Bayesian-Logit Model for Risk Assessment in Coronary Artery Bypass Grafting

Guillermo Marshall, PhD, A. Laurie W. Shroyer, PhD, Frederick L. Grover, MD, and Karl E. Hammermeister, MD

Denver Department of Veterans Affairs Medical Center, Department of Preventive Medicine and Biometrics, Department of Medicine, and Division of Cardiothoracic Surgery, University of Colorado Health Sciences Center, Denver, Colorado

Predictive models for the assessment of operative risk using patient risk factors have gained popularity in the medical community as an important tool for the adjustment of surgical outcome. The Bayes' theorem model is among the various models used to predict mortality among patients undergoing coronary artery bypass grafting procedures. Comparative studies of the various classic statistical techniques, such as logistic regression, cluster of variables followed by a logistic regression, a subjectively created sickness score, classification trees model, and the Bayes' theorem model, have shown that the Bayes' model is among those with the highest predictive power. In this study, the Bayes' theorem model is reformulated as a logistic equation and extended to include qualitative and quantitative risk factors. We show that the resulting model, the Bayesian–logit model, is a mixture of logistic regression and linear discriminant analysis. This new model can be created easily without complex computer programs. Using 12,712 patients undergoing coronary artery bypass grafting procedures at the Department of Veterans Affairs Continuous Improvement in Cardiac Surgery Study between April 1987 and March 1990, the predictive power of the Bayesian–logit model is compared with the Bayes' theorem model, logistic regression, and discriminant analysis. The ability of the Bayesian–logit model to discriminate between operative deaths and operative survivors is comparable with that of logistic regression and discriminant analysis.

(Ann Thorac Surg 1994;57:1492–500)

Predictive models for surgical outcomes, such as mortality and postoperative complications, have been used to compute risk-adjusted rates and assess the quality of cardiac surgery care performed by hospital and individual surgeons [1, 2]. In particular, the Continuous Improvement in Cardiac Surgery Study (CICSS) of the Department of Veterans Affairs (VA), since its inception in 1987, has developed predictive models and used risk-adjusted mortality rates to monitor quality of care among the 43 VA medical center programs performing cardiac operations nationwide.

The classic logistic regression model is one of the most commonly used modeling alternatives for predicting dichotomous outcomes, such as mortality and other surgical outcomes. More recently, biostatisticians and clinical researchers have proposed alternative methods to create predictive models. Included in this list are a Bayes' theorem model [3, 4], an additive model [5], a model consisting in cluster analysis combined with clinical judgment followed by logistic regression [6, 7], classification and regression trees [7–9], a model consisting in principal components analysis to reduce the number of predictor followed by a logistic regression model [6, 7], a linear discriminant analysis [10], and a subjective sickness score [7]. There has also been interest in comparing the predictive power of these alternative models. Harrell and colleagues [6, 9], Cook and Goldman [8], and more recently Marshall and colleagues [7] have compared these different strategies to develop models for predicting adverse outcomes. The predictive power of a model can be measured using a general *c*-index [6], which for binary outcomes are reduced to the area under the receiver operator characteristic (ROC) curve as described by Hanley and McNeil [11].

The Society of Thoracic Surgeons National Cardiac Surgery Database has used the Bayes' theorem model to predict mortality among patients undergoing coronary artery bypass procedures. In this study, we review the Bayes' theorem model to predict operative mortality in patients undergoing cardiac operations. We show that the model can be reformulated using the same logistic equation used in the classic logistic regression model to produce a new model, the Bayesian–logit model. In fact, these models only differ in the methodology used to estimate the regression coefficients, but not in the structural relationship between the risk factors and the outcome. Also, the Bayesian–logit model produces coefficients for covariates indicating the strength of their relationship with the outcome variable. This result, which is not available with the standard Bayes' theorem model, should make the Bayesian–logit model more attractive to clinicians and increases the potential use and interpretation of the model. We also show that the regression coefficients can be easily computed using marginal frequency tables, instead of the iterative procedures required

Presented at the Fortieth Annual Meeting of the Southern Thoracic Surgical Association, Panama City Beach, FL, Nov 4–6, 1993.

Address reprint requests to Dr Marshall, Cardiac Research (151), Department of Veterans Affairs Medical Center, 1055 Clermont, Denver, CO 80220.

Table 1. Sample Size, Number of Events, and Percent Operative Mortality in the Learning and Test Samples and in the Total Population

Sample	Size	Events	Percent Mortality
Learning	6,317	285	4.51%
Test	6,395	297	4.64%
Total	12,712	582	4.58%

in logistic regression analysis. The Bayesian–logit model is able to incorporate quantitative variables in the same form as used in linear or logistic regression models, thus avoiding the need for arbitrary cut points in continuous variables. The Bayesian–logit model is the mixture of a logistic regression model and a linear discriminant analysis, assuming conditional independence among the risk factors.

Finally, the predictive powers of the logistic model, the Bayes' theorem model, the new Bayesian–logit model, and the linear discriminant model are compared using 12,712 patients undergoing coronary artery bypass procedures entered into CICSS between April 1987 and March 1990.

Material and Methods

Data

During the first 3 years (April 1987 through March 1990) of operation of CICSS [12], we received 15,444 data forms from 43 VA medical centers on patients undergoing cardiac operations. Of these, 12,712 underwent coronary artery bypass grafting (CABG) as the primary procedure, 2,326 underwent valve procedures, and 406 underwent other procedures. The initial efforts at assessing predictive accuracy among these models were confined to the 12,712 patients undergoing CABG, because of the larger population size and likelihood of greater homogeneity of the patient population.

We divided this population randomly into a learning sample (used to create the predictive model) and a test sample (used to test the predictive model) of approximately equal sizes using a pseudo-random number-generating function. The sample size and the number of deaths in the two samples were not forced to be equal to prevent the introduction of artificial correlation between the two samples. Nevertheless, the sample size, the number of events (deaths), and the percent operative mortality were very similar in the two samples as shown in Table 1.

From this large data set, we selected 43 preoperative variables thought to be associated with operative mortality. Of these, 10 variables with more than 20% of data missing were excluded. The remaining 33 variables and their percentage of missing data are shown in Table 2.

The number of patients in the learning data set provides the power to detect an odds ratio of 1.2 in a logistic regression model with various predictor variables. This sample size was calculated using the method described by Hsieh [13], assuming a multiple correlation coefficient of $r = 0.5$ among the predictor variables, with a statistical power of 90%, and a significance level of 5%.

The Bayes' Model

Edwards and associates [3, 4] proposed the use of a model based on the Bayes' theorem as a tool to predict outcome for patients undergoing cardiac operations. The model computes the probability of an operative death given a set of patient risk factors, $X = \{x_1, x_2, \ldots, x_m\}$, such as age and prior heart operation. Using Bayes' theorem, the conditional probability of an operative death given a specific set of risk factors, $P\{death|X\}$, can be calculated as

$$P\{death|X\} = \frac{P\{X|death\} \times P\{death\}}{P\{X|survival\} \times P\{survival\} + P\{X|death\} \times P\{death\}}, \quad (1)$$

if we know the distribution of the risk factors $X = \{x_1, x_2, \ldots, x_m\}$ in the two separated outcome groups, deaths and survivors, represented here by the probability terms $P\{X|death\}$ and $P\{X|survival\}$, respectively. In addition, we need information about the overall mortality rate in the entire population, $P\{death\}$, and the complementary probability $P\{survival\} = 1 - P\{death\}$. For example, among the patients undergoing CABG, 25.7% of the operative deaths had prior cardiac operations, whereas there were only 8.9% among the survivors. If the overall operative mortality rate is 4.5%, the probability (after knowing the risk factors) of operative mortality for a patient with prior cardiac operation is

$$P\{death|\text{prior cardiac operation}\} = \frac{0.257 \times 0.045}{0.089 \times 0.955 + 0.257 \times 0.045} = 0.12,$$

or in terms of proportion equal to 12%.

Considering that $X = \{x_1, x_2, \ldots, x_m\}$ may include a large number of risk factors, the evaluation of the probability $P\{X|death\}$ and the probability $P\{X|survival\}$ involves a multivariate distribution and the correlation among these variables. For simplicity, one can assume that these risk factors are conditionally independent given the operative outcome. The computational implication of this assumption is that now we can evaluate $P\{X|death\}$ as the product of marginal distributions (see Appendix 1), that is, as

$$P\{X|death\} = P\{x_1|death\} \times P\{x_2|death\} \times \cdots \times P\{x_m|death\}.$$

Therefore, probabilities, such as $P\{x_i|death\}$, can be estimated by computing the marginal distribution of x_i among the operative deaths in the same way that we did for prior cardiac operation.

Ann Thorac Surg
1994;57:1492–500

Table 2. Variable Names With Less Than 20% Missing Values in the Learning Sample Used to Assess the Predictive Models, Description of the Variables, and the Percentage of Missing Values in the Learning Sample

Variable Name	Description	Missing Values (%)
AGE	Age (y)	0.2
ANGIOP	Angioplasty ≤7 days of operation[a]	1.5
BSA	Body surface area (m^2)	2.5
CHF	Congestive heart failure[a]	2.3
CM	Cardiomegaly (x-ray)[a]	4.1
COPD	Chronic obstructive pulmonary disease[a]	3.3
CR	Creatinine (mg/dL)	14.2
CURCAB	Current calcium channel blocker use[a]	5.9
CURDIG	Current digoxin use[a]	5.6
CURDIUR	Current diuretic use[a]	5.5
CURRBB	Current beta blocker use[a]	5.9
CURRSMOK	Current smoker[a]	5.3
CVD	Cerebral vascular disease[a]	5.8
CXS	Circumflex (% stenosis)	13.9
DIABETES	Diabetes[a]	5.6
ECGLVH	ECG LVH[a]	7.7
EVERSMOK	Ever smoker[a]	5.3
EXANG	Exertional angina[a]	5.6
HTN	Hypertension[a]	5.6
IVNTG	Intravenous nitroglycerine preoperatively[a]	5.6
LAD	Left anterior descending (% stenosis)	9.4
NYHAFC	New York Heart Association functional classification (I, II, III, IV)	6.2
OLDMI	Old myocardial infarction (>30 d)[a]	4.9
PRIORHS	Prior heart operation	0.3
PRIORITY	Priority of operation (elective, urgent, emergent)	4.2
PROPIABP	Preoperative intraaortic balloon pump[a]	1.9
PULMR	Pulmonary rales[a]	5.9
PVD	Peripheral vascular disease[a]	5.5
RECMI	Recent myocardial infarction (≤30 d)[a]	3.0
RESTANG	Unstable angina[a]	6.4
RESTSTD	Resting ST-segment depression[a]	7.3
RCHS	Right coronary artery (% stenosis)	11.4
SEX	Sex (0 = male, 1 = female)	0.2

[a] 1 = present, 0 = absent.

According to the model used by Edwards and colleagues [3, 4], the indicators x_i in the Bayes' model are chosen as dichotomous variables representing whether or not the risk factor is present; these expressions can be better visualized by creating a 2 × 2 frequency table (Table 3). To evaluate the Bayes' model, the individual probabilities $P\{x_i = 1|death\}$ and $P\{x_i = 1|survival\}$ are estimated by the observed ratios $a_i/(a_i + b_i)$ and $c_i/(c_i + d_i)$, respectively.

Table 3. Bayes' Model 2 × 2 Frequency Table

Outcome	Risk Factor: Present ($x_i = 1$)	Risk Factor: Absent ($x_i = 0$)
Death	a_i	b_i
Survival	c_i	d_i

Reformulating the Bayes' Theorem Model

Edwards and colleagues [3, 4] point out that model (1) is a valuable clinical tool for predictive purposes, but with sufficient mathematical complexity to require a computer program to solve the equation. In this study, we propose to reformulate the Bayes' theorem model (1) into a simpler and familiar form, that is, using a logistic equation. Without making any assumption, it is possible to formulate the Bayes' theorem model as

$$P\{death|X\} = \frac{\exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m\}}{1 + \exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m\}}, \quad (2)$$

where $\exp\{\beta_i\}$, as in the logistic regression model, is the odds ratio of operative mortality associated with the ith risk factor x_i, and α is a constant depending on the overall prevalence of operative mortality (Appendix 1). The odds

ratio can be calculated from the 2 × 2 frequency table (see Table 3) as

$$\exp\{\beta_i\} = \frac{a_i \times d_i}{b_i \times c_i}.$$

Although the estimation of α is important for absolute risk assessment of operative mortality, it is not needed for the determination of the model's predictive power or for discrimination purposes. In the case of prior cardiac operation, the frequencies of the 2 × 2 contingence table are $a_i = 73$, $b_i = 211$, $c_i = 537$, and $d_i = 5{,}479$ and therefore, the estimated odds ratio for this risk factor is $\exp\{\hat{\beta}_i\} = 3.53$. The estimated regression coefficient can be obtained by taking the natural logarithm of the odds ratio, that is, $\hat{\beta}_i = 1.26$.

Expression (2) shows that the Bayes' theorem model does not differ from a classic logistic regression model except in the approach used to estimate the regression coefficients. It is not difficult to show that the approach used in logistic regression to estimate the regression coefficients reduces to the approach used for the Bayes' model when the correlation between the risk factors is ignored. In the Bayes' theorem model, $\exp\{\beta_i\}$ represents the unadjusted odds ratio associated with the ith risk factor, whereas in classic logistic regression, $\exp\{\beta_i\}$ represents the same odds ratio adjusted by the remaining factors in the model. More technically, in the logistic regression model the parameters are estimated using the maximum likelihood method from a multidimensional 2^{m+1} table provided that the totals of the response variable in the 2^m table are constant. In the Bayes' theorem model the regression coefficients are also estimated using the maximum likelihood method, but from the m marginal tables generated by cross-tabulating the response variable with each of the m risk factors.

There are several advantages to writing the Bayes' theorem model as a logistic equation model. First, the form of the logistic equation (2) is more familiar to clinicians, as logistic regression models are widely used in the clinical literature. Second, the regression coefficients, βs, have meaningful clinical interpretations as odds ratios, a familiar concept in medicine. Finally, the third advantage is that the coefficients of the model can be estimated from simple frequency tables.

The Bayesian–Logit Model

Thus far we have assumed that the set of risk factors, $X = \{x_1, x_2, \ldots, x_m\}$, includes simple indicators of the presence ($x_i = 1$) or absence ($x_i = 0$) of a particular baseline characteristic. In this context, quantitative variables such as age must be categorized using arbitrary intervals. In this case, each category of age will have an associated indicator variable representing whether or not the patient's age falls into that particular category. For example, $x_1 = 1$ if age is less than 60 years old and $x_1 = 0$ if age is 60 years or older. This transformation of quantitative or continuous variables to categorical variables creates a number of problems, including the arbitrary selection of the size of the intervals and the selection of breakpoints. These problems can be solved if we allow continuous variables to enter into the Bayes' theorem model in a fashion similar to other classic regression models such as linear or logistic regression analyses. However, before we include continuous variables into the model, we must assume an underlying probability distribution for these factors. Following classic statistical methods, we assume that the underlying distribution of these variables is normal. Departures from this assumption are not expected to change significantly the results as the same results can be obtained without normality using more powerful and complex mathematical techniques. For practical purposes we assume that deaths and survivors have different risk factor means but common variance.

Consider μ_{i1}, μ_{i2}, and σ_i^2 as the population means and common variance of the continuous variable x_i for the deaths and survivors, respectively. The regression coefficient β_i associated with this continuous variable is the difference between the two means divided by the pooled variance, that is,

$$\beta_i = \frac{\mu_{i1} - \mu_{i2}}{\sigma_i^2}. \tag{3}$$

See Appendix 2 for more detail of the derivation of this result. If all indicators $\{x_1, x_2, \ldots, x_m\}$ were continuous, model (2) would still have a logistic equation form, but the constant α would become a function of the means and the variances of the different risk factors. An estimate of β_i can be obtained from the sample means and the common sample variance of the two groups. For example, we found that the average age of deaths in the learning sample was 63.8 years old and the average age among survivors in the same sample was 61.3 years old. The pooled variance was 62.6; therefore, the estimated regression coefficient for age is

$$\hat{\beta} = \frac{63.8 - 61.3}{62.6} = 0.04.$$

The Bayesian–logit model is a special case of a linear discriminant model when the risk factors are all continuous variables. In linear discriminant analysis the estimates of the model are obtained using maximum likelihood based on the multivariate normal distribution of the risk factors. The estimates are not only a function of the means and the variances, but also the correlation among all risk factors. The estimates of the regression coefficients in the Bayesian–logit model are obtained using maximum likelihood methods based on the univariate distribution of each individual risk factor, ignoring the correlation with the remaining risk factors.

Developing the Predictive Model

For model development purposes, the Bayes' theorem model and the Bayesian–logit model can use univariate screening to select the risk factors that will be incorporated into the models. To select the risk factors in the Bayes' theorem model, a χ^2 test for a 2 × 2 frequency table can be used as it is assumed that all risk factors are categorical. The Bayesian–logit model can use the same

Ann Thorac Surg
1994;57:1492–500

test for categorical risk factors, but the two sample t test for quantitative risk factors.

Among the 33 predictor variables shown in Table 2 competing to be included in the model, only those variables with $p < 0.01$ were selected. There are many reasons to be conservative in selecting the risk factors in this application. First, the sample size is large enough to select clinically meaningful factors, and second, we expect that many of the selected risk factors contain duplicative information to predict operative mortality. Alternative methods, such as checking the correlation matrix of the risk factors, can reduce the number of variables in the model.

The logistic regression and discriminant analysis model used to compare the Bayesian–logit model were developed using stepwise procedures and including the 33 risk factors shown in Table 2. The predictive power of these models was assessed by using the c-index [6, 7, 9] in the learning and testing sample.

Results

The Bayes' Theorem Model

Using the learning sample of 6,317 CABG procedures performed from April 1987 to September 1992, 19 risk factors were found to be significant at the level of $p < 0.01$ using univariate screening. These significant risk factors are: congestive heart failure, current digoxin use, current diuretic use, preoperative use of intravenous nitroglycerine, resting ST-segment depression, preoperative intraaortic balloon pump, recent myocardial infarction, peripheral vascular disease, cerebral vascular disease, unstable angina, electrocardiographically determined left ventricular hypertrophy, cardiomegaly, old myocardial infarction, pulmonary rales, prior cardiac operation, chronic obstructive pulmonary disease, priority of operation, New York Heart Association (NYHA) functional classification, and age. Age was categorized using the following intervals: younger than 50, 51 to 60, 61 to 70, and older than 70 years old. Indicator variables to represent the different categories of NYHA functional classification (II, III, and IV) and priority of operation (urgent and emergent) were also created. A total of 24 indicator variables were used to create the Bayes' theorem model by estimating the conditional probability of having the risk factor present given the outcome of the procedure as described by Edwards and colleagues [3, 4].

The Bayesian–Logit Model

The same 19 risk factors found to be significant using univariate screening and included in the Bayes' theorem model were included in the Bayesian–logit model. Sixteen of these selected risk factors were indicator variables, representing the presence or absence of a particular risk factor. However, three of these risk factors are ordinal variables: NYHA functional class, priority of operation, and age. By plotting mortality rates for the categories of each of these ordinal risk factors, we conclude that operative mortality increases linearly when the ordinal categories also increase. The results suggest that these three risk factors can be included as continuous variables in the Bayesian–logit model. By doing this, we reduce from 24 the number of indicator variables in the Bayes' theorem model to 19 variables in the Bayesian–logit model.

Table 4. Frequencies of the Presence or Absence of Risk Factors Among Deaths and Survivors, Respectively, and Unadjusted Odds Ratios for Each of the Categorical Risk Factors

Risk Factor	Deaths		Survivors		$e^{\beta} = \frac{a \times d}{b \times c}$
	Present a	Absent b	Present c	Absent d	
CHF	80	198	836	5058	2.44
CM	71	204	797	4988	2.18
COPD	96	181	1551	4278	1.46
CURDIG	53	212	691	5008	1.81
CURDIUR	104	164	1412	4287	1.93
CVD	62	207	787	4894	1.86
ECGLVH	50	218	694	4870	1.61
IVNTG	103	167	1053	4640	2.72
OLDMI	177	95	3126	2612	1.56
PRIORHS	73	211	537	5479	3.53
PROPIABP	28	256	181	5732	3.46
PULMR	50	222	367	5303	3.25
PVD	97	174	1294	4402	1.90
RECMI	80	199	1037	4814	1.87
RESTANG	189	85	3207	2433	1.69
RESTSTD	81	188	968	4619	2.06

See Table 2 for abbreviations.

Table 4 shows the frequencies of the 2 × 2 contingency table (see Table 3) used to derive the regression coefficients and the odds ratios for the indicator variables. Table 5 shows the mean and variance of the continuous variables used to derive the regression coefficients and the odds ratios.

Logistic Regression Model

In developing a logistic regression model to predict operative mortality we included the 33 risk factors shown in Table 2 in a stepwise procedure without any clinical judgment or forcing variables into the model. Only seven risk factors were found to be significant at the level of $p < 0.01$ as independent predictors of mortality using the learning sample. These risk factors are as follows: prior cardiac operation, priority of operation, pulmonary rales,

Table 5. Mean of Risk Factors and Common Variables Among Deaths and Survivors, Respectively, and Unadjusted Odds Ratios for Each Quantitative Risk Factor

Risk Factor	Deaths Mean μ_1	Survivors Mean μ_2	Pooled Variance σ^2	$e^{\beta} = \exp\left\{\frac{\mu_1 - \mu_2}{\sigma^2}\right\}$
AGE	63.8	61.3	62.68	1.04
NYHAFC	3.4	3.0	0.65	1.94
PRIORITY	1.7	1.4	0.38	2.46

See Table 2 for abbreviations.

Table 6. Regression Coefficients (β), Odds Ratios, and c-Indices of the Logistic Regression Model, Discriminant Analysis, and Bayesian–Logit Model Using the Same Risk Factors Found in the Stepwise Logistic Regression Procedure

	Logistic Regression Model		Discriminant Analysis		Bayesian–Logit Model	
Factor	β	Odds Ratio	β	Odds Ratio	β	Odds Ratio
PRIORHS	1.142	3.13	1.840	6.30	1.261	3.53
PRIORITY	0.450	1.57	0.602	1.83	0.898	2.45
PULMR	0.723	2.06	1.363	3.91	1.180	3.25
NYHAFC	0.453	1.57	0.379	1.46	0.663	1.94
AGE	0.034	1.03	0.029	1.03	0.040	1.04
PVD	0.359	1.43	0.418	1.52	0.640	1.90
CVD	0.210	1.23	0.304	1.36	0.622	1.86
c-Index						
Learning sample		0.740		0.724		0.722
Testing sample		0.710		0.705		0.702

See Table 2 for abbreviations.

New York Heart Association functional class, age, peripheral vascular disease, and cerebral vascular disease.

The Predictive Power

The *c*-index was calculated for the Bayes' theorem model, the Bayesian–logit model, and the logistic regression model using the learning and the testing sample. It is well known that the learning sample consistently overestimates the *c*-index as the model is evaluated with the same data from which it was developed. The *c*-index for the Bayes' theorem model, the Bayesian–logit model, and the logistic regression model were 0.713, 0.718, and 0.740 in the learning sample, respectively, and 0.695, 0.701, and 0.710 in the testing sample, respectively. Although the *c*-index differences are statistically significant, for both clinical decision making and risk adjustment these *c*-indices seem to be very similar.

The Discriminant Analysis Model

Discriminant analysis is a similar modeling technique to logistic regression that allows the classification of individuals coming from two different groups or subpopulations, which in this context are survivors and deaths. The classic assumption in discriminant analysis is that the distribution of the risk factors is normal; however, this assumption can be relaxed. The major difference with respect to logistic regression is in the focus of the problem. In discriminant analysis the outcome variables are the risk factors and the independent variable is the group. The estimate of the regression coefficients has a closed form solution and can be found using multiple linear regression programs.

The Bayesian–Logit Versus the Classic Regression Models

One of the features of this new Bayesian–logit model is its similarity to logistic regression and discriminant analysis in terms of both the model equation, and the regression coefficients and the odds ratios. Therefore, we developed a Bayesian–logit model and a discriminant analysis model with the same seven risk factors found to be significant in the stepwise logistic regression. The Bayesian–logit model has consistently higher odds ratios compared with logistic regression, and all except the odds ratios of two risk factors are higher compared with discriminant analysis (Table 6). These results are expected as the seven risk factors included in the model have pairwise positive correlations. These pairwise correlations adjust the values of regression coefficients in logistic regression and discriminant analysis. In the Bayesian–logit model these correlations are ignored and the regression coefficients are not adjusted for the presence of other risk factors.

The predictive power of these three models is very similar, particularly for the testing sample (see Table 6). Note that the Bayesian–logit model has a better *c*-index with the seven risk factors found in the stepwise logistic regression than with the 19 risk factors found to be significant using univariate screening. This is the result of using redundant information, which can be avoided by including the set of risk factors with the highest predictive power base on univariate analysis. Theoretically, the number of risk factors that should be included in the model can be determined by observing the *c*-index of the testing sample after sequentially adding a new variable.

Table 7 shows the calibration of the Bayesian–logit model by illustrating the expected and observed mortality in the different deciles of risk groups. The calibration was done using both the learning and testing sample, showing no major discrepancies. This result shows that the model correctly estimates the risk of mortality in all risk groups except in the sicker group where the model tends to overestimate the true risk of the patient. This result is mainly because the regression coefficients of the Bayesian–logit model, as well as the Bayes' model, are not adjusted by the presence of other risk factors.

Comment

The Importance of Predictive Models in Risk Adjustment

Unadjusted mortality rates have been used as a measure of quality of care to compare the performance of hospitals

Ann Thorac Surg
1994;57:1492–500

Table 7. Expected Versus Observed Mortality for the Bayesian–Logit Model

Deciles of Risk	Learning Sample		Testing Sample	
	Expected	Observed	Expected	Observed
1	0.46%	1.60%	0.46%	1.08%
2	0.76%	1.10%	0.75%	2.11%
3	1.08%	2.28%	1.07%	1.67%
4	1.41%	2.04%	1.38%	1.66%
5	1.92%	2.01%	1.87%	3.71%
6	2.59%	2.58%	2.57%	3.52%
7	3.64%	4.37%	3.59%	5.17%
8	5.38%	3.80%	5.51%	4.91%
9	9.40%	10.62%	9.22%	7.64%
10	26.31%	14.52%	25.16%	14.13%

for nearly a century. However, raw mortality does not account for differences in patient mix across hospitals or providers. Therefore, the use of unadjusted mortality rates may produce a significant bias in comparing quality of care. Recently, there has been an increasing interest in using risk-adjusted mortality indicators by developing predictive models of operative mortality. The Department of Veterans' Affairs and the state of New York have taken the initiative of collecting hospital mortality rates with the purpose of improving the quality of patient care [1, 2]. These initiatives have captured the attention of the media and it is expected to increase the attention of the public and the medical community with the new trends in health care reforms.

Clinical researchers and biostatisticians have proposed different methodologic approaches to develop predictive models. Without doubt, classic logistic regression is one of the most common statistical techniques used to predict adverse outcomes. Although Bayes' theorem model is not as commonly used as logistic regression, it has comparable predictive power to other well-known classic statistical approaches. One of the advantages of logistic regression with respect to the Bayes' theorem model is the comparative simplicity of the model equation and the fact that results can be easily interpreted by clinicians as odds ratios. The Bayes' theorem model has a significantly more complicated model equation and it is not clear how to interpret clinically the model results with respect to individual predictor variables.

Reformulating the Bayes' Theorem Model

The Bayes' theorem model could become a more attractive model for predicting adverse outcomes by reformulating it using the logistic equation (2), because the contribution of individual covariates can be interpreted from the results. However, it is important to emphasize that by reformulating the Bayes' theorem model we have learned that it is a constrained version of logistic regression. Although it is easier to compute, it can never exceed logistic regression in terms of predictive capability. This constraint is imposed by the assumption of conditional independence, which means that the model ignores the interrisk factor correlations.

The Bayesian–Logit Model

The Bayesian–logit model allows clinicians to include quantitative risk factors linearly in the model instead of breaking the variables into arbitrary intervals. This feature is particularly important in predicting the outcome of patients undergoing cardiac operation, as the effect of age on operative mortality has been shown to be continuous rather than constant by intervals [1, 7]. By using the Bayesian–logit model we can also compare results with other statistical techniques such as logistic regression and discriminant analysis. Another advantage of using the Bayesian–logit model over the Bayes' theorem model is the fact that the Bayesian–logit model is a constrained model of a mixture of logistic regression and discriminant analysis, which is not widely used as it is significantly more complex to compute. Therefore, the model is not a restrictive version of any other commonly used model as is the case with the Bayes' theorem model, which we show to be a subset of logistic regression. It is also important to mention that both the Bayes' theorem model and the Bayesian–logit model are different applications of the same Bayes' theorem, also known as the Bayes' rule, which was named after Thomas Bayes (1702–1761).

To compare performance of the different models in predicting outcomes we have reported the *c*-indices. On the basis of the results of this analysis, the predictive power of the Bayes' theorem model, the Bayesian–logit model, and the logistic regression model is comparable, and any differences are unlikely to be clinically significant. However, these results do confirm what we obtained using analytical derivation. The advantage of these two Bayesian models with respect to logistic regression is the simplicity of parameter estimation. This advantage, however, can be questionable with today's powerful computer hardware and widely available software that are able perform the complex numerical analysis required for logistic regression. In the immediate future, we must explore other computer-intensive models, such as neural networks, that may produce even better predictive results.

Appendix 1

The conditional independence Bayes' theorem model can be written as

$$P\{death|X\} = \frac{\prod_{i=1}^{m} P\{x_i|death\} \times P\{death\}}{\prod_{i=1}^{m} P\{x_i|survival\} \times P\{survival\} + \prod_{i=1}^{m} P\{x_i|death\} \times P\{death\}},$$

where $P\{x_i|death\}$ and $P\{x_i|survival\}$ represent the conditional probability that the risk factor x_i is present ($x_i = 1$) or absent ($x_i = 0$) given one of the two possible operative outcomes: death and

survival, respectively. If we divide both the numerator and the denominator by

$$\prod_{i=1}^{m} P\{x_i|survival\} \times P\{survival\},$$

the first term in the denominator, the probability of operative death given the risk factors X is

$$P\{death|X\} = \frac{f(X)}{1 + f(X)}, \quad (4)$$

where

$$f(X) = \prod_{i=1}^{m} \frac{P\{x_i|death\}}{P\{x_i|survival\}} \times \frac{P\{death\}}{P\{survival\}}. \quad (5)$$

For categorical risk factors, the distribution of the risk factors given the two outcomes can be written as $P\{x_i|death\} = p_{i1}^{x_i} \times (1 - p_{i1})^{1-x_i}$ and $P\{x_i|survival\} = p_{i2}^{x_i} \times (1 - p_{i2})^{1-x_i}$, respectively, where p_{i1} and p_{i2} represent the conditional probability that the risk factor x_i is present given the outcomes are death and survival, respectively. If we replace these expressions back into equation (5), $f(X)$ reduces to

$$f(x) = \prod_{i=1}^{m}\left(\frac{p_{i1} \times (1 - p_{i2})}{p_{i2} \times (1 - p_{i1})}\right)^{x_i} \times \prod_{i=1}^{m}\frac{1 - p_{i1}}{1 - p_{i2}} \times \frac{P\{death\}}{P\{survival\}},$$

$$= \exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m\}$$

where

$$\alpha = \sum_{i=1}^{m}\log\left(\frac{1 - p_{i1}}{1 - p_{i2}}\right) + \log\left(\frac{P\{death\}}{P\{survival\}}\right)$$

and

$$\beta_i = \log\left(\frac{p_{i1} \times (1 - p_{i2})}{p_{i2} \times (1 - p_{i1})}\right).$$

Replacing $f(x) = \exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m\}$ in equation (4) we prove that the conditional independence Bayes' theorem model can be reformulated as equation (2).

Appendix 2

For continuous risk factors, we assume that the distribution of the risk factor is normal. The probability function of the risk factor x_i given the outcome of the operation is death is defined as

$$P\{x_i|death\} = c_i \times \exp\left\{-\frac{(x_i - \mu_{i1})^2}{2\sigma_i^2}\right\},$$

where c_i is a constant that only depends on x_i but not on the operative outcome. The probability function $P\{x_i|survival\}$ has the same form except that the constant μ_{i1} is replaced by μ_{i2}. The constants μ_{i1}, μ_{i2}, and σ_i^2 represent the population means and common variance of the continuous risk factor x_i for the deaths and survivors, respectively.

The ratio between the two conditional probability functions as shown in equation (5) in Appendix 1 is

$$\frac{P\{x_i|death\}}{P\{x_i|survival\}} = \exp\left\{-\frac{(x_i - \mu_{i1})^2}{2\sigma_i^2} + \frac{(x_1 - \mu_{i2})^2}{2\sigma_i^2}\right\}$$

$$= \exp\left\{\frac{(\mu_{i1} - \mu_{i2})}{\sigma_i^2} \times x_i - \frac{(\mu_{i1} - \mu_{i2})}{\sigma_i^2} \times \frac{(\mu_{i1} + \mu_{i2})}{2}\right\}$$

and therefore, the regression coefficient β_i is

$$\beta_i = \frac{(\mu_{i1} - \mu_{i2})}{\sigma_i^2},$$

and where

$$-\frac{(\mu_{i1} - \mu_{i2})}{\sigma_i^2} \times \frac{(\mu_{i1} + \mu_{i2})}{2}$$

is added to the constant α (see Appendix 1).

This work was supported by the Health Services Research and Development Service and the Office of Quality Management of the Department of Veterans Affairs.

References

1. Hammermeister KE, Johnson R, Marshall G, Grover FL. Continuous assessment and improvement in quality of care: a model from the Department of Veterans Affairs cardiac surgery. Ann Surg 1994;219:281–9.
2. Hannan EL, Kilburn H, O'Donnell JF, Lukacik G, Shields EP. Adult open heart surgery in New York State: an analysis of risk factors and hospital mortality rates. JAMA 1990;264: 2768–74.
3. Edwards FH, Graeber GM. The theorem of Bayes as a clinical research tool. Surgery 1987;165:127–9.
4. Edwards FH, Albus RA, Zajtchuk R, et al. Use of a Bayesian statistical model for risk assessment in coronary artery surgery. Ann Thorac Surg 1988;45:437–40.
5. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation 1989;79(suppl I):3–12.
6. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Statistics Med 1984;3:143–52.
7. Marshall G, Grover FL, Henderson WG, Hammermeister KE. Assessment of predictive models for binary outcomes: an empirical approach using operative death from cardiac surgery. Statistics Med 1994 (in press).
8. Cook EF, Goldman L. Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. J Chronic Dis 1984;37:721–31.
9. Harrell FE Jr, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems and suggested solutions. Cancer Treat Rep 1985;69:1071–7.
10. Kennedy JW, Kaiser GC, Fisher LD, et al. Multivariate discriminant analysis of the clinical and angiographic predictors of operative mortality from the Collaborative Study in Coronary Artery Surgery (CASS). J Thorac Cardiovasc Surg 1980;80:876–87.
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
12. Grover FL, Hammermeister KE, Burchfiel C, and Cardiac Surgeons of the Department of Veterans Affairs. Initial report of the Veterans Administration preoperative risk assessment study for cardiac surgery. Ann Thorac Surg 1990;50:12–28.
13. Hsieh FY. Sample size for logistic regression. Statistics Med 1989;8:795–802.

Ann Thorac Surg
1994;57:1492–500

DISCUSSION

DR RICHARD E. CLARK (Pittsburgh, PA): I welcome this type of investigative work, albeit in mathematics, exactly as we would any innovative experiment in the animal laboratory. It is very important to explore how we can treat our data sets in a predictive way that will stand the test of time.

What we have seen here today is a marriage, in part, between the logistic and the Bayesian approaches to building predictive models of perioperative mortality. Doctor Fisher, a renowned American biostatistician, has said that all statistics come from the same trunk, which gives forth several major branches. Doctor Marshall has tried to blend two major branches.

The reason we use the Bayesian system in the STS National Database is that it can accommodate a very large number of variables. In your model, for example, you only had seven variables and then you tried it with 18. The STS risk stratification system in use in more than 650 hospitals uses 23 and it can accommodate more. In addition, the Bayesian system is forgiving for incomplete data for some variables in contrast to the logistic regression method. I congratulate you on the innovative effort.

DR MARSHALL: Thank you, Dr Clark, for your comments. I think one of the most important conclusions is that all these models have the same predictive power, and, as Dr Clark mentioned, one of the most important advantages of the Bayesian model is that if you have missing values, if you have missing information on a patient, you still can compute the predictive mortality, whereas in logistic regression you have no information about the predictive mortality, unless you impute a value.

The Third International Conference on Circulatory Support Devices for Severe Cardiac Failure

Lawrence Convention Center
Pittsburgh, Pennsylvania, October 28–30, 1994

Circulatory Support 1994 will be a comprehensive multidisciplinary meeting focusing on clinical applications for the entire spectrum of circulatory support devices. Surgeons, cardiologists, anesthesiologists, intensivists, perfusionists, engineers, and operating and intensive care nurses are encouraged to attend as individuals or as a team.

The meeting is sponsored by The Society of Thoracic Surgeons under the direction of its Ad Hoc Committee on Circulatory Support and Thoracic Transplantation. An additional, **optional** program will be scheduled for Friday, October 28th, that will feature a site visit to either of two regional medical centers. Allegheny General Hospital will be visited in the morning to review their cardiomyoplasty program. An afternoon program will feature a site visit to the University of Pittsburgh to review their mechanical circulatory support program. Space will be limited for these special programs.

The general program will begin on Saturday, October 29, at 8:00 AM and will conclude on Sunday, October 30, at 2:45 PM. Session topics will review clinical and experimental systems that focus on pulsatile ventricular assist, total artificial heart, nonpulsatile ventricular assist, ECMO for cardiopulmonary support, pediatric blood pump development, and pneumatic artificial heart segments; also clinical problems sessions will be presented that involve univentricular versus biventricular support, infection, thromboembolism and bleeding, patient selection, and extended bridge to permanent implant. Other sessions will cover dynamic myoplasty, a video review of selected devices, and a panel session dealing with topical regulatory issues.

The meeting will include a poster session and commercial exhibits including manufacturers and products that will be reviewed during the meeting.

Further details on this meeting will be mailed to all members of The Society in the future and may be requested from the Society Headquarters at 401 N Michigan Ave, Chicago, IL 60611.