 when n_j is large.(2,21)

Unlike standard LR, the ML approach then evaluates hospitals according to their shrunken mortality (SM), namely

$$SM = \lambda_j OM + (1 - \lambda_j) EM.$$

This calculation produces a weighted estimate “shrinking” the effect of each hospital toward the expected group mean, most notably when n_j is small and therefore λ_j is small.(2,3) The CI is shrunken less than the mortality estimates, so there should theoretically be fewer outliers, especially among smaller hospitals.

Outcome comparisons

Our main purpose was to investigate the ability of the models constructed from Year N (2002, 2003, and 2004, respectively) prediction samples to predict observed mortality in the validation samples from Year N+2 (2004, 2005, and 2006, respectively). Briefly, the predictive performances at the hospital level for the logistic regression model and the ML logistic regression model were evaluated using three measures: The first measure was the mean absolute difference between the observed and expected mortality rates in Year N+2; the second was the root mean square error (RMSE) of the differences between observed and expected mortality rates in Year N+2; and the third was the percentage of hospitals whose expected numbers of deaths in Year N+2 were within 95% Poisson confidence intervals based on the observed numbers of deaths in the same year.

The following steps were followed to obtain each hospital's expected mortality rate in Year N+2 using logistic regression models: First, each patient's expected probability of death in Year N was calculated using the parameters of the model and the patient's risk factors in the same year; then each hospital's expected mortality rate (EMR) in Year N was computed as the average expected probability of death for all patients in that hospital. Each hospital's observed mortality rate (OMR) in Year N was calculated by dividing the number of observed deaths by the number of patients admitted to that hospital in Year N. Next, we calculated the ratio of OMR to EMR for each hospital in Year N, which will be referred to as the O/E value or relative performance measure.

The logistic regression model and hospital-specific O/E value in Year N were then used to calculate the expected probability of death for each patient in Year N+2, based on each patient's risk factors. The expected probability of death for a patient in a hospital in Year N+2 was calculated as the product of predicted probability of death in year N+2 using the Year N model and O/E for the hospital in Year N. The EMR for each hospital in Year N+2 was then calculated as the average expected probability of death for all patients in that hospital in that year. OMR and the O/E value were calculated for Year N+2 as they had been for Year N.

For ML logistic regression models, the coefficients of the models along with hospital-specific residuals for the intercept in Year N were similarly used to calculate each patient's expected probability of death in Year N+2, based upon each patient's risk factors. The EMR and O/E for hospitals were then computed as described above. More detailed description of the methods for comparing predictive performance of ML and standard logistic regression models can be found in Hannan et al.(10)

In the prediction samples, hospitals were considered outliers if their predicted mortality rates were not within the 95% confidence intervals based on their observed mortality rates for standard LR, or if their shrunken residuals of intercepts in MLLR were significantly different from zero ($p < 0.05$). We also constructed standard and ML models from the validation samples, using the same variables that had been used in the corresponding prediction samples. For each year and modeling method, we determined whether each hospital was a performance outlier using the same methods applied to the prediction samples for each type of statistical model. The outlier status of each hospital in each validation sample was compared to its status in the corresponding prediction sample from two years previously. Analogous to a diagnostic test, we measured "sensitivity" as the proportion of actual outliers in Year N+2 that would have been predicted by the same method using the data from Year N.

RESULTS

From the 2002 NIS, we identified 306,303 cases admitted to 989 hospitals with a principal diagnosis of injury. Of the 989 hospitals, 346 were considered ineligible because their 2002 injury case volume was less than 25, their proportion transferred to another acute-care hospital was greater than 10%, or they lacked discharge vital status for 1% or more of their cases. All 28,315 cases from these 346 ineligible hospitals were therefore excluded. Among the 277,988 cases in the 643 eligible hospitals from 2002 (Table 1), 1.9% were excluded because they had been transferred to another acute-care hospital, and 1.7% because they could not be assigned an AIS severity. 172 of the 643 hospitals eligible in the 2002 NIS also met similar criteria in the 2004 NIS and constituted the "comparable" group for this pair of years.

Similar distributions and exclusions were seen for the 2003 and 2004 prediction samples and their corresponding 2005 and 2006 validation samples. There were 148 comparable hospitals in the 2003-2005 samples and 171 in the 2004-2006 samples. The ranges of hospital volume and mortality were similar in the prediction and validation samples (Table 2).

Models constructed from the 2002, 2003, and 2004 prediction samples demonstrated similar fixed effects for patient-level covariates, regardless of the modeling method (Table 3). Increased hospital mortality was associated with progressively older decades of age after 45, and with progressively higher AIS, especially with head injury. Age under 15 and motor vehicle mechanism were significant in the model constructed from 2002 data, but not in subsequent years. Transfer status, mechanisms of injury, and age groups other than those specified were not significant when entered as patient-level covariates. Estimates of the fixed components of the random-intercept ML models were very similar to the estimates from standard logistic regression (Table 3).

As described above, the predictive abilities of the models were compared in three ways for each pair of years: The average absolute difference between the observed mortality and the predicted mortality; the root mean square error between observed and predicted mortality; and the percentage of hospitals whose observed mortality was within a predicted 95% confidence interval. For each comparison, and for each pair of years, the ML approach was superior (Table 4). The robustness of these findings was investigated by eliminating those hospitals with no recorded mortality in either the prediction sample or the validation sample, and the ML approach remained superior (Table 4).

Table 5 depicts all possible combinations of outliers identified by one or both modeling methods in the 2002-04 prediction (Year N) samples and/or the 2004-06 (Year N+2) validation samples. Standard LR models, using observed mortality, identified as outliers 39 hospitals in one or more of the prediction samples, and 30 hospitals in one or more of the validation samples. Multilevel LR models, using shrunken mortality, identified as outliers only 31 hospitals in the prediction samples and 18 hospitals in the validation samples. ML regression identified more high outliers (mortality greater than expected), while standard regression identified many more low outliers (mortality less than expected). Of the 18 outliers identified in the validation samples by MLLR, 14 had been identified in the corresponding prediction samples for a sensitivity of 78%. Of the 30 outliers identified in the validation samples by standard LR, 11 had been identified in the corresponding prediction samples for a sensitivity of 37%.

DISCUSSION

The voluntary participation of trauma centers in collection and submission of registry data is motivated by a desire to define a measurable standard of care, as well as the desire of individual centers to demonstrate their achievement of the standard (within the bounds of inevitable random variation). The assumption is that properly adjusted institutional performance in the recent past can be used to predict institutional performance in the immediate future. These predictions can be compared with a “benchmark”, most often the similarly predicted average performance of all hospitals in the same class.

Once a measure of relative hospital performance has been developed, it may be used (or abused) for competitive or regulatory purposes, and even publicized in a “scorecard” or “report card”.⁽²²⁾ To produce the desired goal of quality improvement, the measure must therefore be perceived by the professional community as clinically and statistically valid. Diligent efforts to standardize data collection and quality are clearly important, but so is the

implementation and explanation of a statistical method that properly accounts for random variation in outcomes.(1,9)

In the present study, we sought to determine whether the additional complexity of ML modeling might be worthwhile compared to standard methods of logistic regression. While theoretical considerations should lead us to expect fewer outliers using ML methods, this does not guarantee that the correct number of outliers will be identified. We therefore sought to test the methodologies in the practical sense of how well a model predicted future performance.

The structure of NIS and other hospitalization data is multilevel or hierarchical, with patients clustered within hospitals. Implicit assumptions of such clustering are that patients in a given hospital tend to be more like one another, in characteristics important to the analysis, than they are like patients at other hospitals. Multilevel models explicitly acknowledge that a small sample of patients from a hospital will predict, on average, the hospital effect with a larger error than a sample with a larger number of patients. This methodology potentially reduces prediction errors caused by small sample sizes and reduces mislabeling of hospitals as outliers simply because of small patient sample sizes.

Even a zero mortality may not be an accurate predictor of future performance, especially if it is based on a relatively small number of cases.(23)

Accumulating, submitting, and processing registry data at a state or national level typically takes at least two years, so we reasoned that a model constructed from data in Year N would typically be applied to predicting the performance of hospitals in Year N+2. The NIS offered the opportunity to see in retrospect how well the models would have worked if they had been applied for this purpose to hospitals that happened to be in the samples for both Year N and Year N+2.

Estimates of various patient cofactors were very similar in each year (Table 3), and were also similar to those obtained by Clark and Winchell(17) using other administrative data. In particular, we note again that the effect of increasing AIS is much more pronounced for head injuries than for injuries in other parts of the body. This finding argues against the use of a single calculated Injury Severity Score and in favor of some method that separates these effects by body region.

The fixed effects shown in Table 3 were also nearly identical regardless of whether a standard or a ML model was estimated. It is therefore not surprising that Cohen et al.(11) found no shrinkage and no difference between the results from MLLR and standard LR when the hospital effects were “estimated using only the fixed portion of the model”. However, the determination of u_{0j} and evaluation of the shrunken estimates are essential features of the ML approach, so we do not feel that their study should be taken as evidence that it has no advantages over standard LR.

For the purposes of simplifying this comparative study, we did not use any hospital-level predictors (e.g., teaching status), although these could easily be added. In particular, a registry-based comparison could include hospital level variables indicating trauma center status (Level I, ACS Verification, etc.), which is not possible in standard regression. One caution after applying hospital-level variables in a ML model is that estimates will then be shrunken toward the mean of any subgroup sharing the same hospital-level predictors, which may or may not be the intention of the analysis.

A significant limitation of our study is the reliance on NIS data that were not specifically collected for quality improvement. Administrative data in general do not have the depth of

descriptive detail that might be available in a registry or other dedicated database. Indeed, a previous comparison of ML and standard mortality models using New York clinical cardiac surgical data did not find that ML methods were preferable.⁽¹⁰⁾ While this may reflect differences between cardiac and trauma patients, it may also be true that better methods of risk adjustment at the patient level would reduce the proportion of variability assigned to hospitals with differing patient populations.

Another important limitation of the NIS (and other databases recording only outcomes at the time of hospital discharge) is that differences in survival at the time of discharge may not reflect a true difference in survival, since earlier discharge from the hospital may lead to some deaths not being observed and recorded in hospital data. This is particularly true of older patients, who may be discharged to nursing or rehabilitation facilities earlier or later in different regions of the country.⁽²⁴⁾ Thus, small differences in hospital survival should not necessarily be interpreted as differences in overall survival, and the apparent consistency of a model to predict this outcome may be influenced by its ability to predict practice patterns rather than its ability to predict clinical quality.

Comparison of standard regression to ML regression must consider the assumptions made by each model, and the purposes for which they are applied. For the theoretical reasons discussed above and elsewhere, the ML approach should be expected to be more reliable, because it shrinks the observation from one year toward what would be expected in the future assuming that the hospitals are truly exchangeable. On the other hand, standard regression does not allow individual hospitals to “borrow” data from the group as a whole, and might be preferable if the purpose of a model is to determine which hospitals should be rewarded (or penalized) for their actual performance in a given year.

With our samples of patients and hospitals, standard regression predicted more low outliers, while ML regression predicted more high outliers. This difference may be attributed to the distribution of hospital means, which does not necessarily follow a normal distribution as assumed by the ML model. Standard logistic regression evaluates performance with respect to the overall mean probability of mortality, while ML regression evaluates performance with respect to the mean of the hospital effects. Other samples of patients and hospitals may or may not affect the results in the same way, and arguments may be advanced whether one or the other mean is more appropriate.

Regardless of the differences we found between the two methods, neither was very accurate in predicting outliers using data from two years previously. Also, the difference between observed rates and predicted rates in the validation samples was quite large in relation to the overall mortality rates. Hospital mortality is a relatively infrequent outcome, and its variability over two years in a given hospital may be too large to allow effective prediction by any statistical method, especially when small differences among hospital outcomes may result from variability in inclusion criteria or hospital length of stay. Further studies using other databases may be useful to demonstrate or refute the contribution of ML models to the process of hospital quality improvement for trauma or other conditions.

Acknowledgments

Supported by Grant #R01HS015656 from the Agency for Healthcare Research and Quality, and a supplemental grant from the Maine Medical Center Research Strategic Plan.

References

1. Iezzoni, LL., editor. Risk Adjustment for Measuring Healthcare Outcomes. Chicago: Health Administration Press; 1997.

2. Raudenbush, SW.; Bryk, AS. Hierarchical Linear Models. London: Sage; 2002.
3. Snijders, TAB.; Bosker, RJ. Multilevel Analysis. London: Sage; 1999.
4. Gatsonis C, Normand SL, Liu C, et al. Geographic variation of procedure utilization. A hierarchical model approach. *Med Care*. 1993; 31(5 Suppl):YS54–YS59. [PubMed: 8492586]
5. Normand SL, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: Issues and applications. *J Am Statist Assoc*. 1997; 92:803–814.
6. DeLong ER, Peterson ED, DeLong DM, et al. Comparing risk-adjustment methods for provider profiling. *Stat Med*. 1997; 16:2645–2664. [PubMed: 9421867]
7. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: Comprehensive review and statistical critique. *Ann Thorac Surg*. 2001; 72(6):2155–2168. [PubMed: 11789828]
8. Moerbeek M, van Breukelen GJ, Berger MP. A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *J Clin Epidemiol*. 2003; 56(4):341–350. [PubMed: 12767411]
9. Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes. *Circulation*. 2006; 113(3):456–462. [PubMed: 16365198]
10. Hannan EL, Wu C, DeLong ER, Raudenbush SW. Predicting risk-adjusted mortality for CABG surgery: Logistic versus hierarchical logistic models. *Med Care*. 2005; 43(7):726–735. [PubMed: 15970789]
11. Cohen ME, Dimick JB, Bilimoria KY, et al. Risk adjustment in the American College of Surgeons National Surgical Quality Improvement Program: A comparison of logistic versus hierarchical modeling. *J Am Coll Surg*. 2009; 209:687–693. [PubMed: 19959035]
12. Rating the severity of tissue damage. I. The abbreviated scale. *JAMA*. 1971; 215(2):277–280. [PubMed: 5107365]
13. Clark, DE.; Hahn, DR.; Osler, TM. ICD programs for injury categorization. 2009. Available at <http://ideas.repec.org/s/boc/bocode.html>. Version 2.1
14. Recommended framework for presenting injury mortality data. *MMWR Morb Mortal Wkly Rep*. 1997; 46:1–30. [PubMed: 9011775]
15. Houchens, R.; Chu, B.; Steiner, C. HCUP Methods Series Report #2007-01. U.S. Agency for Healthcare Research and Quality; 2007. Hierarchical Modeling using HCUP Data. Available at <http://www.hcup-us.ahrq.gov/reports/methods.jsp>
16. Clark DE, Ahmad S. Estimating injury severity using the Barell matrix. *Inj Prev*. 2006; 12:111–116. [PubMed: 16595426]
17. Clark DE, Winchell RJ. Risk adjustment for injured patients using administrative data. *J Trauma*. 2004; 57:130–140. [PubMed: 15284563]
18. Champion HR, Copes WS, Sacco WJ, et al. The Major Trauma Outcome Study: Establishing national norms for trauma care. *J Trauma*. 1990; 30:1356–1365. [PubMed: 2231804]
19. Breslow, NE.; Day, NE. Volume II – The Design and Analysis of Cohort Studies. New York: Oxford University Press; 1987. Statistical Methods in Cancer Research.
20. Pollock DA. Summary of the discussion: Trauma registry data and TRISS evidence. *J Trauma*. 1999; 47:S56–S58. [PubMed: 10496612]
21. Clark DE, Hannan EL, Raudenbush SW. Using a hierarchical model to estimate risk-adjusted mortality for hospitals not included in the reference sample. *Health Services Research*. 2010 in press.
22. Kaplan RS, Norton DP. The balanced scorecard -- measures that drive performance. *Harv Bus Rev*. 1992; 70:71–79. [PubMed: 10119714]
23. Dimick JB, Welch HG. The zero mortality paradox in surgery. *J Am Coll Surg*. 2008; 206:13–16. [PubMed: 18155563]
24. Gorra AS, Clark DE, Mullins RJ, DeLorenzo MA. Regional variation in hospital mortality and 30-day mortality for injured Medicare patients. *World J Surg*. 2008; 32:954–959. [PubMed: 18224464]

Table 1

Characteristics of Injury Cases Included in the NIS Sample for 2002

	Eligible hospitals (n=643)		Comparable hospitals (n=172)	
	All cases (n=277,988)	Included cases (n=245,776)	All cases (n=76,034)	Included cases (n=66,991)
Age, y, mean	54.6	54.9	54.1	54.4
Male, %	50.3	50.0	50.1	50.6
Burn wound, %	1.4	0.1	1.5	0.1
Stab wound, %	3.1	3.3	3.0	3.3
Gunshot wound, %	2.2	2.4	2.3	2.6
Motor vehicle, %	20.9	22.1	22.4	23.7
Transfer in, %	3.6	3.3	3.9	3.7
Transfer out, %	1.9	0.0	1.7	0.0
Head AISmax 3, %	5.2	5.6	5.8	6.2
Head AISmax 4, %	3.8	4.0	4.0	4.2
Head AISmax 5, %	0.5	0.5	0.5	0.6
Body AISmax 3, %	34.4	36.4	36.1	38.3
Body AISmax 4, %	1.0	1.0	1.3	1.4
Body AISmax 5, %	0.6	0.7	0.8	0.8
No valid AIS, %	1.7	0.0	1.9	0
Died, %	2.5	2.7	2.8	3.0

The principal reason for case exclusion was patients recorded as elective admissions. "Eligible" hospitals had at least 25 injured patients, fewer than 10% transferred to another acute care hospital, and fewer than 1% whose vital status was unspecified at the time of discharge. "Comparable" hospitals were those eligible by this definition in both 2002 and 2004 NIS samples.

Table 2

Characteristics of Comparable Hospitals in the Prediction and Validation Samples

	n	Mean volume (range)	Mortality, % (range)
Prediction samples			
2002	172	389 (41 – 3337)	1.99 (0 – 6.31 %)
2003	149	335 (28 – 2353)	2.07 (0 – 5.77 %)
2004	171	493 (26 – 2612)	2.00 (0 – 7.14 %)
Validation samples			
2004	172	404 (33 – 3695)	1.82 (0 – 7.14 %)
2005	149	345 (27 – 2874)	1.83 (0 – 6.10 %)
2006	171	510 (31 – 2861)	2.00 (0 – 5.94 %)

Table 3
Covariate Effects (Beta Coefficients) for Standard LR and Random-Intercept Multilevel LR Models Predicting Hospital Mortality

	Standard logistic regression			Multilevel logistic regression		
	2002	2003	2004	2002	2003	2004
Age 0-14 y	-0.73 (.19)			-0.67 (.20)		
Age 45-54 y	0.48 (.10)	0.60 (.15)	0.56 (.09)	0.49 (.10)	0.62 (.15)	0.56 (.09)
Age 55-64 y	0.84 (.11)	0.96 (.15)	0.75 (.10)	0.87 (.11)	0.98 (.15)	0.77 (.10)
Age 65-74 y	1.37 (.10)	1.59 (.13)	1.37 (.08)	1.44 (.10)	1.64 (.12)	1.42 (.08)
Age 75-84 y	1.81 (.09)	1.88 (.11)	1.71 (.07)	1.92 (.09)	1.95 (.12)	1.79 (.08)
Age 85+ y	2.12 (.09)	2.32 (.11)	2.05 (.08)	2.24 (.09)	2.40 (.12)	2.18 (.08)
Male	0.40 (.05)	0.46 (.07)	0.40 (.05)	0.38 (.06)	0.46 (.07)	0.39 (.05)
Head AISmax 3	2.34 (.07)	2.50 (.10)	2.36 (.06)	2.26 (.08)	2.45 (.10)	2.26 (.06)
Head AISmax 4	2.48 (.08)	2.67 (.09)	2.43 (.07)	2.40 (.08)	2.64 (.09)	2.35 (.07)
Head AISmax 5	5.03 (.13)	5.52 (.17)	5.35 (.11)	4.96 (.13)	5.50 (.17)	5.31 (.11)
Body AISmax 3	0.95 (.06)	1.11 (.07)	1.01 (.05)	0.93 (.06)	1.10 (.07)	1.01 (.05)
Body AISmax 4	2.00 (.13)	1.86 (.22)	1.74 (.12)	1.93 (.13)	1.79 (.22)	1.70 (.12)
Body AISmax 5	2.17 (.17)	2.64 (.26)	2.32 (.15)	2.12 (.17)	2.60 (.26)	2.28 (.16)
Fall	-0.40 (.07)	-0.31 (.07)	-0.34 (.05)	-0.42 (.07)	-0.32 (.07)	-0.33 (.06)
Firearm	1.49 (.12)	1.35 (.19)	1.46 (.11)	1.35 (.12)	1.27 (.20)	1.39 (.11)
Motor vehicle	0.21 (.08)			0.11 (.08)		
Constant	-4.28 (.04)	-4.44 (.05)	-4.28 (.03)	-4.45 (.05)	-4.51 (.06)	-4.43 (.05)
Level 2 variance	-	-	-	0.106	0.068	0.105

Standard errors are given in parentheses. Terms for which a 95% confidence interval included zero were eliminated from the models in 2003 and 2004.

Table 4

Comparisons of the Ability of Standard LR and Random-Intercept Multilevel LR Models of Mortality to Predict Actual Outcomes in the Same Hospitals in the NIS Sample Two Years Later

	Standard logistic regression				Multilevel logistic regression			
	2002	2003	2004		2002	2003	2004	
All comparable hospitals:	(n=172)	(n=149)	(n=171)		(n=172)	(n=149)	(n=171)	
Absolute error (%), observed-expected	1.12	1.28	1.04		0.89	0.96	0.85	
Root-mean-square error, %, observed-expected	1.51	1.82	1.54		1.14	1.28	1.17	
Percentage of hospitals whose predicted deaths were within 95% CI of observed	80.2	82.6	88.3		95.4	95.3	97.7	
Hospitals with some mortality:	(n=139)	(n=117)	(n=145)		(n=139)	(n=117)	(n=145)	
Absolute error, %, observed-expected	1.01	0.98	0.84		0.77	0.75	0.69	
Root-mean-square error, %, observed-expected	1.33	1.28	1.16		1.03	1.07	0.97	
Percentage of hospitals whose predicted deaths were within 95% CI of observed	84.9	88.0	92.4		95.7	94.0	97.2	

For each measurement in each year, the difference between standard LR and random-intercept multilevel LR models was significant ($p < 0.001$, Wilcoxon signed rank test).

Table 5

Outlier Status of Hospitals in the Validation Samples from 2004–2006, Compared to the Outlier Status of the Same Hospitals in the Prediction Samples from 2002–2004

	Observed outlier status, by method and year											
	Neither method			Standard only			Multilevel only			Both methods		
	2004	2005	2006	2004	2005	2006	2004	2005	2006	2004	2005	2006
Predictions												
Outlier by neither method 2 years before	147	140	151	5 L	2 H	4 L	2 H*					2 H*
Outlier by standard only 2 years before	3 H 9 L	1 H 3 L	4 L	1 H 1 L		1 L						
Outlier by multilevel only 2 years before	2 H		7 H				2 H* [†]	1 H* [†]		1 H* [†]	1 H* [†]	1 H* [†]
Outlier by both methods 2 years before	2 H	2 H	2 H 2 L				2 H* [†]		1 H* [†]	2 H* [†]	1 H* [†]	2 H* [†]

Each cell shows the number of hospitals that were high (H) or low (L) outliers using each modeling method. Standard LR models identified 39 outliers in one or more of the prediction samples (rows 2 and 4). Multilevel LR models identified 31 outliers in the prediction samples (rows 3 and 4).

* Eighteen outliers identified in the validation samples by MLLR.

[†] Of those, 14 had been identified in the corresponding prediction samples for a sensitivity of 78%. The corresponding sensitivity for standard LR was 11/30=37%.