GTS

Internal validation of risk models in lung resection surgery: Bootstrap versus training-and-test sampling

Alessandro Brunelli, MD,[a] and Gaetano Rocco, MD[b,*]

Objective: The objective of the present analysis was to compare the performance of a lung resection mortality model developed by means of logistic regression and bootstrap analysis with that of multiple mortality models developed by using the traditional training-and-test method from the same dataset.

Methods: Eleven mortality models (1 developed by means of logistic regression and bootstrap validation and the other 10 developed by means of the traditional training-and-test random splitting of the dataset) were generated by the data of unit A (571 patients submitted to major lung resection). The performances of each of the 11 mortality models were then evaluated by assessing the distribution of the respective c-statistics in 1000 bootstrap samples derived from unit B (224 patients).

Results: The first model (logistic regression and bootstrap analysis) had good discrimination among the 1000 bootstrap external samples (c-statistics >0.7 in 80% of samples and >0.8 in 38% of samples). Among the 10 training-and-test models, only one model had a similar performance, whereas the others had a poorer discrimination.

Conclusions: The traditional training-and-test method for risk model building proved to be unreliable across multiple external populations and was generally inferior to bootstrap analysis for variable selection in regression analysis. Therefore the use of bootstrap analysis must be recommended for every future model-building process.

Risk stratification and outcome analysis can be used to judge effectiveness of care and to assist providers in quality improvement activities, such as cost containment, patients' education, effectiveness of care studies, and improvement in provider practices.

Regression analyses are the analytic techniques most commonly used for risk modeling. However, the resultant models are useful only if they reliably predict outcomes for patients by determining significant risk factors associated with the outcome of interest. A problem might arise from this dependence on risk factor analysis. Different investigators evaluating the same predictors through regression analysis might obtain heterogeneous results because of methodologic discrepancies and inadvertent biases introduced in the statistical elaboration.[1]

In the early 1980s, computer-intensive computational techniques, termed bootstrap methods, were popularized.[2-6] Bootstrap analysis is a simulation method for statistical inference, which, if applied to regression analysis, can provide variables that have a high degree of reproducibility and reliability as independent risk factors of the given outcome.

In fact, the predictive validity of a model can be assessed not only in one randomly split set of patients, as in the traditional training-and-test method, but also in perhaps hundreds or, typically, 1000 new different samples of the same number of patients as the original database obtained by means of resampling with replacement.

We hypothesized that the traditional training-and-test method for model building might generate models that are heavily biased by the characteristics of the patients

From the Unit of Thoracic Surgery,[a] "Umberto I" Regional Hospital, Ancona, Italy, and the Division of Thoracic Surgery,[b] Sheffield Teaching Hospital, Sheffield, United Kingdom.

Received for publication Nov 29, 2005; revisions received Jan 3, 2006; accepted for publication Feb 6, 2006.

Address for reprints: Alessandro Brunelli, MD, Via S. Margherita 23, Ancona 60129, Italy (E-mail: alexit_2000@yahoo.com).

*Dr Rocco is currently affiliated with the Division of Thoracic Surgery, National Cancer Institute, Pascale Foundation, Naples, Italy.

J Thorac Cardiovasc Surg 2006;131:1243-47

0022-5223/$32.00

doi:10.1016/j.jtcvs.2006.02.002

GTS

Abbreviations and Acronyms

CAD = coronary artery disease
DLCO = carbon monoxide lung diffusion capacity
FEV_1 = forced expiratory volume in 1 second
ppoDLCO = predicted postoperative carbon monoxide lung diffusion capacity
$ppoFEV_1$ = predicted postoperative forced expiratory volume in 1 second

who are sampled to derive and test them. The external performance of these types of models could be extremely variable and therefore totally unreliable. On the other hand, by using the entire dataset for model construction and bootstrap analysis for validation and variable selection, a more robust and stable model would be obtained, which can be more reliably applied to external patients.

Therefore the objective of the present study was to compare the performance of a mortality model adjusted for the covariates contributing to the risk of death developed from the entire dataset of patients submitted to major lung resection in one single unit and validated by using the bootstrap procedure with that of multiple mortality models developed by using the training-and-test method from the same dataset. To this purpose, each model was assessed in 1000 external bootstrap samples derived from another set of patients operated on in another unit during the same period.

Patients and Methods

A population of 571 patients undergoing major lung resection (479 lobectomies-bilobectomies and 92 pneumonectomies) from January 2000 through December 2004 in a thoracic surgery unit (unit A) was used to develop mortality models. Mortality was considered as that occurring within 30 days from the operation or over a longer period if the patient was still hospitalized.

Two different model-building approaches were used. The first method (model A) consisted of using the entire dataset for model construction. The following variables were initially evaluated for possible association with postoperative mortality: age, body mass index (in kilograms per square meter), type of operation (lobectomy vs pneumonectomy), type of disease (benign vs malignant), neoadjuvant chemotherapy, presence of coronary artery disease (CAD), forced expiratory volume in 1 second (FEV_1), carbon monoxide lung diffusion capacity (DLCO), predicted postoperative FEV_1 ($ppoFEV_1$; calculated by using the formula),

$$[Preoperative\ FEV_1 / Number\ of\ preoperative\ functioning\ segments] \times Number\ of\ postoperative\ functioning\ segments$$

and predicted postoperative DLCO (ppoDLCO, calculated by using the formula).

$$[Preoperative\ DLCO / Number\ of\ preoperative\ functioning\ segments] \times Number\ of\ postoperative\ functioning\ segments$$

The number of functioning segments was estimated by means of computed tomography, bronchoscopy, and quantitative perfusion lung scanning. Pulmonary function tests were performed according to the American Thoracic Society criteria. DLCO was measured by using the single-breath method. Results of spirometry were collected after bronchodilator administration and were expressed as a percentage of predicted value for age, sex, and height according to the European Community for Steel and Coal prediction equations.[7]

Survivors and nonsurvivors were initially compared by means of univariate analyses performed with the unpaired Student *t* test or the Mann-Whitney test for numeric variables and the χ^2 test or the Fisher exact test for categoric variables. Multicollinearity among variables was obviated by using only one variable (selected by means of bootstrap analysis) in a set of variables with a correlation coefficient greater than 0.5 in the regression analysis. Variables with a *P* value of less than .1 at univariate analysis were used as independent variables in a stepwise logistic regression analysis (dependent variable of mortality). A *P* value of less than .1 was selected for variable retention in the final regression model. The model was then validated by means of bootstrap analysis. In the bootstrap procedure 1000 samples of 571 patients were sampled with replacement. Stepwise logistic regression analysis was applied to every bootstrap sample. The stability of the final model was assessed by comparing the frequency of occurrence of the variables of the final model in the bootstrap samples. If the predictors occurred in more than 50% of the bootstrap models, they were judged to be reliable and were retained in the final model.[8] Unreliable variables, if present, were removed from the final model.

The second method (model B) consisted of the traditional training-and-test splitting method. The dataset was randomly split into 2 sets of patients. The first set (60% of the database) was used to develop the model. The same variables used in the first method were initially evaluated for possible association with postoperative mortality. Screening for univariate associations and multicollinearity was performed in the same way in the second method as described for the first method. Variables with a *P* value of less than .1 at univariate analysis were used as independent variables in a stepwise logistic regression analysis (dependent variable of mortality). A *P* value of less than .1 was selected for variable retention in the final regression model, for which the calibration and discrimination was assessed with the remaining 40% of patients (test set) by using the Hosmer-Lemeshow goodness-of-fit statistic and by using the c-statistics or area under the receiver operating characteristic curve.[9-12] The proportion of patients sampled in the training and test samples (60% and 40%, respectively), was selected in accordance with recently published analyses on risk modeling in lung surgery.[13]

Ten models were developed by repeating the training-and-test method 10 times. Therefore a total of 11 mortality models were obtained. The performance of each of these models was assessed 1000 times, each time using a bootstrap sampling of 224 patients drawn with replacement from the database of unit B, by evaluating the distribution of the c-statistics in these samples.[9-12]

Prospective, electronic, quality-controlled, clinical databases at the 2 participating centers were used for the analysis of data. The study was approved by the local institutional review boards, and informed consent concerning prospective data collection was obtained from all patients. The authors had access to the primary

TABLE 1. Characteristics of the patients in the 2 units used to derive (unit A) and externally validate (unit B) the 11 mortality models

Variables	Unit A (571 patients)	Unit B (224 patients)	*P* value
Age	67 (9.6)	61.3 (12.2)	<.0001*
Male sex, n (%)	450 (79%)	144 (64%)	<.0001†
BMI, kg/m^2	26.1 (4.2)	25.6 (4.5)	.12*
FEV_1, %	85.5 (19.5)	84 (18.2)	.3*
DLCO, %	76.6 (18.7)	81.4 (18.2)	.004*
ppoFEV_1, %	67.4 (17.1)	62.4 (17.8)	.0004*
ppoDLCO, %	60.4 (16.6)	60.6 (16.6)	.6*
Pneumonectomy, n (%)	92 (16%)	42 (19%)	.4†
Coronary artery disease, n (%)	70 (12%)	34 (15%)	.3†
Malignant disease, n (%)	550 (96%)	185 (83%)	<.0001†
Neoadjuvant chemotherapy, n (%)	79 (14%)	25 (11%)	.3†
Mortality, n (%)	25 (4.4%)	10 (4.5%)	.9†

Results are expressed as means ± standard deviations unless otherwise specified. *BMI,* Body mass index; *FEV_1,* forced expiratory volume in 1 second; *DLCO,* carbon monoxide lung diffusion capacity; *ppoFEV_1,* predicted postoperative forced expiratory volume in 1 second; *ppoDLCO,* predicted postoperative carbon monoxide lung diffusion capacity. *Mann-Whitney test. †χ^2 test.

data, directed the analyses, and made all decisions pertaining to the article and its submission for publication.

All tests were 2-tailed and were entirely performed with the Stata 8.2 statistical software (Stata Corp, College Station, Tex).

Results

The characteristics of the patients in the 2 units analyzed in this study are shown in Table 1. The 2 populations differed in age, sex, type of disease, and pulmonary function parameters.

The first statistical method used to develop and validate the mortality model (by using bootstrap analysis) yielded the following regression equations:

$$Model\ A: InR/1 + InR = -6.3 + 0.09 * Age\ (Bootstrap\ frequency,\ 86\%) - 0.048 * ppoFEV_1\ (Bootstrap\ frequency,\ 71\%)$$

(Hosmer-Lemeshow statistics, 7.2; $P = .5$; c-statistic, 0.76). The expression InR/1 + InR represents the probability of dying because in the logistic regression equation the logarithm of the odds of the outcome (termed the logit or log odds) is used as the dependent variable.

The second statistical method (training and test) repeated 10 times yielded the following different mortality models:

- *Model B*1: $InR/1 + InR = -0.66 - 0.05 * ppoDLCO + 1.04 * CAD$ (test set: Hosmer-Lemeshow statistics, 5.9; $P = .7$; c-statistic, 0.67).

TABLE 2. Distribution of different variables and their coefficients in the 11 regression mortality models

Models	Intercept	Age	ppoFEV_1, %	ppoDLCO, %	Pneumonectomy	CAD
A	−6.3	0.09	−0.048			
B1	−0.66			−0.05		1.04
B2	−3.355	0.056		−0.066		
B3	−0.55		−0.06			
B4	−10.1	0.1			1.64	
B5	−8.26	0.108	−0.05			1.79
B6	−10.2	0.096				1.32
B7	−11.4	0.113			1.86	
B8	−10.9	0.108			1.45	
B9	−10.7	0.147	−0.049			1.11
B10	0.64		−0.067			1.62

ppoFEV_1, Predicted postoperative forced expiratory volume in 1 second; *ppoDLCO,* predicted postoperative carbon monoxide lung diffusion capacity; *CAD,* coronary artery disease.

- *Model B*2: $InR/1 + InR = -3.355 + 0.056 * Age - 0.066 * ppoDLCO$ (test set: Hosmer-Lemeshow statistics, 18; $P = .02$; c-statistic, 0.57).
- *Model B*3: $InR/1 + InR = -0.55 - 0.06 * ppoFEV_1$ (test set: Hosmer-Lemeshow statistics, 14.2; $P = .07$; c-statistic, 0.61).
- *Model B*4: $InR/1 + InR = -10.1 + 0.1 * Age + 1.64 * Pneumonectomy$ (test set: Hosmer-Lemeshow statistics, 8; $P = .4$; c-statistic, 0.65).
- *Model B*5: $InR/1 + InR = -8.26 + 1.79 * CAD + 0.018 * Age - 0.05 * ppoFEV_1$ (test set: Hosmer-Lemeshow statistics, 25.5; $P = .001$; c-statistic, 0.67).
- *Model B*6: $InR/1 + InR = -10.2 + 1.32 * CAD + 0.096 * Age$ (test set: Hosmer-Lemeshow statistics, 21; $P = .07$; c-statistic, 0.63).
- *Model B*7: $InR/1 + InR = -11.4 + 0.113 * Age + 1.86 * Pneumonectomy$ (test set: Hosmer-Lemeshow statistics, 21.6; $P = .06$; c-statistic, 0.64).
- *Model B*8: $InR/1 + InR = -10.9 + 0.108 * Age + 1.45 * Pneumonectomy$ (test set: Hosmer-Lemeshow statistics, 13.7; $P = .09$; c-statistic, 0.64).
- *Model B*9: $InR/1 + InR = -10.7 + 0.147 * Age - 0.049 * ppoFEV_1 + 1.11 * CAD$ (test set: Hosmer-Lemeshow statistics, 23; $P = .03$; c-statistic, 0.68).
- *Model B*10: $InR/1 + InR = 0.64 + 1.62 * CAD - 0.067 * ppoFEV_1$ (test set: Hosmer-Lemeshow statistics, 23.8; $P = .02$; c-statistics, 0.59).

The distribution of the different predictors in the 11 mortality models is shown in Table 2.

Table 3 and Figure 1 show the distribution of the c-statistics of each model in 1000 bootstrap samples derived from an external series of patients operated on in another unit (unit B) during the same period. Model A (regression plus bootstrap) had good discrimination among the bootstrap external samples,

GTS

TABLE 3. Distribution of ROC areas of each model in 1000 bootstrap samples derived from unit B

Models	% ROC area >0.5	% ROC area >0.7	% ROC area >0.8	95% CL
Model A	100	80	38	0.58-0.94
Model B1	80	11	0	0.4-0.79
Model B2	100	92	38	0.65-0.89
Model B3	89	31	9	0.37-0.86
Model B4	99	59	18	0.62-0.95
Model B5	100	51	7	0.56-0.86
Model B6	85	11	1	0.53-0.87
Model B7	99	59	15	0.62-0.95
Model B8	99	62	17	0.63-0.95
Model B9	99	64	17	0.58-0.9
Model B10	75	16	0	0.32-0.79

Model A is derived from the entire dataset of unit A and validated by means of bootstrap bagging, and models B1 through B10 are derived from a randomly selected training set of unit A (60% of patients) and validated in a test set of unit A (the remaining 40% of patients). *ROC,* Receiver operating characteristic; *CL,* confidence limit.

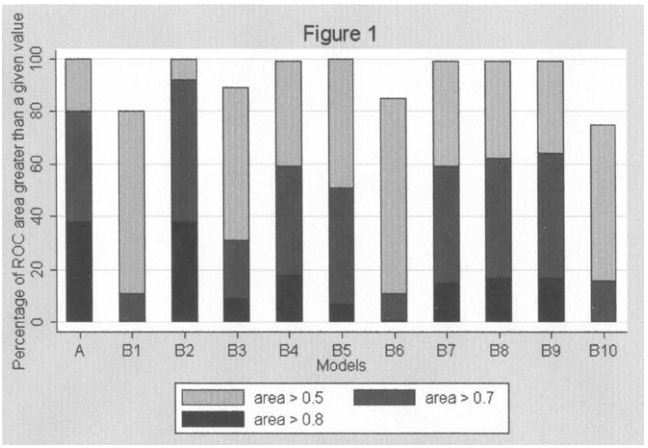

Figure 1. Distribution of the receiver operating characteristic *(ROC)* areas of each model derived from unit A in 1000 bootstrap samples drawn from unit B.

with a c-statistic greater than 0.7 in 80% of the samples and greater than 0.8 in 38% of the samples. Among the 10 training-and-test models, only model B2 had a similar performance, whereas the other models had a much poorer discrimination.

Discussion

Risk modeling, which is the base of risk stratification, is essential for quality improvement activities in managed care systems. It serves multiple purposes, such as provider profiling, cost containment, patient counseling about the operative risk, planning of postoperative advanced care management, and construction of efficacy studies. The importance of risk models is such that their reliability and reproducibility in populations other than the one from which they were derived must be absolutely proved. Otherwise, any clinical and administrative decision based on these models can be critically flawed by their instability in external populations.

We hypothesized that the traditional training-and-test method for model building, consisting of a random splitting of the database into a derivation set from which to construct the model and a test set in which to assess its calibration and discrimination, might be subject to sampling noise. To this purpose, we repeated 10 training-and-test sessions, producing 10 corresponding mortality models. Seventy percent of these models included different combinations of variables. The performance of each of these models was assessed 1000 times, each time using a bootstrap sampling of 224 patients drawn from the dataset of another unit. The distribution of the c-statistics was extremely variable from one model to another, and in general, their performances in external samples were only modest. The development of risk-adjusted models by the method of training and testing appears therefore completely unreliable.

Bootstrap analysis was recently proposed as a breakthrough method for internal validation of surgical regression models.[8,14] The main advantage of this technique is that the entire dataset can be used for model building, which would yield more robust models, especially in moderate-size databases and for rare outcomes (eg, mortality after major lung resection).[15] Furthermore, the predictive validity of the model can be assessed not only in one randomly split set of patients but also typically in 1000 new different samples of the same number of patients as the original database obtained by means of resampling with replacement. By using this method, we constructed and validated a mortality model, which, when assessed in 1000 bootstrap samples drawn from another unit, performed better than the majority of the models developed by using the training-and-test method. This shows that the bootstrap procedure can yield stable models across multiple populations, warranting its use as a standard instrument in future model-building analyses. Yet a search of PubMed performed over the last 5 years yielded, at the time of this writing, only 16 surgical articles (published in the English literature and dealing with human subjects) that used logistic regression analysis and bootstrap for its validation (0.003% of the total number of surgical articles that used logistic regression analysis and were published during the same period). It is clear that although bootstrap technology has broken down important barriers to surgical clinical research,[8] its importance appears still largely underestimated by most surgeons. This might be due to the paucity of readily available high-quality statistical software incorporating this analysis or the lack of understanding of the methodology, which might make surgeons perceive this statistical technique itself as a barrier to their interpretation of clinical data analysis reports. In this regard

a specific statistical training focusing on a reliable evaluation of the surgical outcome would be of help to disseminate a culture of quality improvement practice among surgeons.

On the basis of our results, we regard the process of developing risk models or risk factors without bootstrap validation as unreliable, obsolete, and resembling more an art than a science.[8] In view of the unreliability of the training-and-test method, previously published reports using it should be interpreted with caution.

Bootstrap analysis can formalize the development of model building, removing much of the human biases associated with regression analysis, providing a balance between selecting risk factors that are not reliable (type I error) and overlooking variables that are reliable (type II error), and introducing a concrete measure of reliability of the risk factors.[8] For this reason, the use of bootstrap analysis must be recommended for every future surgical model-building process.

References

1. Naftel DC. Do different investigators sometimes produce different multivariable equations from the same data? *J Thorac Cardiovasc Surg.* 1994;107:1528-9.
2. Diaconis P, Efron B. Computer-intensive methods in statistics. *Scientific American.* 1983;248:96-109.
3. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife and cross-validation. *Am Stat.* 1983;37:36-48.
4. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
5. Breiman L. Bagging predictors. *Machine Learning.* 1996;26:123-40.
6. Efron B. The bootstrap and modern statistics. *J Am Stat Assoc.* 2000; 95:1293-6.
7. Quanjer PhH, Tammeling GJ, Cotes JE, Pedersen OF, Peslin R, Yernault JC. Lung volumes and forced ventilatory flows. Report Working Party. Standardization of lung function tests. European Community for Steel and Coal. Official statement of the European Respiratory Society. *Eur Respir J.* 1993;6(suppl 16):5-40.
8. Blackstone EH. Breaking down barriers: helpful breakthrough statistical methods you need to understand better. *J Thorac Cardiovasc Surg.* 2001;122:430-9.
9. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med.* 1984;3:143-52.
10. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15:361-87.
11. Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons; 1989.
12. Grunkemeier GL, Jin R. Receiver operating characteristic curve analysis of clinical risk models. *Ann Thorac Surg.* 2001;72:323-6.
13. Berrisford R, Brunelli A, Rocco G, Treasure T, Utley M. The European Thoracic Surgery Database project: modelling the risk of in-hospital death following lung resection. *Eur J Cardiothorac Surg.* 2005;28:306-11.
14. Grunkemeier GL, Wu Y. Bootstrap resampling methods: something for nothing? *Ann Thorac Surg.* 2004;77:1142-4.
15. Harrel FE. Regression modelling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.