Harvard-MIT Division of Health Sciences and Technology
HST.947: Medical Artificial Intelligence
Prof. Peter Szolovits
Prof. Lucila Ohno-Machado

Evaluation of Support Vector Machine Risk Modeling over Time in Interventional Cardiology

Michael E. Matheny

1. BACKGROUND

In the last decade, significant emphasis has been placed on the development of statistical models to help predict risk in various patient populations. In addition to providing the basis for quality scorecards,[1,2] these risk profiles can be helpful on the procedural level to both patients and physicians. Numerous studies have shown that subjective prediction of risk by clinicians tends to be poor at very low and very high probabilities.[3,4] The use of various statistical methods can provide an objective estimation of outcome risk.

Percutaneous coronary intervention (PCI) is one of the most common procedures in cardiology, and is associated with significant morbidity and mortality. PCI is a high volume procedure with significant morbidity and mortality. The risk of adverse outcomes varies widely based on patient co-morbidities. Early attempts to build statistical models to predict mortality were hampered by the inclusion of non-standardized variables. In response, the American College of Cardiologists developed the National Cardiovascular Data Registry (ACC-NCDR), a standardized data set with clearly defined criteria, and a number of states now require mandatory reporting ACC-NCDR data.

The most common current modeling technique for this procedure is logistic regression (LR). There are currently a number of well-known LR mortality risk models, developed largely over the last 15 years.[5,6,7,8,9,10] However, most of these models include variables observed during the evaluation or intervention portions of the procedure. As a result, these models tend to be more applicable to help triage patients post-procedurally to an appropriate level of monitoring.

Application of these models on an individual patient level has been shown to be inaccurate when applied to different patient populations, or the same patient population after some time has passed, even when the overall population level performance of the model is maintained. There are a number of explanations for this, including changing medical practice[11], differing patient demographics, and different access to resources.

We seek to compare logistic regression modeling with a support vector machine to evaluate each model type's robustness over time with respect to discrimination and calibration.

2. METHODS

Source Data

Brigham & Women's Hospital (BWH), Boston Massachusetts has maintained a detailed database of all cases of PCI since 1997. The dataset is based on the ACC-NCDR dataset,[12] and participates in the mandatory Massachusetts Adverse Reporting System (MARS). All catheterization laboratory procedures performed are included in the database, and real-time data acquisition is accomplished through a dedicated team of trained nurses, physicians and technologists. A total of 5,383 PCI procedures were

recorded between January 01, 2002 and October 30, 2004 on all patients who underwent PCI at BWH. There were 78 (1.45%) deaths in the sample for this period.

This data set was sorted in two manners, sequentially and random. Each sorted data set was split into a two-thirds training set (3588) and one-thirds test set (1795). The sequentially sorted data was split by cases occurring on October 27, 2003.

Variable Selection & Statistical Evaluation

A thorough literature search was performed, and all previously identified risk factors for PCI were considered for inclusion in this study. [13,14,15,16,17,18,19] This list of variables was then filtered to include only those that could be known prior to the start of the procedure. The final variable set used for all models in this study is shown in Table 1.

Evaluation of all models was done with χ^2 and maximum log likelihood methods. Discrimination was assessed with the area under the receiver operating characteristic (ROC). [20,21,22] Calibration was evaluated with Hosmer-Lemeshow goodness-of-fit (HL-P) χ^2- estimates using deciles.[23]

Logistic Regression Model Development & Feature Selection

Backward stepwise logistic regression was performed using STATA 8.2 (College Station, TX). Initial feature selection was done using an exclusion threshold based on a residual Wald chi-square p value of 0.1. Parameters were removed in a systematic stepwise fashion based upon ROC and HL-P results. Initially, variables were removed singly one after another starting with the ones with the largest P value searching for the greatest HL-P value improvement. Once the optimal HL-P value was found for a single variable removal, that variable was dropped from the model, and the process repeated. Table 2 shows a partial evaluation chain of feature removal and evaluation measures. Table 1 lists those features removed from the model during this selection process.

Feature	*Selected*	*Feature*	*Selected*
Age	Yes	Hemodynamic Shock	Yes
Female	Yes	Hx Cardiac Arrest	Yes
Body Mass Index		Tachycardia	Yes
Smoker	Yes	Hemodynamic Instability	Yes
Hyperlipidemia	Yes	Pre-Proc IABP	Yes
Hypertension	Yes	AMI within 24 hours	Yes
Diabetes	Yes	AMI	
Fam Hx of CVD	Yes	Unstable Angina	Yes
Hx of COPD	Yes	Chronic Angina	Yes
Hx of CVD	Yes	Creatinine > 2.0	Yes
Hx of PVD	Yes	CHF on Presentation	
Hx of CHF		NYHA CHF Class 3 or 4	Yes
Prior MI	Yes	Thrombolytics Given	Yes
Prior PCI	Yes	AMI on Presentation	Yes
Prior CABG	Yes	NYHA CHF Class	Yes
Procedure Urgency		**LV Ejection Fraction**	

Table 1: Included Variables for All Model Evaluations, and Feature Selection for Logistic Regression Models. Some Variables are continuous, and were discretized for the models. Variables in bold were discretized.

Feature	ROC	HL P
All	0.952	0.0358
-BMI	0.952	0.0706
-EF	0.945	0.0004
-arrest	0.951	0.0602
-hyperlipid	0.9408	0.0001
-BMI,EF	0.9482	0.0743
-BMI, Urgency	0.949	0.1066
-BMI, Urgency, CHF Hx	0.956	0.956

Table 2: Partial feature removal evaluation chain with Area under the Receiver Operating Characteristic Curve and Hosmer-Lemeshow Goodness-of-Fit test results.

Support Vector Machine Model Development

Support Vector Machine (SVM) model methodology is based on a number of mathematical and statistical works.[24,25] GIST 2.1.1 (Columbia University, New York, NY) was used for development of the models specifically because of its output of both a classifier and a discriminant function. Polynomial kernels were evaluated from a power of 1 to 6, and are represented by d in the following kernel equation:

$$K(X,Y) = (X \circ Y + 1)^d$$

Radial-based kernels[26] were evaluated from an absolute width function from 0.25 to 2, and are represented by s in the following kernel equation:

$$K(X,Y) = e^{\frac{-\|X-Y\|^2}{2s^2}}$$

SVMs generally give outputs as a classifier {-1,1}, however some work has been done to provide a probabilistic output. Various methods have been used, and the one explored in this study is applying a sigmoid to the discriminant output (distance from the hyperplane).[27] This is done by using the discriminant as the only covariate in a logistic regression model and evaluating that model.

3. RESULTS

Logistic Regression Model

Full backward stepwise variable selection technique was performed using exclusion thresholds of 0.05, 0.10, and 0.15 on both sequential and random training data sets. The ROCs for the sequential training set varied from 0.936 to 0.949, and from 0.900 to 0.926 for the random training set. The HL-P values ranged from less than 0.001 to 0.004 for the sequential test set. The HL-P values were less than 0.001 for a threshold of 0.05, and 0.140 for thresholds of 0.10 and 0.15. These results are summarized in Table 3.

		Training		Test	
		ROC	HL	ROC	HL
	0.15	0.946	0.672	0.894	<0.001
SEQ	**0.10**	0.949	0.488	0.904	<0.001
	0.05	0.936	0.704	0.889	0.004
	0.15	0.926	0.269	0.920	0.140
RND	**0.10**	0.926	0.269	0.920	0.140
	0.05	0.900	0.095	0.899	<0.001

Table 3: Summary of ROC and Hosmer-Lemeshow Goodness-of-Fit testing for both data sets using stepwise backwards logistic regression with variable exclusion thresholds

Support Vector Machine Models

The sequential training data was used to develop support vector machines for a range of polynomial and radial-based kernels. The ROC for the polynomial kernels from 1 to 6 was 0.970 to 0.994. The only kernel that was not adequately fitted to the training set was for a power of 6 and the HL-P was 0.049. The remainder of the polynomial kernel HL-P's ranged from 0.503 to 0.999. These SVMs were then applied to the sequential test set, and the kernels of power 1 and 2 failed to calibrated (0.002-0.002), while the kernels of power 2 to 6 maintained calibration (0.067-0.738).

The same methods were applied using a radial-based kernel with a width factor ranging from 0.25 to 2.0. All kernel models were able to achieve ROCs from 0.974 to 1.000 on the training data, and 0.889 to 0.910 on the test data. All models were able to achieve calibrated HL-P values in a range of 0.502 to 1, and those kernels with a width factor from 0.25 to 1.00 achieved calibration with HL-P values of 0.111 to 0.601. Two radial-based width factors failed calibration, 1.5 and 2.0 with HL-P values of 0.001 each.

	Training		Test	
	ROC	HL	ROC	HL
Lin	0.970	0.503	0.896	0.003
P2	0.991	0.966	0.907	0.002
P3	0.994	0.999	0.909	0.067
P4	0.992	0.997	0.907	0.163
P5	0.987	0.818	0.899	0.713
P6	0.976	0.049	0.885	0.738
R 0.25	1	1	0.889	0.111
R 0.50	1	1	0.909	0.601
R 0.75	1	1	0.910	0.200
R 1.00	0.997	1	0.910	0.246
R 1.50	0.970	0.502	0.904	0.001
R 2.00	0.974	0.817	0.904	0.001

Table X: Support Vector Machine Area under the Receiver Operating Characteristic Curve and Hosmer-Lemeshow Goodness-of-Fit evaluation for sequential training and test sets across a range of polynomial and radial-based kernels

The randomized training data was applied in an identical way to develop support vector machines. The training data ROCs on the polynomial kernels of power 1 to 6 range from 0.963 to 0.997, and range from 0.862 to 0.903 on the test data. All polynomial kernel powers but 6 were calibrated on the training data with HL-P values

from 0.616 to 1.000, and failed to calibrate with 0.013 for a power of 6. All polynomial powers were successfully calibrated on the test data with HL-P values ranging from 0.521 to 0.856. All radial-based width factors used to develop models on the training set had ROCs of 0.895 to 1.000, and all were calibrated with HL-Ps from 0.961 to 1.000. When these models were evaluated on the test data, the ROC ranged from 0.891 to 0.911, and width factors from 0.50 to 2.00 were calibrated with HL-P values from 0.199 to 0.810. The width factor 0.25 failed to retain calibration with a HL-P value of 0.046.

	Training		Test	
	ROC	**HL**	**ROC**	**HL**
Lin	0.963	0.616	0.862	0.817
P2	0.992	0.920	0.900	0.754
P3	0.995	0.999	0.901	0.617
P4	0.996	1.000	0.903	0.521
P5	0.996	0.903	0.878	0.749
P6	0.997	0.013	0.871	0.856
R 0.25	0.999	1	0.891	0.046
R 0.50	1	1	0.908	0.593
R 0.75	1	1	0.910	0.199
R 1.00	0.997	1	0.911	0.542
R 1.50	0.992	0.961	0.907	0.810
R 2.00	0.895	0.961	0.898	0.232

Table X: Support Vector Machine ROC and Hosmer-Lemeshow Goodness-of-Fit evaluation for random training and test sets across a range of polynomial and radial-based kernels

4. DISCUSSION

All of the models had excellent discriminatory performance for both the training and test datasets. This indicates that the risk of death is well modeled on a population level by the available pre-procedural data, and supports the supposition that the variables collected well characterize the outcome. The consistent degradation of discrimination from the training to the test set is an expected finding and is comparable for all models.

The overall trend for the randomized data sets to remain calibrated across the logistic regression and support vector machine models suggests that both methodologies can adequately model the data in this domain without secular trends. However, these shifts in data registries are well documented, and are most commonly attributed to changes in data recording and changes in clinical practice. Developing robust models to maintain good case level estimations is important in real clinical applications for this reason.

Initially, a LR model was created using all features included in the data set. However, this model failed to calibrate on the training set. After significant feature subset selection, a model was found that would calibrate on the training set using the standard method. Because this used literature knowledge and expert knowledge of variable selection, this could be considered a more involved method of modeling than the SVM method that was used.

The logistic regression model failed to remain calibrated across all thresholds for the sequential test data, but did remain calibrated for a majority of the randomized test

data, suggesting that the feature selection and modeling technique were not robust with respect to changes over time of the data.

All of the support vector machine models were developed on the full feature set. This has both benefits and costs. First, this allows a less supervised model development and less dependence on expert domain knowledge. However, some variables could be mostly noise or not related to the outcome of interest. Feature selection could potentially further improve the SVM models.

The support vector machine was calibrated on both training data sets except the 6^{th} power of the polynomial kernel, suggesting that it was able to find a hyperplane that separated the data fairly well for most kernels. A subset of parameter ranges for each kernel maintained calibration on the sequential test data. This suggests that support vector machines could be more robust in terms of maintaining calibration over time for this domain. However, this relies heavily on the Hosmer-Lemeshow goodness-of-fit test as the measure of calibration. In addition, it is difficult to use the ROC and HL-P to find an optimal fit for a training set that will maintain calibration in the test set.

Overall, this study is promising in the pursuit of a risk modeling technique that is more robust in terms of retaining calibration over time. Additional work will need to be done with a different data set to further support this finding. In addition, exploration will be made into more rigorous methods of calibration evaluation.

REFERENCES

[1] Topol EJ, Block PC, Holmes DR, Klinke WP, Brinker JA.. Readiness for the scorecard era in cardiovascular medicine. Am J Cardiol 1995;75:1170-3.

[2] McNeil BJ, Pedersen SH, Catsonis C. Current issues in profiling quality of care. Inquiry 1992;29:298-307.

[3] Poses RM, Smith WR, McGlish DK, Huber EC, Clemo FLW, Schmitt BP, Alexander-Forti D, Racht EM, Colenda CC, Centor RM. Physician's Survival Predictions for Patients with Acute Congestive Heart Failure. Arch Intern Med 1997;157:1001-7.

[4] Perkins HS, Jonsen AR, Epstein WV. Providers as Predictors: Using Outcome Predictions in Intensive Care. Crit Care Med 1986;14:105.

[5] O'Connor GT, Malenka DJ, Quinton H, Robb JF, Kellett MA, Shubrooks S, Bradley WA, Hearne MJ, Watkins MW, Wennberg DE, Hettleman B, O'Rourke DJ, McGrath PD, Ryan T, VerLee P. Multivariate Prediction of In-Hospital Mortality After Percutaneous Coronary Interventions in 1994-1996. J Am Coll Cardiol 1999;34:681-691.

[6] Hannan EL, Arani DT, Johnson LW, Kemp HG, Lukacik G. Percutaneous Transluminal Coronary Angioplasty in New York State. JAMA 1992;268:3092-3097.

[7] Hannan EL, Racz M, Ryan TJ, McCallister BD, Johnson LW, Arani DT, Guerci AD, Sosa J, Topol EJ. Coronary Angioplasty Volume-Outcome Relationships for Hospitals and Cardiologists. JAMA 1997;279:892-898.

[8] Moscucci M, Kline-Rogers E, Share D, O'Donnell M, Maxwell-Eward A, Leengs WL, Kraft P, DeFranco AC, Chambers JL, Patel K, McGinnity JG, Eagle KA. Circulation 2001;104:263-268.

[9] Shaw RE, Anderson V, Brindis RG, Krone RJ, Klein LW, McKay CR, Block PC, Shaw LJ, Kathleen Hewitt, Weintraub WS. J Am Coll Cardiol 2002;39:1104-12.

[10] Ellis SG, Weintraub W, Holmes D, Shaw R, Block PC, King SB. Relation of Operator Volume and Experience to Procedural Outcome of Percutaneous Coronary Revascularization at Hospitals With High Interventional Volumes. Circulation 1997;95:2479.

[11] Williams DO, Holubkov R, Yeh W, Bourassa MG, Al-Bassam M, Block PC, Coady P, Cohen H, Cowley M, Dorros G, Faxon D, Holmes DR, Jacobs A, Kelsey SF, King SB 3rd, Myler R, Slater J, Stanek V, Vlachos HA, Detre KM Percutaneous coronary intervention in the current era compared with 1985-1986: the National Heart, Lung, and Blood Institute Registries Circulation 2000;102(24):2910-4.

[12] Cannon CP, Battler A, Brindis RG, Cox LJ, Ellis SG, Every NR, Flaherty JT, Harrington RA, Krumholz HM, Simoons ML, Van de Werf FJJ, Weintraub WS. ACC key data elements and definitions for measuring the clinical management and outcomes of patients with acute coronary syndromes: reference guide: a report of the American College of Cardiology Task Force on Clinical Data Standards (Acute Coronary Syndromes Writing Committee) 2001. http://www.acc.org/clinical/data_standards/ACS/acs_index.htm.

[13] O'Connor GT, Malenka DJ, Quinton H, Robb JF, Kellett MA, Shubrooks S, Bradley WA, Hearne MJ, Watkins MW, Wennberg DE, Hettleman B, O'Rourke DJ, McGrath PD, Ryan T, VerLee P. Multivariate Prediction of In-Hospital Mortality After Percutaneous Coronary Interventions in 1994-1996. J Am Coll Cardiol 1999;34:681-691.

[14] Hannan EL, Arani DT, Johnson LW, Kemp HG, Lukacik G. Percutaneous Transluminal Coronary Angioplasty in New York State. JAMA 1992;268:3092-3097.

[15] Hannan EL, Racz M, Ryan TJ, McCallister BD, Johnson LW, Arani DT, Guerci AD, Sosa J, Topol EJ. Coronary Angioplasty Volume-Outcome Relationships for Hospitals and Cardiologists. JAMA 1997;279:892-898.

[16] Moscucci M, Kline-Rogers E, Share D, O'Donnell M, Maxwell-Eward A, Leengs WL, Kraft P, DeFranco AC, Chambers JL, Patel K, McGinnity JG, Eagle KA. Circulation 2001;104:263-268.

[17] Shaw RE, Anderson V, Brindis RG, Krone RJ, Klein LW, McKay CR, Block PC, Shaw LJ, Kathleen Hewitt, Weintraub WS. Development of a risk adjustment mortality model using the American College of Cardiology-National Cardiovascular Data Registry (ACC-NCDR) experience: 1998-2000 J Am Coll Cardiol 2002;39:1104-12.

[18] Ellis SG, Weintraub W, Holmes D, Shaw R, Block PC, King SB. Relation of Operator Volume and Experience to Procedural Outcome of Percutaneous Coronary Revascularization at Hospitals With High Interventional Volumes. Circulation 1997;95:2479.

[19] Resnic FS, Ohno-Machado L, Selwyn A, Simon DI, Popma JJ. Simplified Risk Score Models Accurately Predict the Risk of Major In-Hospital Complications following Percutaneous Coronary Intervention. Am J Cardiol 2001;88:5-9.

[20] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Diagn Radiol 1982;143:29-36.

[21] Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285-1293.

[22] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43

[23] Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol. 1982;115(1):92-106.

[24] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings on the fifth annual workshop on Computational learning theory. 1992 Pittsburgh, PA ACM Press

[25] Brown MPS, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS 2000;97:262-7.

[26] Scholkopf B, Kah-Kay S, Burges CJC, et al. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans Sig Proc 1997;45:2758.

[27] Platt Jt. Probabilistic outputs for svms and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers. MIT Press, 1999.