© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2009.01074.x
METHODS BRIEF

Using a Hierarchical Model to Estimate Risk-Adjusted Mortality for Hospitals Not Included in the Reference Sample

*David E. Clark, Edward L. Hannan, and
Stephen W. Raudenbush*

Objective. To provide a method for any hospital to evaluate patient mortality using a hierarchical risk-adjustment equation derived from a reference sample.

Data Source. American College of Surgeons National Trauma Data Bank (NTDB).

Study Design. Hierarchical logistic regression models predicting mortality were estimated from NTDB data. Risk-adjusted hospital effects obtained directly from models using standard software were compared with approximations derived from a summary equation and data from each individual hospital.

Principal Findings. Theoretical approximations were similar to results using standard software.

Conclusions. To allow independent verification, agencies using reference databases for hospital mortality “report cards” should publish their risk-adjustment equations. Similar hospitals not in the reference database may also use the published equations along with the approximations described to evaluate their own outcomes using their own data.

Key Words. Mortality, risk adjustment, hierarchical models, multilevel models

There is increasing interest in methods to evaluate hospital performance, especially patient mortality, with respect to some group standard. Hospital “profiles” for cardiac surgery have been published in several regions, leading to careful study of the associated statistical issues (Gatsonis et al. 1995; Normand, Glickman, and Gatsonis 1997; Hannan et al. 2005). Extending this approach, the Agency for Healthcare Research and Quality (AHRQ) has published Inpatient Quality Indicators (IQI), described at <http://qualityindicators.ahrq.gov>, that allow hospitals to compare their performance in several areas to the outcomes expected from the reference databases of the AHRQ Healthcare Cost and Utilization Project (HCUP). The American College of

Surgeons (ACS) has also been a leader in this effort, by supporting the development of a National Trauma Data Bank (NTDB) to combine registry data from multiple trauma centers and compare their outcomes.

Many hospital performance evaluations today rely on multilevel (ML) or hierarchical logistic regression modeling. While these methods can have theoretical advantages in modeling institutional mortality and are increasingly accessible due to software advances, they can present challenges as well. For example, the AHRQ or ACS NTDB data may be used to compute performance measures for a set of hospitals, but other hospitals not included in the original analysis may also be interested in comparing their outcomes to these standards. The purpose of this paper is to derive a method by which such a comparison can be made.

The method we propose could be useful in several contexts. Analysts can use this method to compute performance measures for hospitals that were not included in an original analysis. Hospitals that were included in the original analyses can also use this method to verify “report cards” provided by an agency conducting the analysis. Hospitals or other observers can use this method to derive updated scores based on the original model but incorporate newer data that become available. Finally, experiments can be conducted to see how the scores might change if certain aspects of the data were to change.

METHODS

Let us consider the problem of comparing hospitals with respect to their mortality rates after statistically adjusting for the background characteristics of their patients. Define the outcome variable for patient i as $y_{ij} = 1$ if that patient died and $y_{ij} = 0$ if not, for $i = 1, \dots, n_j$ patients in hospital j with $p_{ij} = \text{prob}(y_{ij} = 1)$, the probability of death. We expect this probability to depend upon patient background characteristics $\mathbf{x}_{ij} = (x_{1ij}, x_{2ij}, \dots)$ and also upon a hospital-specific contribution b_j . The aim is to get a good estimate of b_j for each hospital, so that each hospital can be compared with a set of reference hospitals. If hospital j is one in a large sample of hospitals, we might formulate a

Address correspondence to David E. Clark, M.D., Department of Surgery, Maine Medical Center, 887 Congress Street, Suite 210, Portland, ME 04102; e-mail clarkd@mmc.org. Edward L. Hannan, Ph.D., is with the Department of Health Policy, School of Public Health, State University of New York, University at Albany, One University Place, Rensselaer, NY. Stephen W. Raudenbush, Ph.D., is with the Department of Sociology, University of Chicago, Chicago, IL.

logistic regression model for this purpose (Iezzoni et al. 1992):

$$\begin{aligned}\log(p_{ij}/(1 - p_{ij})) &= \text{logit } p_{ij} \\ &= \mathbf{x}_{ij}^T \mathbf{b} + b_j = b_1 x_{1ij} + b_2 x_{2ij} + \dots + b_m x_{mij} + b_j\end{aligned}\quad (1)$$

where $\mathbf{b} = (b_1, b_2, \dots, b_m)$ are constant coefficients and b_j is a fixed effect associated with hospital j . The odds ratio obtained from exponentiating b_j may be used as a measure of effect assuming that the covariates in the model correctly incorporate all other sources of variability affecting mortality. The significance of an estimate of b_j may be judged by a confidence interval or p -value obtained from maximum likelihood theory (Hosmer and Lemeshow 2000).

Allowance for “clustering” of patient characteristics has led to the development of “hierarchical” or “ML” models (Normand, Glickman, and Gatsonis 1997; Snijders and Bosker 1999; Raudenbush and Bryk 2002; Goldstein 2003). The simplest ML model of mortality, known as the “random intercept model,” uses the logit link function just as in equation (1), but allows the intercept term to vary randomly by hospital. The specific hospital intercepts b_j are regarded as distributed around a mean b_0 with a variance v_2 to be estimated from the data. In most applications, analysts have assumed that the hospital effects are independently normally distributed, that is,

$$b_j = b_0 + u_j, u_j \sim N(0, v_2) \quad (2)$$

so that u_j is a hospital-specific random effect with respect to the overall mean intercept b_0 .

In the ML theory, the hospital’s contribution to mortality predicted by this equation is a “shrunk” estimate that weights the observed mortality by its reliability; that is, the hospital-specific estimate of u_j is shrunk or “pulled” toward zero, with the hospitals producing the least data for estimation experiencing the greatest shrinkage. Theoretical considerations and empirical evidence (Raudenbush and Bryk 2002) suggest that, on average, the shrunk estimates will tend to be more accurate than those based on the fixed effects model. The ML method has been used for numerous comparisons of inter-hospital or intersurgeon differences (Gatsonis et al. 1995; DeLong et al. 1997; Moerbeek, van Breukelen, and Berger 2003). The major drawback to the use of ML modeling has been computational complexity, which is only recently being overcome.

The ML approach offers several theoretical advantages, including (1) appropriately modeling the hierarchical nature of the data and consequent correlation of outcomes within hospitals; (2) reducing the incidence of out-

lying mortality estimates based on scant data; and (3) allowing for incorporation of hospital characteristics in the model. In addition, ML models produce estimates for hospitals with no observed mortality (whose performance must be shrunk at least slightly toward the mean), whereas standard regression will fail due to collinearity.

Let us now consider the problem of obtaining a reasonable shrunk estimate for some hospital, call it hospital q , that did not contribute data for the reference sample from which estimates of the parameters $(b_0, b_1, b_2, \dots, b_m, v_2)$ were obtained. If certain assumptions to be described below are met, we can compute a good estimate as follows (see Appendix A for the derivation):

1. Estimate the parameters $(b_0, b_1, b_2, \dots, b_m, v_2)$ from the hospitals $j = 1, \dots, J$ in the reference dataset using whatever ML software is most practical and effective. Let us denote these estimates as $(\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_m)$. This step can be carried out, and the results published, by a central agency like the ACS with access to the entire database; the remaining steps of the algorithm can be carried out by any hospital q , with access only to its own data and the published model, regardless of whether hospital q is in the reference database.
2. For each patient i in hospital q , compute a provisional risk-adjusted probability of mortality $p_{iq}^{(0)} = 1/[1 + \exp\{-(\hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \dots + \hat{b}_m x_{mi} + \hat{b}_q^{(0)})\}]$. The estimate $\hat{b}_q^{(0)}$ of the specific hospital effect is equal to the sum of the estimated intercept from the reference sample, that is, $\hat{b}_0$, and a provisional estimate of the random effect, denoted $\hat{u}_q^{(0)}$. It would typically make sense to set the initial value $\hat{u}_q^{(0)} = 0$.
3. Compute the next estimate of the random effect as follows:
 - 3a. Estimate the *raw effect* of hospital q at iteration 1 as

$$r_q^{(1)} = \frac{\sum_{i=1}^{n_q} y_{iq} - \sum_{i=1}^{n_q} p_{iq}^{(0)}}{\sum_{i=1}^{n_q} p_{iq}^{(0)} (1 - p_{iq}^{(0)})}$$

- 3b. Calculate an approximate Level-1 variance (on the logit scale) at iteration 1 as

$$v_{1q}^{(1)} = 1 / \sum_{i=1}^{n_q} p_{iq}^{(0)} (1 - p_{iq}^{(0)})$$

3c. Partition the total variance so that it can be used as a “shrinkage factor,” calculating the quantity $\lambda_q^{(1)}$ as

$$\lambda_q^{(1)} = \frac{v_2}{v_2 + v_{1q}^{(1)}}$$

3d. Compute a corrected (“shrunk”) estimate of the hospital random effect as

$$u_q^{(1)} = \lambda_q^{(1)}(r_q^{(1)} + u_q^{(0)})$$

3e. Now update the patient-specific predicted probability of mortality as

$$p_{iq}^{(1)} = 1/[1 + \exp\{-(\hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \dots + \hat{b}_m x_{mi} + \hat{b}_q^{(1)})\}]$$

where $\hat{b}_q^{(1)} = \hat{b}_0 + u_q^{(1)}$.

4. Reiterate steps 3a–3e, estimating $(r_q^{(k)}, v_{1q}^{(k)}, \lambda_q^{(k)}, u_q^{(k)})$ from $(\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_2)$ and $(p_{iq}^{(k-1)}, u_q^{(k-1)})$ until the estimate $u_q^{(k)}$ at iteration k is negligibly different from the estimate $u_q^{(k-1)}$ at iteration $k-1$.

The final estimate of the risk-adjusted hospital mortality effect can be considered as $u_q^{(k)}$, and the final estimate of the shrinkage factor can be considered as $\lambda_q^{(k)}$. An approximate variance of u_q will be $v_2(1 - \lambda_q)$, which can be used to construct confidence intervals or test pairwise differences between hospitals (Snijders and Bosker 1999). This measure of effect on the logit scale may be exponentiated to approximate an observed/expected (O/E) ratio for each hospital (DeLong et al. 1997; Hannan et al. 2005).

One key assumption underlying this procedure is that the random intercept model generating the data for the reference sample of hospitals $j = 1, \dots, J$ is also the correct model for generating the data for hospitals $q = 1, \dots, Q$ of interest but not in the reference sample. This assumption would appear plausible when hospital q draws its patients from a population that is similar to those represented by the reference sample, and it can be checked if a summary of background information on patients in the reference sample is available. For this reason, registries or databases to be used for interhospital comparisons should report summary background data on patients.

A second assumption is that the number of hospitals in the reference sample is large enough to assume that the parameters $(b_0, b_1, b_2, \dots, b_m, v_2)$ are only negligibly different from their estimates. These assumptions are in ad-

dition to the usual assumptions of the random intercept model (normality of random effects, constant random effects variance, linearity of the log-odds of mortality in the covariates).

VALIDATION AND RESULTS

Our strategy for validating the procedure was first to compute estimates for a large reference sample containing data from hospitals $j = 1, \dots, J$. We took the shrinkage estimators of the hospital-specific random effects u_j , based on standard software, as the criterion for evaluating our procedure. Next, for each hospital j , we used the parameter estimates $(\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{v}_2)$ to compute hospital-specific effects as described in steps 1–4 above. We reasoned that if our procedure is correct, it should essentially reproduce the criterion estimates generated by the software.

NTDB data for 2001–2005 were obtained from the ACS in compliance with its standard Data Use Agreement. Relevant patient variables were extracted, and independent variables were centered on their population means. ML logistic regression models predicting hospital mortality for cases of blunt trauma were created using the “xtmelogit” command in *Stata* (version 10, Stata Corp., College Station, TX). The number of quadrature points was started at 1 (Laplace approximation) and increased in increments of 2 until the variance of the random effect did not change at three significant digits. After estimation of the random effects model, a posterior modal estimate of the effect for each hospital, and its standard error, were generated using the “predict, reffects” and “predict, reses” commands in *Stata*. These estimates were confirmed using the Laplace approximation in *HLM6* (Scientific Software International, Lincolnwood, IL). The hospital-specific effects and their variances were obtained after estimation in *HLM6* from the Level-2 residual file as the variables “ebintrcp” and “pv00.”

ML models and hospital effects estimated by *Stata* and *HLM6* were virtually identical. Hospital effects (u_{0j}) obtained using the algorithm described above (steps 1–4) converged within a few iterations to the effects reported by the software packages.

DISCUSSION

Analytical methods for hospital profiling must be practical and statistically valid, and they must be presented in such a way that they can be understood,

and therefore trusted, by the institutions being evaluated. In many cases, hierarchical (ML) methods are now being used to generate estimates of hospital performance. However, it is not uncommon to encounter situations where it would be desirable to compute a performance score for a hospital that was not included in the original analysis, or to provide institutions with the ability to verify their published scores. We have presented a method for deriving such estimates.

The idea of publishing a risk adjustment equation to enable trauma centers to compare their mortality to a reference database originated with the Major Trauma Outcome Study (MTOS) coordinated by the ACS more than 20 years ago (Champion et al. 1990). As successor to the MTOS, the NTDB has steadily increased the number of contributing hospitals, but a third of the major trauma centers in the nation still find themselves unable to participate due to state regulations, financial constraints, personnel transitions, legal concerns, or other temporary or permanent factors. A method such as we describe would still allow these institutions to use the NTDB as a quality improvement tool, using ML statistical models if desired.

The AHRQ has taken a similar approach in the publication of their IQI. Any hospital may input its own data into the IQI computer programs and compare its performance in several areas to the outcomes expected from the HCUP databases. Since 2007, the equivalent of a ML or hierarchical method has been imbedded in the IQI risk-adjustment software, although the derivations of the specific models and algorithms are explained only in general terms.

These methods assume that if one more hospital were added to the reference database, the regression coefficients derived without that hospital would change very little. Therefore, they are valid only if the reference database is very large compared with the size of any one hospital, and only if the patient population at the additional hospital is sufficiently similar to that from which the reference database was derived (Shahian and Normand 2008). To improve validity, separate risk-adjustment equations could be provided for different geographic regions, hospital characteristics, etc., or additional variables reflecting these differences can be included in the equations.

We have described methods and theoretical justifications by which shrunken risk-adjusted hospital effects on mortality can be obtained for hospitals not necessarily included in a reference database, assuming their patient populations and data management are similar. In the interest of maximum transparency in the process of hospital profiling, we support the practice of making risk-adjustment equations public so that hospitals in a reference

database may use these methods to verify their evaluations and other hospitals may evaluate themselves with respect to this standard.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was supported in part by grant R01HS015656 from the Agency for Healthcare Research and Quality, and a supplemental grant from the Maine Medical Center Research Strategic Plan. Dr. Clark has been a member and chairman of the National Trauma Data Bank Subcommittee of the American College of Surgeons Committee on Trauma. Dr. Raudenbush is an author of the *HLM6* software.

Disclosures: None.

Disclaimers: None.

APPENDIX A: Theoretical Background

Let $y_{ij} = 1$ if patient i in hospital j dies and 0 if that patient survives. We assume that the probability of death for this patient may be expressed as

$$p_{ij} = 1 / (1 + \exp\{-x_i^T \beta + u_j\})$$

where u_j is a hospital-specific random effect distributed as $\text{Normal}(0, v_2)$. The parameters (β, v_2) have presumably been estimated from an appropriate large-scale database in which hospital j is not necessarily included. Indeed, we assume that the reference database is so large that the estimates of (β, v_2) are essentially equivalent to the parameters.

We can model the outcome as

$$y_{ij} = p_{ij} + e_{ij}, E(e_{ij}) = 0, \text{Var}(e_{ij}) = p_{ij}(1 - p_{ij}) \quad (\text{A1})$$

Expanding p_{ij} in a first-order Maclaurin series, we have

$$\begin{aligned} p_{ij} &\approx \hat{p}_{ij} + \left. \frac{dp_{ij}}{du_j} \right|_{u_j=0} u_j \\ &= \hat{p}_{ij} + \hat{p}_{ij}(1 - \hat{p}_{ij})u_j \end{aligned} \quad (\text{A2})$$

Substituting (A2) into (A1) yields the linear model

$$y_{ij} \approx \hat{p}_{ij} + \hat{p}_{ij}(1 - \hat{p}_{ij})u_j + e_{ij} \quad (\text{A3})$$

Moving all observables to the left-hand side and setting $w_{ij} = \hat{p}_{ij}(1 - \hat{p}_{ij})$, we

have the linearized dependent variable

$$y_{ij}^* = \frac{y_{ij} - \hat{p}_{ij}}{w_{ij}} \approx u_j + \frac{e_{ij}}{w_{ij}} \quad (\text{A4})$$

Note that the error term in this model has mean zero and variance

$$\text{Var}\left(\frac{e_{ij}}{w_{ij}}\right) = \frac{p_{ij}(1 - p_{ij})}{w_{ij}^2} \approx \frac{1}{w_{ij}} \quad (\text{A5})$$

This gives us a weighted least squares problem solved by using the inverse variance w_{ij} as the weight in estimating equation (A4):

$$\begin{aligned} \hat{u}_j &= \frac{\sum_{i=1}^{n_j} w_{ij} y_{ij}^*}{\sum_{i=1}^{n_j} w_{ij}} = \frac{\sum_{i=1}^{n_j} (y_{ij} - \hat{p}_{ij})}{\sum_{i=1}^{n_j} w_{ij}} \\ \text{Var}(\hat{u}_j) &= \frac{1}{\sum_{i=1}^{n_j} w_{ij}} \equiv v_{1j} \end{aligned} \quad (\text{A6})$$

We then assume that the true u_j and the estimate $\hat{u}_j$ are bivariate normal, that is

$$\begin{pmatrix} u_j \\ \hat{u}_j \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_2 & v_2 \\ v_2 & v_2 + v_{1j} \end{pmatrix}\right] \quad (\text{A7})$$

from which we can obtain the conditional distribution of u_j given $\hat{u}_j$ as

$$u_j | \hat{u}_j \sim N[\lambda_j \hat{u}_j, v_2(1 - \lambda_j)] \quad (\text{A8})$$

where $\lambda_j = v_2/(v_2 + v_{1j})$. We may then use the conditional mean and variance of u_j given $\hat{u}_j$ from (A8) to make inferences about the unknown u_j .

The approximation can be improved by expanding equation (A1) in a Taylor series around $u_j = \hat{u}_j$:

$$y_{ij} \approx p_{ij}^{(2)} + p_{ij}^{(2)}(1 - p_{ij}^{(2)})(u_j - \hat{u}_j) + e_{ij} \quad (\text{A9})$$

where $p_{ij}^{(2)} = 1/(1 + \exp\{x_i^T \beta + \hat{u}_j\})$. Now we can compute a new linearized dependent variable

$$y_{ij}^* = \frac{y_{ij} - \hat{p}_{ij}^{(2)}}{w_{ij}} + \hat{u}_{0j} \approx u_j + \frac{e_{ij}}{w_{ij}} \quad (\text{A10})$$

where we now use $w_{ij} = p_{ij}^{(2)}(1 - p_{ij}^{(2)})$. We now compute the weighted least squares estimate

$$\hat{u}_j = \frac{\sum_{i=1}^{n_j} w_{ij} y_{ij}^*}{\sum_{i=1}^{n_j} w_{ij}}, \quad \text{Var}(\hat{u}_j) = \frac{1}{\sum_{i=1}^{n_j} w_{ij}} \equiv v_{1j} \quad (\text{A11})$$

Inference proceeds as before, and the process of iteration can continue until convergence.

REFERENCES

- Champion, H. R., W. S. Copes, W. J. Sacco, M. M. Lawnick, S. L. Keast, L. W. Bain Jr., M. E. Flanagan, and C. F. Frey. 1990. "The Major Trauma Outcome Study: Establishing National Norms for Trauma Care." *Journal of Trauma* 30: 1356–65.
- DeLong, E. R., E. D. Peterson, D. M. DeLong, L. H. Muhlbaier, S. Hackett, and D. B. Mark. 1997. "Comparing Risk-Adjustment Methods for Provider Profiling." *Statistics in Medicine* 16: 2645–64.
- Gatsonis, C. A., A. M. Epstein, J. P. Newhouse, S. L. Normand, and B. J. McNeil. 1995. "Variations in the Utilization of Coronary Angiography for Elderly Patients with an Acute Myocardial Infarction. An Analysis Using Hierarchical Logistic Regression." *Medical Care* 33: 625–42.
- Goldstein, H. 2003. *Multilevel Statistical Models*. London: Hodder Arnold.
- Hannan, E. L., C. Wu, E. R. DeLong, and S. W. Raudenbush. 2005. "Predicting Risk-Adjusted Mortality for CABG Surgery. Logistic versus Hierarchical Logistic Models." *Medical Care* 43: 726–35.
- Hosmer, D. W. Jr., and S. Lemeshow. 2000. *Applied Logistic Regression*. 2d Edition. New York: John Wiley & Sons.
- Iezzoni, L. I., A. S. Ash, G. A. Coffman, and M. A. Moskowitz. 1992. "Predicting In-Hospital Mortality. A Comparison of Severity Measurement Approaches." *Medical Care* 30: 347–59.
- Moerbeek, M., G. J. van Breukelen, and M. P. Berger. 2003. "A Comparison between Traditional Methods and Multilevel Regression for the Analysis of Multicenter Intervention Studies." *Journal of Clinical Epidemiology* 56: 341–50.
- Normand, S.-L. T., M. E. Glickman, and C. A. Gatsonis. 1997. "Statistical Methods for Profiling Providers of Medical Care: Issues and Applications." *Journal of the American Statistical Association* 92: 803–14.
- Raudenbush, S. W., and A. S. Bryk. 2002. *Hierarchical Linear Models*. London: Sage.
- Shahian, D. M., and S.-L. T. Normand. 2008. "Comparison of 'Risk-Adjusted' Hospital Outcomes." *Circulation* 117: 1955–63.
- Snijders, T. A. B., and R. J. Bosker. 1999. *Multilevel Analysis*. London: Sage.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.