Development of Statistical Methodologies and Risk Models to Perform Real-Time Safety Monitoring in Interventional Cardiology

By

Michael E. Matheny, M.D.

M.D. (2001)
University of Kentucky, Lexington, KY

Submitted to the Harvard-MIT Division of Health Sciences & Technology in Partial Fulfillment of the Requirements to the Degree of Master of Science in Biomedical Informatics

At the

Massachusetts Institute of Technology

June 2006

Signature of Author ______________________________
Harvard-MIT Division of Health Sciences and Technology
May 12, 2006

Certified by ______________________________
Frederic S. Resnic, MD, MS
Assistant Professor of Medicine
Harvard Medical School
Thesis Supervisor

Accepted by ______________________________
Martha Gray, PhD
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Sciences and Technology
[June 2006]
May, 2006

Development of Statistical Methodologies and Risk Models to Perform Real-Time Safety Monitoring in Interventional Cardiology

By

Michael E. Matheny, M.D.

Submitted to the Harvard-MIT Division of Health Sciences & Technology on May 12, 2006 in Partial Fulfillment of the Requirements to the Degree of Master of Science in Biomedical Informatics

Abstract

Post-marketing surveillance of medical pharmaceuticals and devices has received a great deal of media, legislative, and academic attention in the last decade. Among medical devices, these have largely been due to a small number of highly publicized adverse events, some of them in the domain of cardiac surgery and interventional cardiology. Phase three clinical trials for these devices are generally underpowered to detect rare adverse event rates, are performed in near-optimal environments, and regulators face significant pressure to deliver important medical devices to the public in a timely fashion.

All of these factors emphasize the importance of systematic monitoring of these devices after being released to the public, and the FDA and other regulatory agencies continue to struggle to perform this duty using a variety of voluntary and mandatory adverse event rate reporting policies. Data quality and comprehensiveness have generally suffered in this environment, and delayed awareness of potential problems. However, a number of mandatory reporting policies combined with improved standardization of data collection and definitions in the field of interventional cardiology and other clinical domains have provided recent opportunities for nearly "real-time" safety monitoring of medical device data.

Existing safety monitoring methodologies are non-medical in nature, and not well adapted to the relatively heterogeneous and noisy data common in medical applications. A web-based database-driven computer application was designed, and a number of experimental statistical methodologies were adapted from non-medical monitoring techniques as a proof of concept for the utility of an automated safety monitoring application. This application was successfully evaluated by comparing a local institution's drug-eluting stent in-hospital mortality rates to University of Michigan's bare-metal stent event rates. Sensitivity analyses of the experimental methodologies were performed, and a number of notable performance parameters were discovered. In addition, an evaluation of a number of well-validated external logistic regression models, and found that while population level estimation was well-preserved, individual estimation was compromised by application to external data. Subsequently, exploration of an alternative modeling technique, support vector machines, was performed in an effort to find a method with superior calibration performance for use in the safety monitoring application.

Thesis Supervisor: Frederic S. Resnic
Title: Assistant Professor of Medicine, Harvard Medical School

Biography: Michael E. Matheny received a National Merit Scholarship to attend the University of Kentucky's School of Engineering (1991-1997), and participated as a co-operative student (1994-1995) in activated carbon research at the Center For Applied Energy Research (Lexington, KY). He subsequently graduated Cum Laude in Chemical Engineering. He then attended University of Kentucky's School of Medicine (1997-2001) and was awarded the Robert P. Meriwether Scholarship (2000-2001). Dr. Matheny conducted his Internal Medicine training at St. Vincent's Hospital (Indianapolis, IN), and was awarded 1st place for a research abstract at the Indiana chapter of the American College of Physicians (2003) as well as receiving the Internal Medicine Research Award from St. Vincent's Hospital (2004). He was the recipient of a National Library of Medicine sponsored fellowship in biomedical informatics (2004-2007) conducted with the Decision Systems Group at Brigham & Women's Hospital under the direction of Dr. Robert Greenes. Dr. Matheny is a Diplomat of the American Board of Internal Medicine.

Contents:

Introduction

Post-marketing surveillance of medical devices by the Food and Drug Administration (FDA) has undergone tremendous change in the last few decades.[1-4] These changes were largely due to a small number of highly publicized adverse events; some of them in the domain of cardiac surgery and interventional cardiology.[3, 5-13]

Medical devices are frequently released quickly because of the need to deliver potentially lifesaving medical advances to the public. Rare adverse events are an ongoing concern in this environment because they may not be discovered in pre-marketing trials due to a small sample sizes and a bias towards healthier subjects.[14] To balance this, the FDA has shifted more of its device evaluation to the post-marketing period.[15] This creates the potential for large numbers of patients to be exposed to a new product in the absence of long-term follow-up data, and emphasizes the need for comprehensive methods in post-marketing surveillance.[16]

The data the FDA uses to conduct this surveillance is very heterogeneous, and results from a variety of voluntary and mandatory reporting policies.[1, 6, 14, 17-22] The voluntary policies create significant limitations in event rate recognition through underreporting, bias, and highly variable reporting quality.[14] In response, some agencies have implemented mandatory reporting for medical devices in certain clinical areas, and medical societies have made strides in standardizing data element definitions within their respective domains.[23]

Continued improvements in the quality and volume of reported data have created opportunities for timely and efficient analysis and reporting of alarming trends in patient outcomes. However, standard techniques regarding alerting thresholds and benchmarking methodologies do not exist for post-marketing medical device surveillance. Non-medical industries have been using a variety of automated statistical process control (SPC) techniques for some time for comparable quality control purposes. [28-32] These systems rely on high-quality automated data collection, and apply stringent error thresholds[24] that make direct application of these methods for medical devices difficult.

Statistical process control is a classical frequentist technique typically comparing observed event rates to acceptable rates of adverse events based on previously published or observed empirical data. However, this is very limiting when no or very limited prior empirical data are available. Another methodology, Bayesian updating statistics (BUS), seeks to address this limitation directly through the construction and explicit use of prior probability estimates and provide direct comparison of the final posterior distribution to the prior estimate.

Establishing a standard methodology for the selection and incorporation of prior empirical data to generate alerting thresholds integral is integral to this endeavor. Both SPC and BUS provide for population-level thresholds with static risk stratification. Logistic regression (LR) is the most popular of the modeling techniques that have been developed over the last few decades to help stratify and predict risk for patient subpopulations. These models have been widely used to improve the quality of care,[25] provide institutional quality scorecards[26], provide risk stratification[27] and assist patient selection[28] in research, evaluate futility of care,[29] and to provide individual patient prognostications. The ability of LR to individually risk stratify a patient using a large number of clinical factors could be used to develop a stand-alone monitoring methodology, or be used synergistically with other methods as a risk stratification tool.

Interventional cardiology (IC) is an ideal medical domain to pioneer prospective real-time safety monitoring in medical devices for a number of reasons. Not only has IC been an active area of medical device development in recent years, but it is one of the few domains with high quality comprehensive data. This is largely due to the development of a national standardized data dictionary,[23] and an increase in mandatory electronic data collection and reporting by some state agencies.

In addition, there are a number of well-known LR mortality risk models that have been developed in IC over the last 15 years.[30-35] Several studies of these models have shown good external validation with respect to both calibration and discrimination.[36-40] Others have shown a loss in either discrimination, calibration,[41] or both.[42] This is thought to be primarily related to medical practice and patient composition changes related to geography and time.[43, 44] A study comparing the demographics of percutaneous coronary intervention (PCI) patients in two registries collected twelve years apart found significant differences in a number of important risk factors.[45]

This highlights the need for a thorough evaluation of existing and newly developed LR models for use with a safety monitoring system. Also, exploration of alternative modeling techniques, such as a new machine learning technique called Support Vector Machines (SVM), could potentially improve risk stratification performance in this domain.

The specific goals of this project are to:

1) Develop local logistic regression (LR) risk models, and validate both local and well-known external LR models in order to optimize this modeling method for use in a real-time safety monitoring tool.
2) Explore a recent advance in machine learning, support vector machines, as an alternative risk modeling and stratification method for use in the monitoring application.
3) Develop prospective methods of statistical monitoring and alerting thresholds. Exploration of these methods will include derivations of statistical process control, Bayesian updating statistics, and logistic regression based risk stratification and assessment.
4) Implement a web-based safety monitoring tool in interventional cardiology that allows detailed evaluation of specific outcomes.
5) Perform a sensitivity analysis between statistical process control and Bayesian statistical updating monitoring methods on interventional cardiology data.

Manuscript Organization

This manuscript will be organized into four chapters. Chapter One will describe the design of a new local logistic regression model for the outcome of post-intervention in-hospital death. This model will then be internally validated, and external validation will be conducted on a number of well-known logistic regression models for this outcome from other centers.

Chapter Two which will explore a new machine learning methodology, support vector machines, to determine if the use of this risk modeling method would be a reasonable addition to the safety monitoring tool.

The results of the LR model evaluations were incorporated into the design and development of the statistical methodologies necessary to conduct real-time safety monitoring. The development of these methods, as well as the implementation of a web-based application providing this type of analysis, will be described in Chapter Three.

A sensitivity analysis between statistical process control and Bayesian updating statistics methodologies will be performed in Chapter Four. Actual interventional cardiology data will be used to provide scaled outcome rates for evaluation over a range of baseline event rates and volumes. A conclusion will then summarize the findings in each portion of the work.

Chapter 1: Performance Evaluation of Logistic Regression Risk Model

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,[26, 46] these risk profiles can be helpful on the procedural level to both patients and physicians. Numerous studies have shown that subjective prediction of risk tends to be poor at very low and very high probabilities.[47, 48] The use of various statistical methods can provide an objective estimation of outcome risk.

There has been conflicting literature on whether or not these models can be used outside of their development population. Initial validation is usually based on patients from a given geography and time frame. These evaluations are only directly applicable in that respect, and concerns have been raised about the applicability of a model when patient demographics change with geography, clinical practice changes with time, and disease prevalence changes with both. Some of this concern stems from prior analyses showing deleterious effects on accuracy by changes in geography and time.[43] A study comparing the demographics of percutaneous coronary intervention (PCI) patients in two registries collected twelve years apart found significant differences in age, lesion severity, thrombolytic use, stent use, and death that highlight how much the characteristics of a population can change with a decade of medical advances.[45]

Continuous evaluation of model performance is important to ascertain that classification performance does not degrade with time. Some models are re-developed periodically to adjust for temporal trends.[49] Also important is validation of a model on geographically or temporally distant populations.[50] Constructing a model using a large numbers of patient encounters across a wide variety of geographic areas increases the probability that the model will be suited for different populations, but the only way to determine the model's applicability is to verify the performance empirically in representative sample.

In the field of cardiology, one of the most widely studied areas of risk stratification has been coronary angiography. This article seeks to build on prior work on the applicability of risk models in different geographies and over time. Several prior studies of PCI risk models have shown good external validation with respect to both calibration and discrimination.[36-40] Others have shown a loss in either discrimination, calibration,[41] or both.[42] In the present study we consider the hypothesis that models exhibit differences in discrimination and calibration over space and time.

Methods

Data Collection

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 American College of Cardiology National Data Repository dataset,[23] with a variety of additional, detailed, data elements. The registry is part of the quality assessment and quality improvement program of Brigham & Women's Hospital, and was approved by the hospital Institutional Review Board. 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 5,216 PCI procedures were recorded between January 01, 2002 and September 30, 2004 on all patients who underwent PCI at BWH. This data set serves as the source for the evaluation of each model in this study.

Model Evaluation

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 curve (AUC).[51, 52] A summary of each of the models used is shown in Table 1. Calibration was evaluated with Hosmer-Lemeshow goodness-of-fit χ^2- estimates using deciles.[53] 95% confidence intervals for these parameters were computed with the non-parametric bootstrapping method of STATA (Version 8.2, College Station, TX).[54] These CIs were reported using the percentile method, or bias corrected method if the estimation bias was greater than 25% of the standard error.[55]

Model	Dates		Location	Sample	AUC	HL(p)	Validation Type
NNE 1999	1/1/1994	12/31/1996	NH, ME, MA, VT (7)	15331	0.88	0.09	Bootstrap Resampling
NY 1992	1/1/1991	6/30/1991	NY	5827	0.884	NA	Subset Significance
NY 1997	1/1/1991	12/31/1994	NY	62670	0.892	0.11	Subset Significance
MI 2001	10/1/1999	8/30/2000	Detroit, MI	10796	0.90	0.5	Training/Test
ACC 2002	1/1/1998	9/30/2000	National	100253	0.89	0.133	Training/Test
BWH 2001	1/1/1997	12/31/1999	Boston, MA	2804	0.86	0.11	Training/Test
CC 1997	1/11993	12/31/1994	Cleveland, OH (5)	12985	0.846	NA	Bootstrap Resampling

Table 1: Summary of the training data sets for the models used in this study. Sample = sample size. AUC = area under the receiver operating characteristic. HL(p) = Hosmer-Lemeshow p value.

External Validation of Risk Models

Six external and one local previously described multivariate post-PCI in-hospital mortality risk models were evaluated using the BWH data set: the Northern New England Cooperative Group (NNE 1999),[30] the New York State (NY 1992 & NY 1997),[31, 32] University of Michigan Consortium (MI 2001),[33] the American College of Cardiology-National Cardiovascular Data Registry (ACC 2002),[34] the Cleveland Clinic Foundation Multi-Center (CC 1997),[35] and the Brigham & Women's Hospital (BWH 2001)[56] models. Pair-wise comparison of the area under the ROC curve for each model was performed by Analyse-It (Version 1.71, Leeds, England, UK).

Local Model Development

To test the hypothesis that time and space degrade the accuracy of a risk model, a new local model was developed using the same BWH data that was used to evaluate the discrimination and calibration of existing models. Standard univariate methods were used to generate odds ratios (ORs) with 95% confidence intervals (CIs) and p values to select variables that would be included in the new model.[57] Additionally, all available covariates which have been shown to be univariate risk factors in previous studies were included in the analysis (Table 2). Backward stepwise logistic regression was performed using STATA.[19] Variables were first removed using a residual Wald chi-square p value of 0.1, and then considered for inclusion based on a p value of 0.05. Since there was no independent test set, the evaluation was based on bootstrap resampling with 1000 samples.[58]

Factor	% Pts	% Deaths	OR	95% CI	p
Age					
<50	11.0	0.2	1.00	Ref	
50-59	21.6	0.4	2.55	0.30 - 39.9	0.392
60-69	27.8	0.9	5.20	0.68 - 39.9	0.112
70-79	27.6	1.5	8.91	1.199 - 66.3	0.033
>79	11.9	4.8	29.3	3.98 - 215.5	0.001
Gender					
Male	70.7	1.4	1.00	Ref.	
Female	29.3	1.4	1.02	0.61 - 1.70	0.952
Diabetes	31.7	1.8	1.58	0.99 - 2.5	0.058
PVD	9.5	2.4	1.97	1.05 - 3.69	0.034
COPD	10.6	2.0	1.55	0.81 - 2.97	0.183
Shock	1.7	37.4	82.0	48.1 – 139.8	<0.001
Unstable Angina	4.9	11.6	15.8	9.7 - 25.7	<0.001
Urgency					
Elective	49.9	0.3	1.00	Ref	
Urgent	37.9	0.9	2.98	1.3 - 6.9	0.010
Emergent	11.8	5.7	19.6	9.1 - 42.6	<0.001
Salvage	0.4	45.4	270.3	91 - 803.2	<0.001
LVEF					
>39	91.3	1.1	1.00	Ref	
20-39	7.6	3.6	3.22	1.77 - 5.84	<0.001
<20	1.1	5.5	5.04	1.53 - 16.6	0.008
Tachycardia	2.4	13.5	14.5	8-17 - 25.9	<0.001
Pre-PCI IABP	0.7	19.4	19.3	8.15 - 45.7	<0.001
AMI 24 Hr	10.6	5.2	6.1	3.8 - 9.8	<0.001
Cr > 2.0 mg/dL	5.3	5.0	4.5	2.5 - 8.3	<0.001
CHF	10.1	4.0	3.9	2.3 - 6.5	<0.001
Prior PCI	33.8	0.5	0.28	0.14 - 0.57	<0.001
Prior CABG	1101	1.1	0.76	0.41 - 1.42	0.385
Lesion Risk					
Low	66.3	0.5	1.00	Ref	
High	33.7	3.0	5.5	3.24 - 9.4	<0.001
Intervention					
LAD	42.4	1.9	1.87	1.17 - 3.01	0.010
Disease Location					
Proximal LAD	47.2	2.2	3.34	1.95 - 5.72	<0.001
RCA	52.3	1.7	1.69	1.03 - 2.76	0.036
Diseased Vessels					
0	9.0	0.4	1.00	Ref.	
1	52.6	0.9	2.07	0.49 - 8.8	0.323
2	25.5	1.9	4.5	1.06 - 19.1	0.041
3	12.8	3.0	7.3	1.7 - 31.2	0.008

Table 2: Univariate Association of Factors with In-Hospital Mortality and Registry Demographics. % Pts = percent of sample population. % Deaths = percent of deaths within the sub-population. OR = Odds Ratio. 95% CI = 95% Confidence Interval. p = p Value. PVD = Peripheral Vascular Disease. COPD = Chronic Obstructive Pulmonary Disease. LVEF = Left Ventricular Ejection Fraction. PCI = Percutaneous Coronary Intervention. IABP = Intra-Aortic Balloon Pump. AMI = Acute Myocardial Infarction. Cr = Creatinine. CHF = Congestive Heart Failure. CABG = Coronary Artery Bypass Grafting. LAD = Left Anterior Descending. RCA = Right Coronary Artery.

Results

Local Multivariate Prediction Rule Development

After full backward stepwise variable selection, the variables associated with an increased risk included older age, diabetes, unstable angina, salvage procedure, cardiogenic shock, AMI, and any intervention on the left anterior descending artery as shown in Table 3. The AUC was 0.929 revealing excellent discriminatory ability of the new model, and bootstrap re-sampling the data to obtain a 95% CI of 0.90-0.96 with an SE of 0.017, indicating a good ability to discriminate with respect to the outcome of death. The model had an adequate goodness of fit (HL χ^2=7.61 with 8 d.f., p=0.473).

Factor	OR	95% CI	β	p
Prior PCI	0.30	0.12 - 0.74	-1.20	0.009
Age (yrs)				
60-69	4.41	1.31 - 14.84	1.48	0.016
70-79	8.25	2.58 - 26.34	2.11	<0.001
80+	21.39	6.76 - 66.97	3.06	<0.001
Diabetes	1.82	1.02 - 3.26	0.60	0.042
Unstable	5.46	2.82 - 10.52	1.70	<0.001
Salvage	19.25	5.06 - 73.24	2.96	<0.001
Shock	14.86	7.39 - 29.87	2.70	<0.001
AMI Present	1.72	1.37 - 2.17	0.54	<0.001
Any LAD PCI	1.72	0.97 - 3.07	0.54	0.066

Table 3: Multivariate Analysis of Factors Significantly Associated with In-Hospital Mortality in the New BWH model. OR = Odds Ratio. 95% CI = 95% Confidence Interval. β = Beta Coefficient. p = p Value. Constant (intercept) = -7.777; Hosmer and Lemeshow goodness-of-fit χ^2 = 7.61; p = 0.473; AUC = 0.929. AMI = acute myocardial infarction. LAD = left anterior descending. PCI = percutaneous coronary intervention.

External Validation

The external model performances on the BWH dataset are shown in Table 4. During the study period there were 71 observed deaths (1.36%). BWH 2004 very closely approximated this with 70.5 deaths, NY 1992, CC 1997, and BWH 2001 over predicted, and the remainder under predicted. The AUC indicates excellent discrimination across all models, with the worst being the New York State 1992 model and the best being the new local model. A summary view of the AUC for all models is shown in Figure 1. Of the external models, the best AUC was obtained by the ACC 2002 model.

Curve	Deaths	AUC	95% CI	HL χ^2	95% CI	HL (p)	95% CI
NY 1992	96.7	0.82	0.76 - 0.88	31.1	13.9 - 50.0	<0.001	<0.001 - 0.003
NY 1997	61.6	0.88	0.81 - 0.92	32.2	16.4 - 45.5	<0.001	<0.001 - 0.004
CC 1997	78.8	0.88	0.82 - 0.93	27.8	19.6 - 38.7	<0.001	<0.001 - 0.013
NNE 1999	56.2	0.89	0.84 - 0.94	45.9	31.9 - 67.4	<0.001	<0.001 - <0.001
MI 2001	61.8	0.86	0.81 - 0.90	30.4	16.7 - 43.1	<0.001	<0.001 - 0.011
BWH 2001	136.1	0.89	0.84 - 0.93	39.7	23.2 - 73.3	<0.001	<0.001 - 0.001
ACC 2002	49.9	0.90	0.84 - 0.95	42.0	24.9 - 63.3	<0.001	<0.001 - 0.002
BWH 2004	70.5	0.93	0.89 - 0.96	7.61	1.5 - 14.2	0.473	0.073 - 0.992

Table 4: Summary of discrimination and calibration performance for each model. Deaths = Estimated Deaths. AUC = Area Under the Receiver Operating Characteristic Curve. 95% CI = 95% Confidence Interval. HL χ^2 = Hosmer-Lemeshow χ^2. HL(p) = Hosmer-Lemeshow probability > χ^2 Value.

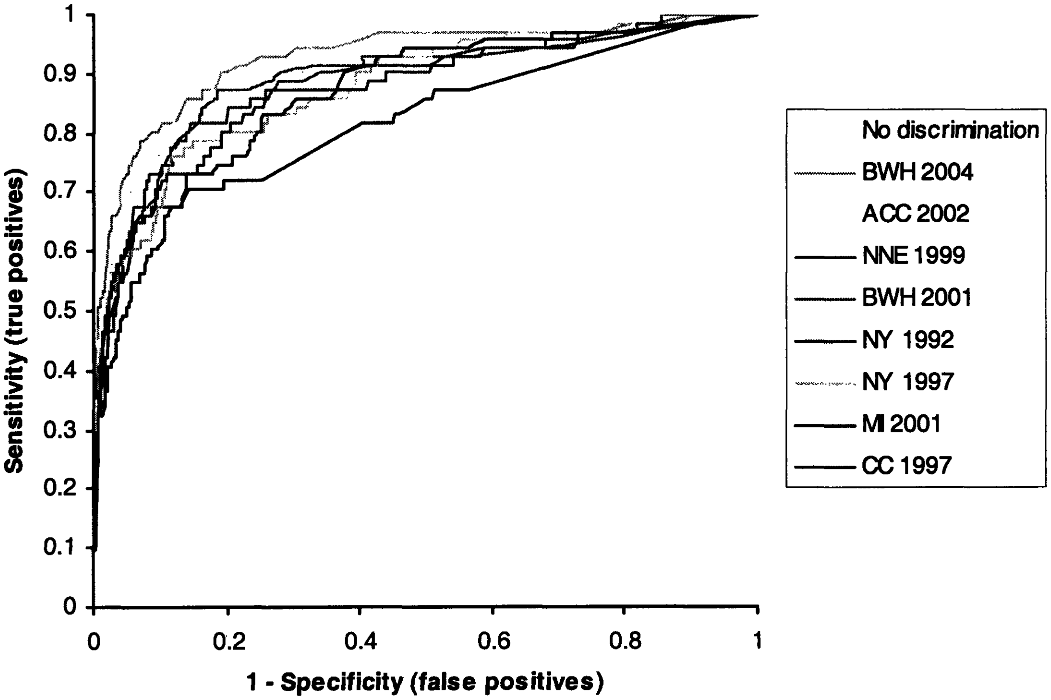


Figure 1: AUC for all models. The grey line shows no discrimination.

Pair-wise AUC comparisons were performed as well, shown in Table 5, by using the method described by Hanley and McNeil.[59] Overall, the best discrimination was obtained by the new local model, which attained significance with respect to every model but ACC 2002. The second best performance was by the external model constructed with the largest training set (ACC 2002), followed by the old local model (BWH 2001). Significant differences were noted between NY 1992 and every model but MI 2001, as well as between MI 2001 and ACC 2002. This indicates that the NY 1992 model, and to a lesser extent the MI 2001 model, is the least discriminatory.

	NY 1992		NY 1997		CC 1997		NNE 1999		MI 2001		BWH 2001		ACC 2002	
	Diff	p	Diff	p	Diff	p	Diff	p	Diff	p	Diff	p	Diff	p
NY 1992														
NY 1997	**0.056**	**0.007**												
CC 1997	0.051	0.101	0.004	0.859										
NNE 1999	**0.062**	**0.013**	-0.007	0.712	0.011	0.644								
MI 2001	0.041	0.165	-0.015	0.485	-0.01	0.627	-0.022	0.310						
BWH 2001	**0.066**	**0.019**	0.011	0.602	0.015	0.551	0.004	0.849	0.026	0.287				
ACC 2002	**0.080**	**0.002**	0.025	0.145	0.03	0.176	0.018	0.254	**0.040**	**0.045**	0.014	0.519		
BWH 2004	**0.105**	**0.001**	**0.049**	**0.007**	**0.053**	**0.011**	**0.043**	**0.048**	**0.064**	**0.003**	**0.038**	**0.050**	0.024	0.176

Table 5: Pair-wise Discrimination Model Comparison. Diff = AUC difference. p = p value of difference.

The Hosmer-Lemeshow goodness-of-fit test reveals poor calibration ($p < 0.05$) for all the models but the newly developed one. Calibration for all models was further explored by plotting the observed to expected frequency of death for each quintile of every model. Figure 2-B is provided to more clearly show the relationships for the low risk population. As shown in Figure 2, the NY 1992 model underestimated the risk of death for low scoring patients, and over estimated this risk for high scoring patients. NY 1997 performed fairly well for low risk patients, but overestimated the probability of death for high risk patients. ACC 2002 performed well under low risk conditions, but significantly underestimated the probability of death for high risk conditions. NNE 1999 consistently under predicted deaths, and CC 1997 as well as BWH 2001 consistently overestimated mortality risk. As expected, BWH 2004 performs well, but since this is not an independent test sample, this result should be interpreted with caution.

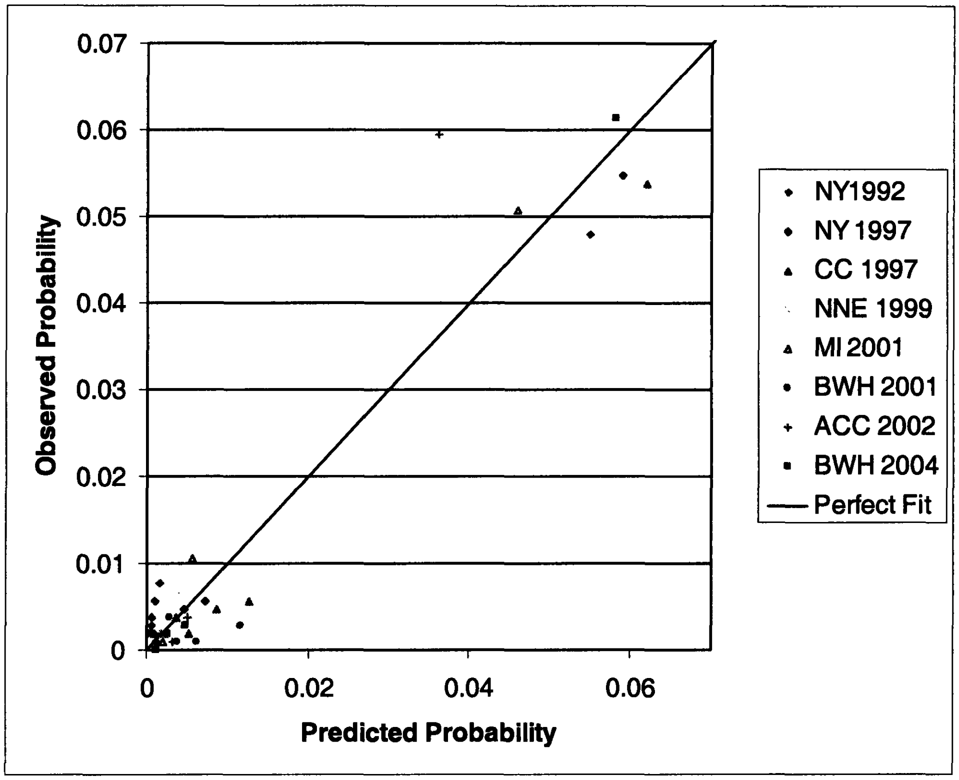


Figure 2-A: The observed and expected mortality rates for each quintile of patient risk. Each risk quintile contains approximately 1050 patients. The diagonal line represents a perfect agreement between observed and expected mortality estimates.

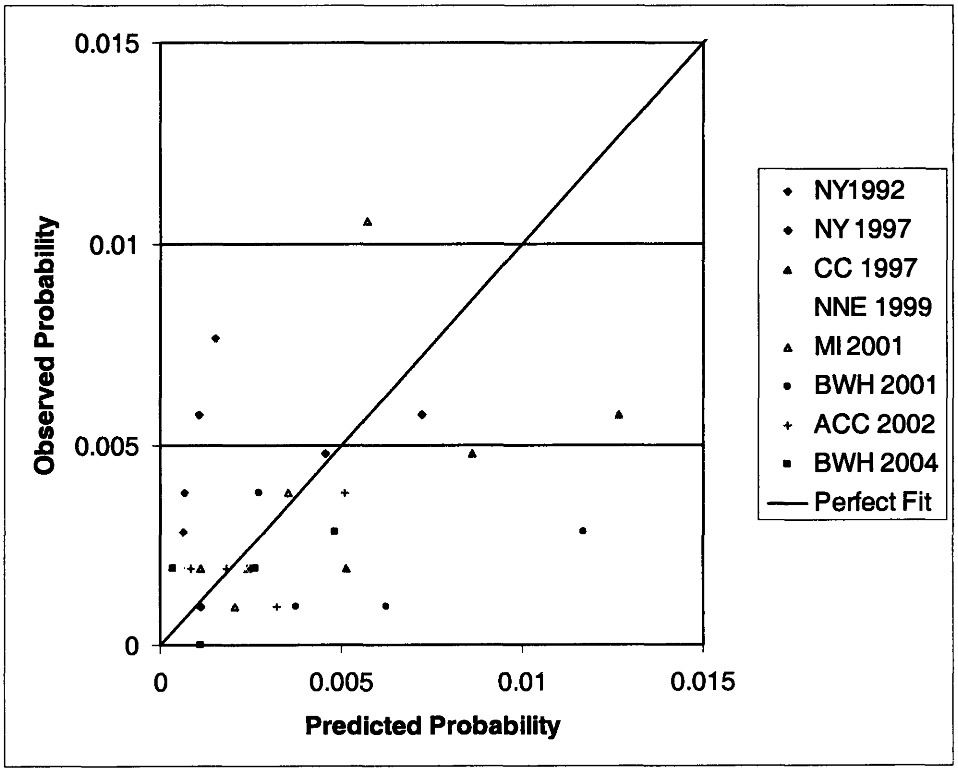


Figure 2-B: Expanded View from 0 to 0.15 of the Observed to Expected Probability Ratios.

Discussion

Interventional cardiology practice has changed significantly over the last decade. Procedural skill development, pharmacology, and device development have all contributed to the evolution of the field, and patient outcomes have changed over that time period in response to these advances. There as been a substantial reduction in risk of death and major adverse cardiac events (MACE)[60] over the past decade. All of these factors create a moving target for any risk stratification model.

All the external models evaluated on the BWH data set showed good discrimination. The model with the worst discrimination was NY 1992, which was to be expected due to the age of the study, and small sample size with which the model was developed. The best external model was the one developed on a national database with the most patient records, suggesting that geographic issues may be related to discrimination. Although these results are promising, it is important to note that discrimination is not the only (and possibly not the most important) factor in determining the applicability of a prognostic model from the perspective of physicians and patients. A model can exhibit perfect discrimination but still be useless for application on individual cases. Good calibration is essential for this type of application. All models, except possibly the one developed locally with recent data, but including a model derived locally, showed poor calibration for our test set, suggesting that time may play an important role in the applicability of a model.

Similar findings have been previously reported. Some techniques have been suggested to recalibrate the model.[61] One of these techniques was employed by Kizer, et al[41] and Peterson, et al[62] with some success, and may offer a strategy to maintain discrimination and improve calibration.

This study supports routine evaluation of any risk model, including aging local models, prior to local implementation. Discrimination was maintained for most risk adjustment models, though those more recently published and those based on the largest original datasets appeared to have the most robust discrimination when applied to a current clinical dataset. The preservation of discrimination supports the use of these models for generic risk stratification, but the poor calibration indicates that they are not useful for application in individual cases: the estimated risk of death for a single patient that is produced by these models is incorrect.

The poor calibration of the prior models suggests that variations in practice and patient demographics as well as clinical features over time have a large effect at the patient level on the risk estimate's accuracy. Further study is required to identify the optimum frequency of model recalibration.

Chapter 2: Performance Evaluation of Support Vector Machine Risk Models

Background

In the last few decades, significant emphasis has been placed on the development of statistical models to help predict risk in various patient populations. These models have been widely used to improve the quality of care,[25] provide institutional quality scorecards[26], provide risk stratification[27] and assist patient selection[28] in research, evaluate futility of care,[29] and to provide individual patient prognostications.

Percutaneous coronary intervention (PCI) is one of the most common procedures in cardiology, and is associated with significant morbidity and mortality. Opportunities for providing objective risk assessment in this domain have been improving since the development of a national standardized data dictionary,[23] and an increase in mandatory electronic data collection and reporting by some state agencies.

The gold standard modeling technique in this domain is logistic regression (LR). There are a number of well-known LR mortality risk models that have been developed over the last 15 years.[30-35] Discrimination of these models is generally high, and has been retained in external validation studies. However, calibration degraded when these models were applied to subsequent local or external data.[74] There are a number of possible explanations for this, including changing medical practice, differing patient demographics, and different access to resources.[44]

Calibration failure primarily affects individual patient prognostication, and negatively impacts risk stratification and any application that relies on individual estimates. Recalibration techniques are being explored to provide adequate calibration over time, but no method has emerged as a standard.

Application of a modeling technique that outperforms LR in terms of calibration can extend the useful life of a risk model before recalibration or model refitting is required. One of the most recent developments in artificial intelligence modeling has been Support Vector Machines (SVM). These models are able to find an optimal separation hyperplane in a multi-dimensional space to perform classification of a dichotomous outcome. To our knowledge, this methodology has not been explored in interventional cardiology, and has potential applications in a number of applications, such as real-time safety monitoring of new medical devices.[75]

In this study, we seek to evaluate and compare the discrimination and calibration performance between a variety of LR and SVM risk models in the evaluation of post-procedural in-hospital mortality in PCI.

Methods

Source Data

Data were collected from Brigham and Women's Hospital (BWH) (Boston, MA) containing all cases (7914) of percutaneous coronary intervention (PCI) performed at the institution from January 1, 2002 to December 31, 2005. The outcome of interest was post-procedural in-hospital death, and there were 124 (1.57%) events during the collection period. The cases were used to generate 100 random data sets. All cases were used in each set, and 5540 were allocated for training and 2374 were allocated for testing. For SVM evaluation, each training set was randomly divided into 3957 *kernel* training and 1583 *sigmoid* training portions. Data element definitions were based on the American College of Cardiology – National Cardiovascular Data Registry (ACC-NCDR) data dictionary.[23] The BWH Institutional Review Board approved this study.

Acute Heart Attack
Age
Body Mass Index
CHF Class
CHF on Presentation
Creatinine > 2.0 mg/dL
Diabetes
Elective Case
Emergent Case
Family Hx Heart Disease
Heart Rate
Hx COPD
Hx PVD
Hx Stroke
Hyperlipidemia
Hypertension
IABP
Prior PCI
Shock
Unstable Angina
Urgent Case

Table 1: Hx = history, COPD = Chronic Obstructive Pulmonary Disease, PVD = Peripheral Vascular Disease, CHF = Congestive Heart Failure, PCI = Percutaneous Coronary Intervention, IABP = Intra-Aortic Balloon Pump.

Variable Selection

After careful literature review, all previously identified risk factors for PCI were selected for inclusion in this study.[30-35, 56] Univariate analysis was then performed with SAS 9.1 (Cary, NC). Variables not significantly associated with the outcome of death were removed from the data (sex, smoking status, prior myocardial infarction, prior CABG, and a history of chronic renal insufficiency). A total of 21 variables were retained for use in model creation and analysis. These variables are listed in Table 1.

Logistic Regression

Model development for LR was performed using the PROC LOGISTIC of SAS. A standard backwards stepwise model selection method was used.[80] In addition, 3-fold cross-validation (CV) was performed on each training set to determine the optimum threshold for feature selection in the backwards stepwise method. The modeling parameters were optimized separately for mean squared error (MSE), Area under the Receiver Operating Characteristic (AUC) curve, and Hosmer-Lemeshow (HL) χ^2 goodness-of-fit values. The thresholds evaluated were 0.05 to 0.50, in 0.05 increments. The optimized threshold parameters were then used to generate a model for each entire training data set, and then applied to the respective test data.

Support Vector Machine

The Support Vector Machine (SVM) models were developed using GIST 2.2.1 (Columbia University, New York, NY). Radial (SVM-R) and polynomial (SVM-P) based kernels were selected for evaluation because of their good performance in other domains. Polynomial kernels transform the feature matrix using the following equation, where X and Y are features (predictors) and the class variable, respectively:

$$K(X,Y) = (X \text{ o} Y + 1)^d$$

The primary kernel parameter is the power, represented as *d* in the above equation. Gaussian radial-based kernels[81] transform the feature matrix using the following equation:

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

The primary kernel parameter is the width function, represented as *w* in the above equation. The other parameter that can be used in both kernels is the cost function, which determines the ratio of error weight between false positives and false negatives. This parameter was fixed constant at a value of 1. Classification SVMs give outputs as a binary classifier (-1, 1) and also as a continuous discriminant (distance from the hyperplane). A method described by Platt[82] allows the generation of a probabilistic outcome by fitting a sigmoid function to the discriminant using independent holdout training data. In this study, we used the corrected Platt algorithm provided by Lin and colleagues.[83]

The parameters of each kernel type were optimized on the kernel training set separately for MSE, AUC and HL χ^2 values by a grid search method on the training set, using 3-fold cross-validation.[84] The sigmoid training set was used to convert discriminant results into probabilities.

The width function for the radial-based kernel ranged from 2^{-4} to 2^4 (2^{-4}, 2^{-3}, 2^{-2}, etc.), and power for the polynomial-based kernel ranged from 1 to 6 by integers. For each respective kernel type and optimization parameter, the best kernel parameter was used to generate a model on the entire kernel training set, and a sigmoid for discriminant conversion was generated using the sigmoid training set. Each of the models was then applied to the respective test data set.

Statistical Evaluation

Discrimination was assessed with the area under the receiver operating characteristic.[51] Calibration was evaluated by the Hosmer-Lemeshow goodness-of-fit χ^2 goodness-of-fit estimates.[78] χ^2 values with 8 degrees of freedom are considered adequately calibrated for values of 15.51 or less (corresponding to a p value >= 0.05). Pair-wise comparison between performance measures was performed using a one-way ANOVA test for summary values with known standard errors.

Results

A summary of test data evaluation for each model type and cross validation optimization parameter is shown in Table 2. Each respective model parameter (p-value threshold for a variable to stay in the model, kernel width factor, and kernel power for LR, SVM-R, and SVM-P, respectively) and performance measure includes the mean values and 95% confidence intervals for the respective model type.

Results of the pair-wise comparison between each optimization type of each modeling method are shown in Table 3. The upper right half of the table contains AUC comparisons, and the lower left half of the table contains HL χ^2 comparisons. p values less than 0.05 indicate a significant difference between the pair.

None of the LR models resulted in AUCs that were significantly different from any other LR model (all $p > 0.05$). All of the LR models had higher AUCs than the radial-based SVMs (all $p < 0.05$). The LR models had significantly higher AUCs when compared to the HL χ^2 optimization method of SVM-P ($p = 0.012$, 0.012, 0.025, and 0.027), but not higher than the AUCs from MSE or AUC methods.

None of the LR models had HL χ^2 values significantly different from any other LR model (all $p > 0.05$). All of the radial-based and polynomial-based SVM models showed significantly lower HL χ^2 values than any of the LR models (all $p < 0.05$). The HL χ^2 optimization method was superior to both the MSE and AUC methods in the radial-based SVM models ($p < 0.001$). The HL χ^2 optimization method in both kernel types was significantly better than all of the other model versions based on MSE and AUC, except the MSE and AUC optimization methods of SVM-P ($p < 0.05$).

Model	**Opt**	**Parameter Mean**	**AUC (95% CI)**	**HL χ^2 (95% CI)**
LR	None	0.10	0.911 (0.905 - 0.916)	101.1 (49.4 - 152.8)
LR	MSE	0.25 (0.22 - 0.28)	0.912 (0.906 - 0.917)	89.8 (53.0 - 126.7)
LR	AUC	0.29 (0.26 - 0.33)	0.912 (0.906 - 0.917)	94.7 (46.0 - 143.4)
LR	HL χ^2	0.17 (0.14 - 0.20)	0.911 (0.905 - 0.916)	99.1 (46.8 - 151.4)
SVM-R	MSE	0.13 (0.08 - 0.17)	0.873 (0.874 - 0.883)	30.3 (27.9 - 32.6)
SVM-R	AUC	0.28 (0.19 - 0.36)	0.894 (0.888 - 0.900)	28.5 (26.0 - 30.9)
SVM-R	HL χ^2	4.53 (3.47 - 5.59)	0.901 (0.895 - 0.908)	16.4 (13.4 - 19.4)
SVM-P	MSE	3.09 (2.97 - 3.21)	0.912 (0.905 - 0.919)	34.1 (12.4 - 55.8)
SVM-P	AUC	2.65 (2.50 - 2.80)	0.915 (0.909 - 0.922)	33.5 (11.8 - 55.1)
SVM-P	HL χ^2	2.98 (2.64 - 3.31)	0.899 (0.891 - 0.907)	20.0 (15.2 - 24.8)

Table 2: Analysis of the test data by model method and cross validation optimization method. LR = backwards step-wise Logistic Regression model, SVM-R = radial-based kernel SVM, SVM-P = polynomial-based kernel SVM. Opt = Cross Validation optimization method. AUC = Area under the Receiver Operating Characteristic Curve. HL χ^2 = Hosmer-Lemeshow Goodness-of-Fit. MSE = mean squared error. For the model parameters, the values refer to the exclusion threshold (p-value) for LR, w for SVM-R, and p for SVM-P.

	LR (MSE)	LR (AUC)	LR (HL)	LR (0.10)	SVM-R (MSE)	SVM-R (AUC)	SVM-R (HL)	SVM-P (MSE)	SVM-P (AUC)	SVM-P (HL)
LR (MSE)	—	0.980	0.724	0.667	**<0.001** ←	**<0.001** ←	**0.014** ←	0.997	0.393	**0.012** ←
LR (AUC)	0.857	—	0.741	0.311	**<0.001** ←	**<0.001** ←	**0.014** ←	0.986	0.379	**0.012** ←
LR (HL)	0.773	0.901	—	0.175	**<0.001** ←	**<0.001** ←	**0.032** ←	0.756	0.239	**0.025** ←
LR (0.10)	0.725	0.858	0.958	—	**<0.001** ←	**<0.001** ←	**0.035** ←	0.715	0.214	**0.027** ←
SVM-R (MSE)	**0.002** ←	**0.009** ←	**0.010** ←	**0.007** ←	—	**<0.001** ↑	**<0.001** ↑	**<0.001** ↑	**<0.001** ↑	**<0.001** ↑
SVM-R (AUC)	**0.001** ←	**0.008** ←	**0.008** ←	**0.006** ←	0.301	—	**0.004** ↑	**<0.001** ↑	**<0.001** ↑	0.354
SVM-R (HL)	**<0.001** ←	**0.002** ←	**0.002** ←	**0.001** ←	**<0.001** ←	**<0.001** ←	—	**0.025** ↑	**0.002** ↑	0.677
SVM-P (MSE)	**0.011** ←	**0.025** ←	**0.024** ←	**0.019** ←	0.729	0.612	0.111	—	0.4378	**0.019** ←
SVM-P (AUC)	**0.010** ←	**0.024** ←	**0.022** ←	**0.018** ←	0.772	0.652	0.124	0.968	—	**0.002** ←
SVM-P (HL)	**<0.001** ←	**0.003** ←	**0.003** ←	**0.002** ←	**<0.001** ←	**0.002** ←	0.204	0.212	0.232	—

Table 3: Summary of *p* values for the pair-wise comparison between different versions of the modeling methods. The top right half of the table contains *p* values for AUC measurements, and the bottom left half of the table contains *p* values for the HL χ^2 measurements. LR = backwards step-wise Logistic Regression model, SVM-R = radial-based kernel SVM, SVM-P polynomial-based kernel SVM. AUC = Area under the Receiver Operating Characteristic Curve. HL = Hosmer-Lemeshow Goodness-of-Fit. MSE = mean squared error. Bold values are statistically significant. ↑ = Column-based model statistically superior. ← = Row-based model statistically superior.

Discussion

All of the models had excellent discriminatory (AUC) performance. This suggests that the clinical data collected in this domain is able to account for the majority of risk for in-hospital post-PCI mortality.

The LR models had superior discrimination to the SVM-R models, and also to the HL χ^2 optimization method of the SVM-P models. This shows that the clinical data are probably linearly separable, and is consistent with prior studies of logistic regression in this domain. In addition, the polynomial kernel was superior to the radial kernel in its ability to discriminate using the MSE and AUC methods.

The MSE method is a common optimization score in regression model development, and the AUC and HL χ^2 values were experimental optimization parameters. To our knowledge, there have been no reports on their use for optimizing variable selection. Discrimination in the LR models was insensitive to the optimization method. The SVM-R models showed a performance improvement with AUC over MSE, and with HL χ^2 over both AUC and MSE. The SVM-P showed a performance improvement with both the MSE and AUC methods over the HL χ^2 method.

Calibration was significantly higher in both of the SVM model types compared to the LR models. Different optimization methods did not significantly impact the LR models. However, the HL χ^2 optimization method provided significant improvement for the SVM-R models. None of the models achieved an upper 95% CI lower than 15.51, revealing that at least a portion of each model type failed to adequately calibrate. However, this is a common finding in other studies that evaluated risk models on a large volume of data sets. This happens because small numbers of models with large HL χ^2 values skew the distribution.

When a model is generated, a balance must be maintained between overfitting and underfitting. Overfitting improves the fit of the model on the training data but reduces its performance on external data. Underfitting generalizes the fit of the model by reducing the complexity of the fit on the training data in order to improve classification or regression results to other data sets. In these data, the AUC and MSE optimization methods significantly overfit the SVM-R models, as shown by the large difference between the mean width factors of HL χ^2 and the other two optimization strategies. In GIST, a decrease in the width factor is associated with an increase in the fit to the training data. However, the model parameters were not substantially different between optimization types for LR or SVM-P.

The primary limitations of this study are the lack of manual manipulation in the modeling process, although this was required by the methodology. In the single model development process, optimization in terms of model parameters, introduction of interaction terms, and feature selection would be manually performed by further optimizing a cross-validation score (MSE, AUC, etc.) if the discrimination was not deemed good or the calibration was inadequate in the training sample.

The parameter selection process used a 3-fold CV method, and the model evaluation used a separate testing sample over 100 randomized data sets. This is related to the nested stratified 10-fold CV method as described by Statnikov and colleagues.[85] The small number of training folds (or inner loop) were utilized because of high computational times of GIST in the relatively large data sets. This may have increased the variance of the scores in the parameter optimization methods, although the large number of data points in each fold likely minimized this problem.

Overall, both the polynomial and radial-based SVMs achieved better calibration than the LR models. Experimentation with different scoring methods used to select parameters for model generation revealed that the regular method of MSE scoring performed as well as the others for LR. However, the HL χ^2 method achieved the best results for both discrimination and calibration for the radial-based kernels, but improved calibration performance at the expense of discrimination for the polynomial-based kernels. This tradeoff requires further exploration.

Use of support vector machine risk models to promote adequate calibration and produce more accurate individualized prognostic estimates for patients is supported by this preliminary study. Future work will include investigating other calibration indices, recalibration methods, and further evaluation of this method for inclusion in an automated real-time safety monitoring system.

Chapter 3: Design and implementation of an Automated Real-Time Safety Monitoring Application

Background

Minimizing harm to patients and ensuring their safety are cornerstones of any clinical research effort. Safety monitoring is important in every stage of research related to a new drug, new medical device, or new therapeutic procedure. This type of monitoring of medical devices, under the auspices of the Food and Drug Administration (FDA), has undergone major changes over the last several decades.[1-4] These changes have largely been due to a small number of highly publicized adverse events.[5-13] The FDA's task is complex; the agency regulates more than 1,700 types of devices, 500,000 medical device models and 23,000 manufacturers.[3, 6, 17, 19, 20, 22, 63] In pre-marketing clinical trials, rare adverse events may not be discovered due to small sample sizes and biases towards healthier subjects.[14] The FDA must balance this concern with the need to deliver important medical advances to the public in a timely fashion. In response to this, the FDA has shifted some of its device evaluation to the post-market period, allowing new devices to reach the market sooner.[15] This creates the potential for large numbers of patients to be exposed to a new product in the absence of long-term follow-up data, and emphasizes the need for careful and thorough post-marketing surveillance.[16]

The current FDA policies in this area include a heterogeneous mix of voluntary and mandatory reporting.[1, 6, 14, 17-22] Voluntary reporting of adverse events creates limitations in significant event-rate recognition through underreporting bias, and highly variable reporting quality.[14] Several state and federal agencies have implemented mandatory reporting for medical devices for specific clinical areas, and national medical societies are making strides to standardize data element definitions and data collection methods within their respective domains.[23, 64, 65] Continued improvements in the quality and volume of reported data have created opportunities for timely and efficient analysis and reporting of alarming trends in patient outcomes.

Non-medical industries (Toronado, HGL Dynamics, Inc., Surrey, GU, WinTA, Tensor PLC, Great Yarmouth, NR) have been using a variety of automated statistical process control (SPC) techniques for quality control purposes for many years.[66-68] These systems rely on automated data collection, and use standard SPC methods of varying rigor. [24] However, automated SPC monitoring has not been widely deployed in the medical domain due to a number of constraints: (a) historically, automated data collection could usually only be obtained for objective data such as laboratory results and vital signs; (b) much of the needed information about a patient's condition is subjective and may be available only in free text in the medical record; and (c) medical source data, due to heterogeneity of clinical factors, typically has more noise than industrial data, and standard industrial SPC metrics may not be directly applicable to medical safety monitoring.

Within the medical domain, the most closely related clinical systems that have been developed to date are those in clinical trial monitoring for new pharmaceuticals. A variety of software solutions (Clinitrace, Phase Forward, Waltham, MA; Oracle Adverse Event Reporting System [AERS], Oracle, Red Shores, CA; Trialex, Meta-Xceed, Inc., Freemont, CA; Netregulus, Netregulus, Inc., Centennial, CO) have been created to monitor patient data relevant to the study trial. These systems rely on standard SPC methodologies, and can provide real-time data monitoring and analysis through internal data standardization and collection for the trial. However, the focus of these systems is on real-time data aggregation and reporting to the FDA.

The increasing availability of detailed electronic medical records and structured clinical outcomes data repositories may provide new opportunities to perform real-time surveillance and monitoring of adverse outcomes for new devices and therapeutics beyond the clinical trial environment. However, the specific monitoring methodologies that balance appropriate adverse event detection sensitivity and specificity remain unclear.

In response to this opportunity, we have developed the Data Extraction and Longitudinal Time Analysis (DELTA) system, and explored both standard and experimental statistical techniques for real-time safety monitoring. A clinical example was chosen to highlight the functionality of DELTA, and to provide an overview of its potential uses. Interventional cardiology was chosen because the domain has a national data field standard,[23] a recent increase in mandatory case reporting from state and federal agencies, and recent device safety concerns publicized by the FDA.

Methods

System General Requirements

The DELTA system was designed to provide real-time monitoring of clinical data during the course of evaluating a new medical device, medication, or intervention. The system was designed to satisfy five principal requirements. First, the system should accept a generic dataset, represented as a flat data table, to enable compatibility with the broadest possible range of sources. Second, the system should perform both prospective and retrospective analysis. Third, the system should support a variety of classical and experimental statistical methods to monitor trends in the data, configured as analytic modules within the system, allowing both unadjusted and risk-adjusted safety monitoring. In addition, the system should support different methodologies for alerting the user. Finally, DELTA should support an arbitrary number of simultaneous datasets, and an arbitrary number of ongoing analyses within each dataset. That is, DELTA should "track" multiple outcomes from multiple data sources simultaneously, thus making it possible for DELTA to serve as a single portal for safety monitoring for multiple simultaneous analyses in an institution.

Source Data and Internal Data Structure

A flat file representation of the covariates and clinical outcomes serves as the basis for all analyses. In addition, a static data dictionary must be provided to DELTA to allow for parsing and display of the source data in the user interface. Necessary information includes whether each field is going to be treated as a covariate, an outcome, and whether it is discrete or continuous.

The system uses a SQL 2000 server (Microsoft Corp., Redmond, WA) for internal data storage, importing all clinical data and data dictionaries from source databases at regular time intervals. This database also stores system configurations, analysis configurations, and results that are generated by DELTA at the conclusion of a given time period. The user interface is web-based, and uses a standard tree menu format for navigation. DELTA's infrastructure and external linkages are shown in Figure 1.

Security of patient data is currently addressed through record de-identification steps[69] performed to the fullest extent possible while maintaining the necessary dataset granularity for

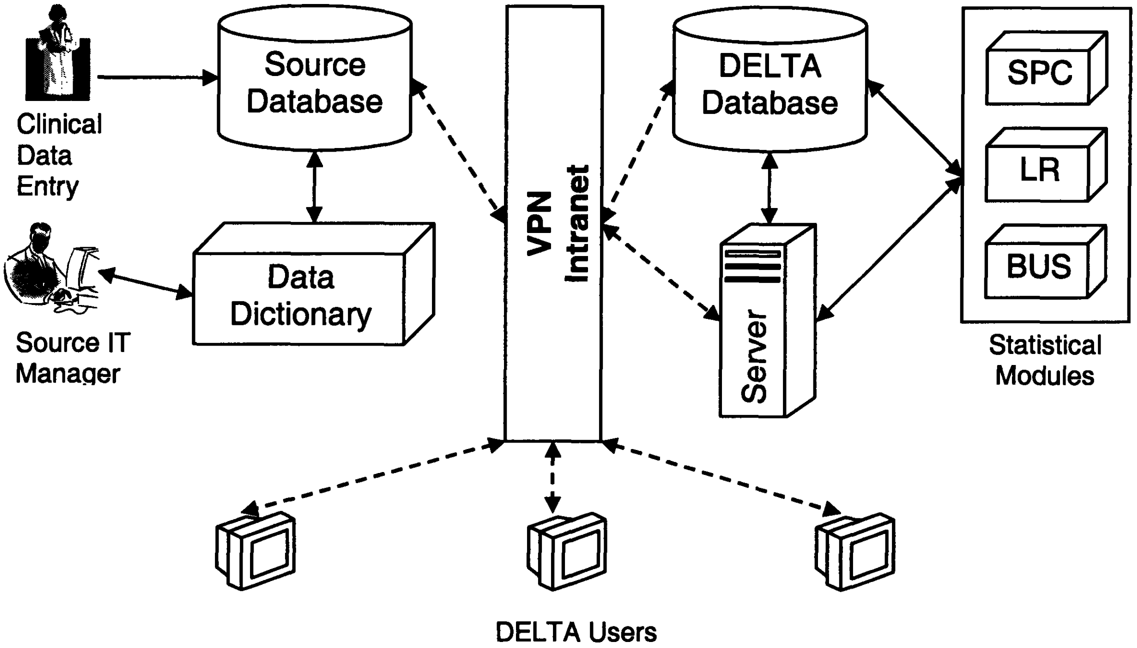


Figure 1: Overall DELTA infrastructure and an example external data source. SPC: Statistical Process Control; LR: Logistic Regression; BUS: Bayesian Updating Statistic.

the risk adjustment models. The system is hosted on the Partners Healthcare intranet, a secure multi-hospital network, accessible at member sites or remotely through VPN.

Statistical Methods

DELTA utilizes a modular approach to statistical analysis that facilitates further expansion. DELTA currently supports three statistical methodologies: statistical process control (SPC), logistic regression (LR), and Bayesian updating statistics (BUS). Discrete risk stratification is supported by both SPC and BUS. Periodic and cumulative analysis of data is supported by SPC and LR, and only cumulative analysis is supported by BUS.

Risk stratification is a process by which a given sample is subdivided into discrete groups based on predefined criteria. This process is used to allow providers to quickly estimate the probability of an outcome for a patient. Statistically, the goal of this process is to create a meaningful separation in the data to allow concurrent and potentially different analyses to be performed on each subset. Criteria are selected based on prior data, typically derived from a logistic regression predictive model, and the relative success of this stratification can be determined by a stepwise increase in the incidence of the outcome in each risk group. The LR method does not offer discrete risk stratification because it incorporates risk stratification on a case level.

Retrospective data analyses traditionally use the entire data set for all calculations. However, in real time data analysis, it is of interest to monitor both recent trends and overall trends in event rates. Evaluation of recent trends will intrinsically have reduced power, because of the reduced sample size, to detect true, significant shifts in event rates. However, such monitoring may serve as a very useful 'first warning' indicator when the cumulative event rate may not yet cross the alerting threshold. This type of alert is not considered definitive, but can be used to encourage increased monitoring of the intervention of interest and heighten awareness of a potential problem. In DELTA, these recent data analyses are termed 'periodic', and can be configured to be performed on a monthly, quarterly, or yearly basis.

SPC is a standard quality control method in non-medical industrial domains. This method compares observed event rates to static alerting boundaries developed from previously published or observed empirical data. Each industry typically requires different levels of rigor in alerting, and selection of confidence intervals (or number of standard errors) establishes this benchmark. In the medical industry, the 95% confidence interval (CI) is considered to be the threshold of statistical improbability to establish a 'true' difference. In DELTA, the 95% CI of proportions by the Wilson method is used to calculate the alerting boundaries for all statistical methods.[70] The proportion of observed events are then compared to these static boundaries, and alerts are generated if they exceed the upper CI boundary. DELTA's SPC module is capable of performing event rate monitoring on multiple risk strata provided that criteria for stratification and benchmark event rates are included for each risk stratum. This method supports comparison of benchmark expected event rates with cumulative and periodic observed event rates.

While simple and intuitive, the SPC methodology does not support case-level risk adjustment. It is also dependent on accurate benchmark data, which may be limited for new procedures or when existing therapies are applied to new clinical conditions.[70]

Logistic regression[71] is a non-linear modeling technique used to provide a probability of an outcome on case-level basis. Within DELTA, the LR method allows for continuous risk-adjusted estimation of an outcome at the case level. The LR model must be developed prior to the initiation of an analysis within DELTA, and is mostly commonly based on previously published and validated models.

Alerting thresholds are established by using the LR model's expected mortality probability for each case. These probabilities are then summated in both periodic and cumulative time frames to determine the 95% confidence interval (CI) of the event rate proportion by the Wilson method. Alerts are generated if the observed event rate exceeds the upper bounds of the 95% CI of a given boundary. This method provides accommodation for high risk patients by adjusting the alerting boundary based on the model's expected probability of death. This is can be very useful in when outcome event rates vary widely with patient co-morbidities. A limitation of this method is that the alerts become dependent on the discrimination (measure of population prediction accuracy) and calibration (measure of small group or case prediction accuracy) of that model.

BUS is an experimental methodology pioneered in non-healthcare industries.[72] This method incorporates Bayes' theorem[73] into a traditional SPC framework by utilizing prior observed data to evolve the estimates of risk. Alerting boundaries are calculated by two methods, both of which are considered cumulative analyses only. The first method includes previous current study data with the prior data used in the SPC method to calculate the 95% CI of the event rate proportion by the Wilson method. This means that the alerting boundary shifts during the course of real-time monitoring due to the influence of the earlier study data.

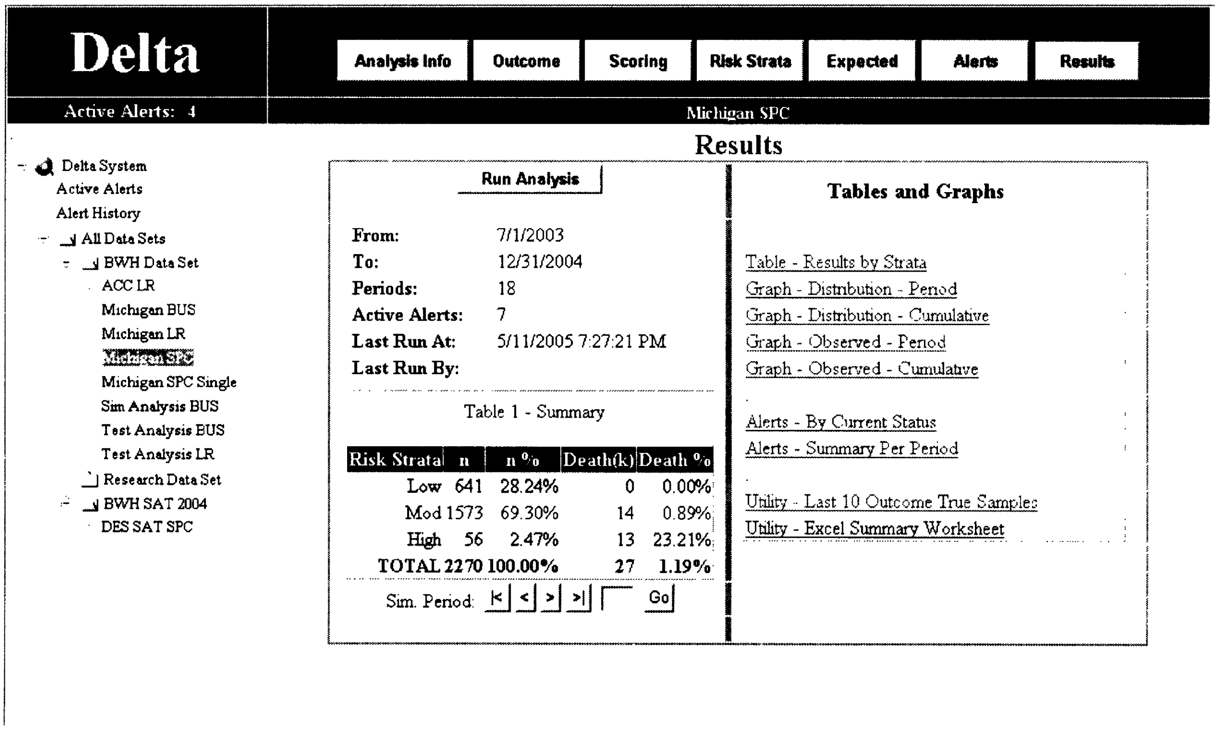
The other alerting method is based on the evolution of the updated risk estimates represented as probability density functions (PDF). In each period, a new PDF is generated based on the cumulative study event rate and baseline event rate. Alerting thresholds are generated by the user specifying minimum percent amount of overlap of the two distributions (by comparison of central posterior intervals).[74] The first comparison PDF is the initial prior PDF, and the second is the previous period's PDF. BUS supports discrete risk stratification.

This method was included in DELTA because it tends to reduce the impact of early outliers in data and complements the other monitoring methods used in the system. It also may be particularly helpful in situations in which limited pre-existing data exist. However, the method is dependent on accurate risk strata development, and on the methods used for weighting of the prior data in the analysis.

User Interface

The user interface is provided via a web browser and was developed in the Microsoft .NET environment, running Microsoft IIS 5.0 Web Server (Microsoft Corp., Redmond, WA). Each data set is represented as a separate folder on the main page, and all analyses for that set are nested under that folder (see Figure 2). At the initiation of an analysis, the user designates the analysis period and starting and stopping dates, selects the statistical module, and selects the outcome of interest. Data filters can be applied to restrict the candidate cases for analysis. Covariates used for risk stratification are selected. Lastly, periodic and cumulative alerts for the statistical method selected can be activated or suppressed based on user preferences. An analysis configuration can be duplicated and modified for convenience in configuring multiple statistical methods to concurrently monitor a data source.

The results screen of DELTA serves as the primary portal to all tables, alerts, and graphs generated from an analysis. Tabular and graphical outputs of the data and specific alerting thresholds by risk strata are available, and an export function is included to allow researchers to perform further evaluation of the data.



From: 7/1/2003
To: 12/31/2004
Periods: 18
Active Alerts: 7
Last Run At: 5/11/2005 7:27:21 PM
Last Run By:

Table 1 - Summary

Risk Strata	n	n %	Death(k)	Death %
Low	641	28.24%	0	0.00%
Mod	1573	69.30%	14	0.89%
High	56	2.47%	13	23.21%
TOTAL	**2270**	**100.00%**	**27**	**1.19%**

Figure 2: DELTA Screenshot showing the results menu screen of the SPC clinical example described in Section 4. The main menu is displayed on the left, and the analysis menu is displayed above the viewing area.

Clinical Example

As an example of the application of DELTA to real-world data, an analysis of the in-hospital mortality following the implantation of a drug-eluting stent was performed. The cardiac catheterization laboratory of Brigham and Women's Hospital has maintained a detailed clinical outcomes database since 1997 for all patients undergoing percutaneous coronary intervention, based on the American College of Cardiology National Cardiovascular Data Repository (ACC-NCDR) data elements.[39]

For risk stratification, the University of Michigan risk prediction model[33] was used since it provides a concise method of comparing all three of DELTA's statistical methods using one reference for prior experience. The previous experience of event rates for all risk strata from this work is listed in Appendix A. A logistic regression model with risk stratification scores are listed in Appendix B. The logistic regression model developed from the data was used to create a discrete risk scoring model. Based on the mortality of patients in the study sample at various risk scores, these data were divided into three discrete risk categories, and the compositions of those categories are listed in Appendix B.

A total of 2,270 drug-eluting stent (DES) cases were performed from July 01, 2003, to December 31, 2004, at our institution, and the outcome in terms of in-hospital mortality was analyzed. These data were retrospectively evaluated in monthly periods for each of the three statistical methodologies. There were a total of 27 observed deaths (unadjusted mortality rate of 1.19%) during the study. Local institutional IRB approval was obtained. Risk stratification of these cases by the University of Michigan model is listed in Table 1, and demonstrates increasing in-hospital mortality risk with 0%, 0.9%, and 23% mortality risk in the low, medium, and high risk strata respectively.

An alternative data set was generated by taking the clinical data above and changing the procedure date from the 8 cases with the outcome of interest in the last 5 periods. The procedures dates were changed by random allocation into one of the first 13 periods. The duration of the monitoring was then shortened to 13 periods. This was done to illustrate alerts when cumulative event rates clearly exceeded established thresholds. The overall event rate for this data set is 1.71% (27/1583), and the risk stratified event rate was 0% (0/446), 1.3% (14/1095), and 31%(13/42) for the low, medium, and high risk strata, respectively.

Risk Strata	Sample	Events	Event Rate
Low	641	0	00.00%
Mod	1573	14	00.89%
High	56	13	23.21%

Table 1: Multiple Risk Strata SPC.

Results

Statistical Process Control

The single risk stratum SPC was configured with no risk stratification covariates. The static alert boundary was a 2.07% (upper 95% CI of 100/5863). Periodic evaluations ranged from 0% to 4.5%. Period 4 exceeded the boundary with a 3.4% (5/148) event rate, and period 10 with a 4.5% (5/110) event rate. Cumulative event rates ranged from 0.9% (2/213) to 1.7% (10/587). No cumulative evaluations had an event rate that exceeded the boundary. The cumulative evaluation is depicted graphically in Figure 3.

Periodic evaluations of the alternative data set ranged from 0% to 4.7%. Period 4 exceeded the boundary with a 4.7% (7/150) event rate, period 7 with a 2.6% (3/117) event rate, period 8 with a 2.6% (3/117) event rate, and period 10 with a 4.5% (5/110) event rate. Cumulative event rates ranged from 0.9% (2/213) to 2.4%(12/490). Periods 4 through 11 had event rates exceeding the 2.07% threshold and generated alerts, and ranged from 2.1% to 2.4%.

Alerting thresholds were calculated for the low, medium, and high risk strata by using the upper 95% CI of the proportion of the event rates of each stratum in the University of Michigan data. The thresholds were 0.3% (1/1820), 1.7% (50/3907), and 44% (49/136), respectively.

There were no events in the low-risk stratum, and no alerts were generated. In the moderate-risk stratum, the periodic observed event rates ranged from 0% to 2.7%. The alerting boundary was exceeded with rates of 2.7% (2/75) in period 5, 2.6% (2/78) in period 10, and 1.9% (2/108) in period 18. The cumulative observed event rates ranged from 0.7% to 1.3%, and never exceeded the upper alert boundary. In the high-risk stratum, the periodic observed event rates ranged from 0% to 100%. The alerting boundary was exceeded with rates of 100% in periods 1 (1/1), 7 (1/1), and 10 (3/3), and by a rate of 50% (4/8) in period 4. The cumulative observed event rates ranged from 16.7% to 100%. The alerting boundary was exceeded by a rate of 100% (1/1) in period 1.

Evaluation of the alternative data set was performed periodic and cumulative alerts. There were no events in the low-risk stratum, and no alerts were generated. In the moderate-risk stratum, the periodic observed event rates ranged from 0% to 3.6%. The alerting boundary was exceeded with rates of 3.6% (3/84) in period 3, 2.1% (2/95) in period 4, 2.7% (2/75) in period 5, 2.5% (2/81) in period 7, and 2.6% (2/78) in period 10. The cumulative observed event rates ranged from 0.7% to 2.0%, and exceeded the alerting boundary in periods 3 through 11. In the high-risk stratum, the periodic event rates ranged from 0% to 100%. The alerting boundary was exceeded with rates of 100% in periods 1 (1/1), 7 (1/), and 10 (3/3), and by a rate of 55.6% (5/9) in period 4. The cumulative observed event rates ranged from 16.7% to 100%. The alerting boundary was exceeded by a rate of 100% (1/1) in period 1.

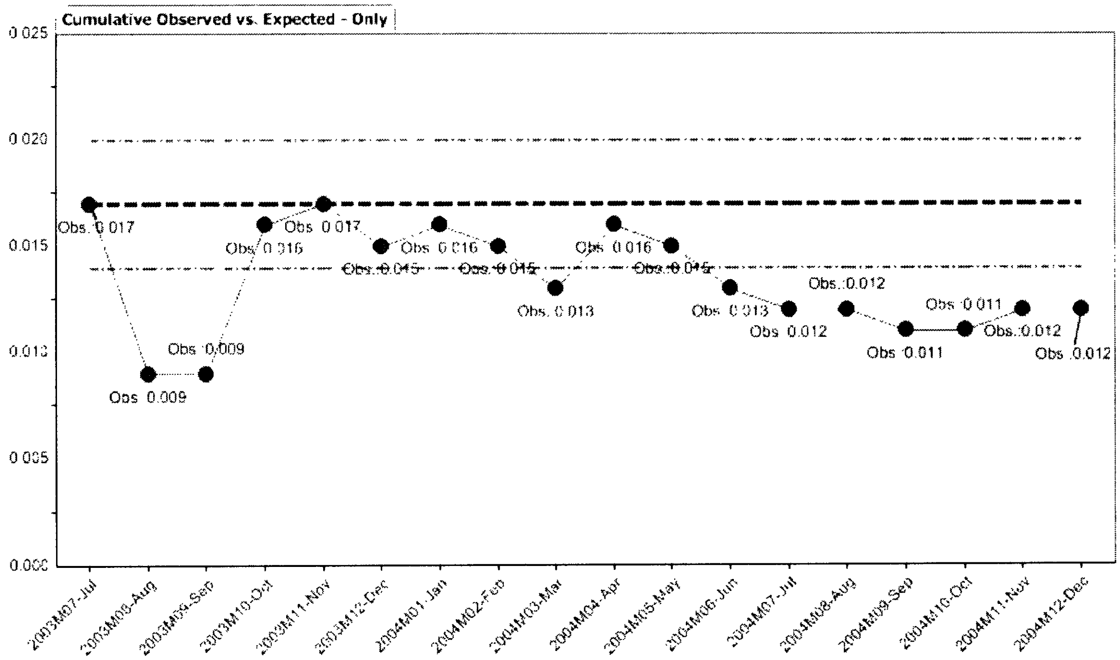


Figure 3: Single-stratum SPC graph showing the cumulative observed event rates versus the static alerting threshold (expected rates) with 95% confidence intervals.

Logistic Regression

Alerting thresholds were calculated on a periodic basis using the expected probability of death for the cases in their respective periods, and the 95% upper CI ranged from 4.9% (1.02/112) to 7.1% (2.51/110). Cumulative-based upper alerting boundaries ranged from 2.3% (29.86/1835) to 5.7% (1.66/115). The overall expected cumulative event rate was 1.75% (39.7/2270).

Periodic event rates ranged from 0% to 4.5%, and no alerts were generated. The two highest periodic event rates of 3.4% (5/148) in period 4 and 4.5% (5/110) in period 10 had upper alerting boundaries of 5.8% (2.98/148) and 7.1% (2.51/110), respectively. Cumulative event rates ranged from a 0.9% (2/213) to 1.7% (10/587) event rate, and the cumulative upper 95% CI was well above the observed event rate throughout the evaluation and are shown in Figure 4.

Alerting thresholds for the alternative data set based on the upper 95% CI ranged from 4.9% (1.02/112) to 7.5% (3.2/117) in the periodic analysis, and from 2.6% (28.17/1583) to 5.7% (1.66/115) in the cumulative analysis. The overall expected event rate was 1.78% (28.17/1583).

Periodic event rates ranged from 0% to 4.7%, and no alerts were generated. The two highest periodic event rates of 4.7% (7/150) in period 4 and 4.5% (5/110) in period 10 had upper alerting boundaries of 6.2% (3.56/150) and 5.1% (2.51/110). Cumulative event rates ranged from 0.9% (2/213) to 2.4% (12/490), and were well below the alert boundaries through the evaluation.

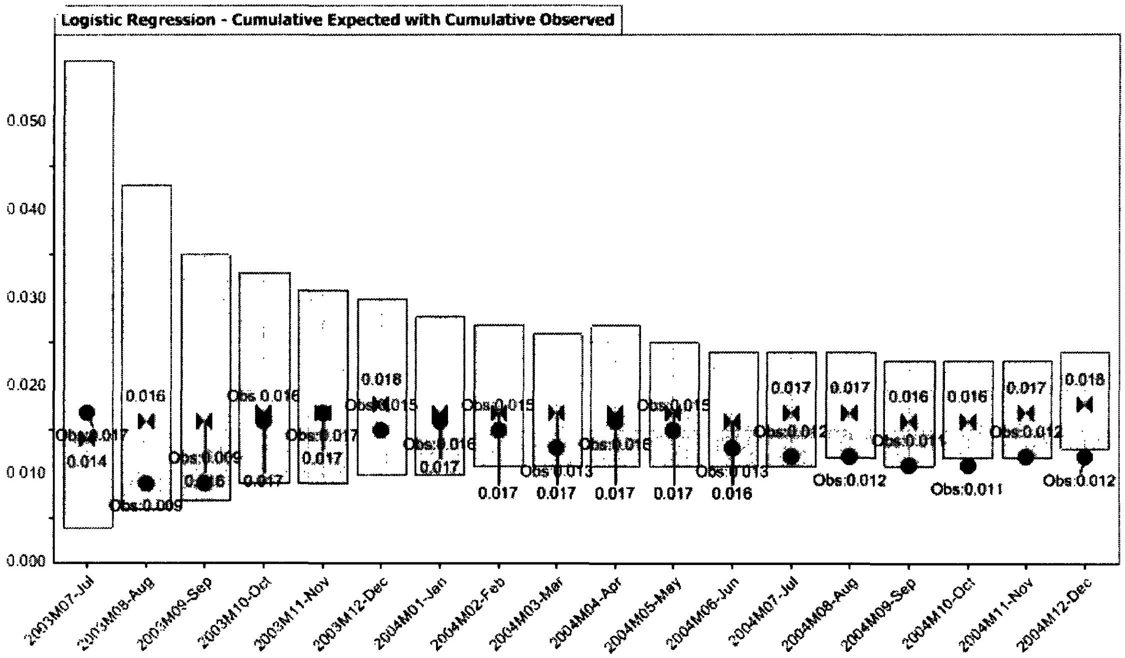


Figure 4: Logistic Regression graph showing the cumulative observed event rate versus the cumulative expected event rate with 95% CI.

Bayesian Updating Statistics

The upper alert boundary varied from 0.2% to 0.3%, from 1.5% to 1.7%, and from 40.2% to 44.4% in the low, medium, and high risk strata, respectively. From all strata, the only alert generated was in the high risk stratum period 1 with an observed event rate of 100% and an upper alert boundary of 44.4% (49/136).

There was a trend towards lower event rates in the PDFs of all risk strata, as illustrated for the high-risk stratum cases in Figure 5. At no time in any strata did the posterior confidence interval overlap fall below the user-specified 80% criteria.

The upper alert boundaries in the alternative data set were the same as the real data set. However, in the moderate-risk stratum, the observed event rates of 1.73% (4/231) in period 3, 1.84% (6/325) in period 4, 2.0% (8/400) in period 5, 1.73% (10/576) in period 7, and 1.71% (14/818) in period 10% exceeded the alert boundaries that ranged from 1.65% to 1.74% for those periods.

The trend towards lower event rates in the PDFs of all risk strata in the real data set was not found in the alternative data set. At no time in any strata did the posterior confidence interval overlap fall below the user-specified 80% criteria.

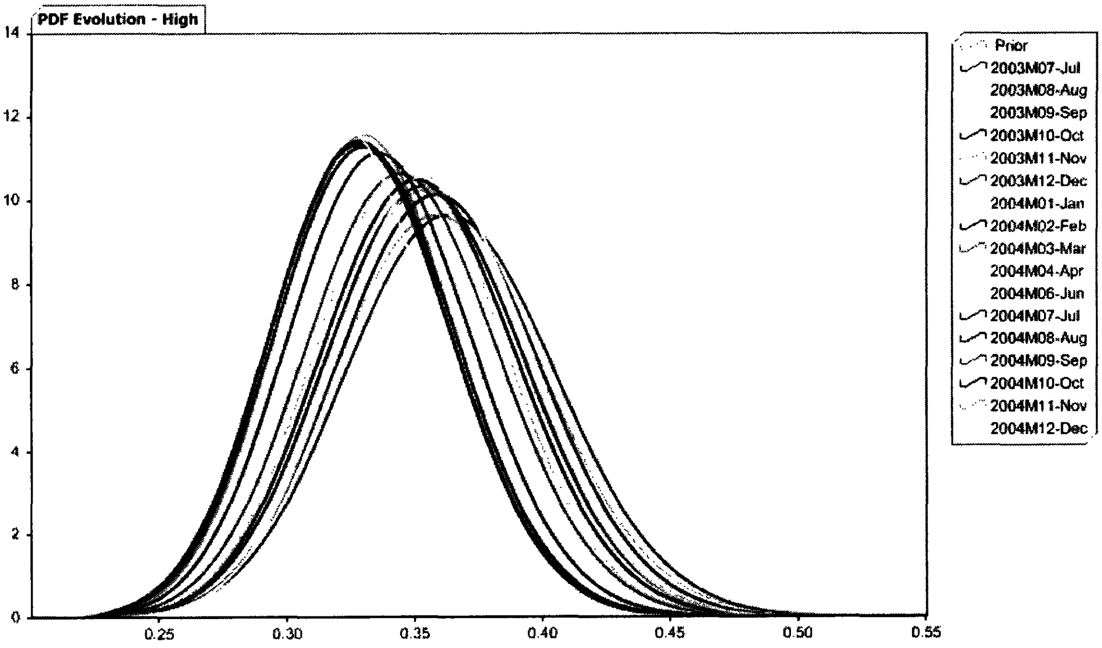


Figure 5: Bayesian Updating Statistics Probability Density Functin (BUS PDF) evolution for high-risk cases, by period.

Discussion

The DELTA system satisfied all pre-specified design requirements, and performed all analyses and graphical renderings within 2 seconds each on the hospital intranet.

The SPC method triggered periodic alerts in both single and multiple risk strata analyses. This method also triggered the first period's cumulative alert in the multiple risk strata, but this can be considered a periodic equivalent alert. Otherwise, there were no cumulative event rate alerts detected by the SPC method. The LR method generated no alarms in either the periodic or cumulative evaluations. The BUS method generated an alert only in the first period of the high-risk stratum. While all BUS alerts are considered cumulative, the alert was generated from one case with a positive outcome for that period.

The alternative data set event rate was elevated manually to generate alarms. The single stratum SPC method alerted to event rates exceeding the threshold for periods 4 through 11. The multi-strata SPC method revealed that the event rate rise of concern was in the moderate risk group, alerting from periods 3 through 11. The LR method generated no periodic or cumulative alerts in the alternative data set. The BUS method agreed that the elevation was primarily of concern in the moderate risk stratum by generating cumulative alerts in periods 3, 4, 5, 7, and 10.

Periodic alerts are very sensitive measures of elevated event rates, but generally lack the statistical power to make a conclusive decision about the safety of a device. These alerts would serve to heighten surveillance and possibly reduce the interval of evaluation for the new device, but would not in of themselves be sufficient to recommend withdrawal of the device. The discrepancy between SPC and LR periodic alerting was because LR attempts to adjust the alerting threshold based on the expected outcome of a given case. If there were a rise in the event rate for a period, SPC would trigger an alert as the rate exceeds the static threshold. However, if the LR model expected the cases to have that outcome, then the method would likely not alert as the alert threshold would be adjusted based on that expectation. The cumulative alerts for this analysis were consistent across statistical methods, and the alerts in the first period were due to a very low number of examined cases. In the alternative set, the LR method has no cumulative alerts, and this could be due to the fact that the events were expected by the model.

In phase 3 randomized controlled trials, there is no previous data to use as a benchmark, and a common method of determining the threshold of stopping the trial is to initially place the threshold at a very statistically improbably number (such as 5 or 6 standard errors from an estimated allowable rate) and gradually reduce the allowable error as the volume of data grows. The allowable rates are generally established by expert consensus and are manually generated on a trial by trial basis.

The benefits of incorporating prior information into the development of alerting thresholds include the ability to develop and establish explicit rules for alerting thresholds. This methodology could then be applied in an objective manner to a wide variety of monitoring applications. This removes the need for an expert consensus to develop the thresholds. However, this objective methodology has limitations. The accuracy of the alerting boundaries is dependent on the source data. In the case of this clinical example, the University of Michigan BMS mortality data and model was established as the benchmark. DELTA then considered mortality event rates statistically significantly above that baseline to be abnormal and of concern. This becomes important when assessing the external validity of the benchmark data with regards to applicability in a different patient population. In addition, applying multiple concurrent statistical methodologies to a monitoring process is meant to guard against specific vulnerabilities one methodology might have to these type of confounding.

SPC is only concerned with the overall event rate in the benchmark source population to establish alerting boundaries, and these are static throughout the analysis. This is the least sensitive to subpopulation variations between the study and baseline populations. Including multiple risk strata in the analysis increases the sensitivity to finding problems in a specific risk group but requires the user to ensure that the study subpopulations using the risk stratification criteria are representative of the source subpopulations. Similar proportions and relative event rate risks between the source data and study data supports the use of stratification in this clinical example.

LR is the most susceptible method to population differences because it provides a case level estimation based on a number of risk factors. In a number of studies these models suffer degraded predictive ability at the case level in disparate populations and as the time from the model's development increases.[74] In the example, the population's event rate was 1.19% and the LR model's expected event rate was 1.75%. This shows that the LR model over-predicted mortality for this population.

BUS carries many of the same benefits and drawbacks of using the aggregate source population's event rate to establish alerting thresholds, but allows for the movement of these thresholds by changing study event rates. This method is the most capable in determining a significant shift in a short period of time.

Overall, the results of the example analysis support that the in-hospital mortality following implantation of DES was acceptably low over the time period studied when compared with the University of Michigan BMS benchmark data. The prototype system currently in use at Brigham and Women's Hospital Cardiac Catheterization Laboratory is in a testing and evaluation phase, and as such, clinicians do not consult the system directly. An evaluation of the current user interface will be conducted to assess DELTA's acceptability in the clinical environment by different health care providers. However, the preliminary results of our testing are encouraging: the DELTA system shows promise in filling a need for automated real-time safety monitoring in the medical domain, and may be applicable to routine safety monitoring for hospital quality assurance, and monitoring of new drugs and devices.

Risk Strata	Risk Score	Sample	Deaths	Death %	Upper 95% CI
Low	0-1.49	1820	1	0.015	0.03
Moderate	1.5-5.49	3907	50	1.28	0.017
High	5.50+	136	49	36.0	0.443
TOTAL		5863	100	1.71	0.0207

Appendix A: Summary of the sample population and outcome of interest (death) per risk strata in the University of Michigan data sample.[46]

Covariate	LR β	Odds Ratio	Risk Score
MI within 24 hours	1.03	2.8	1
Cardiogenic Shock	2.44	11.5	2.5
Creatinine > 1.5 mg/dL	1.70	5.5	1.5
History of Cardiac Arrest	1.29	3.65	1.5
Number of Diseased Vessels	0.44	1.54	0.5
Age 70-79	0.81	2.24	
Age >= 80	0.97	2.65	
Age >= 70			1.0
LV Ejection Fraction <50%	0.51	1.66	0.5
Thrombus	0.52	1.67	0.5
Peripheral Vascular Disease	0.46	1.57	0.5
Female Sex	0.59	1.82	0.5
Intercept	-7.20		

Appendix B: University of Michigan Covariates with Beta Coefficients for the logistic regression model and Risk Scores for the discrete risk stratification.[46] Intercept is the LR model equation intercept.

Chapter 4: Sensitivity Analysis of SPC and Bus Alerting Thresholds

Background

In a prior work,[75] we established the feasibility of a real-time automated monitoring system, and evaluated implementations of Statistical Process Control (SPC)[76] and Bayesian Updating Statistics (BUS)[77] for this application on actual clinical data in the domain of Interventional Cardiology. However, sensitivity analyses to determine the independent and relative performance of these methodologies has not been established, and is required for the subsequent use of this tool.

The purpose of this study is to compare alerting thresholds for SPC and BUS methods using local Interventional Cardiology clinical data.

Methods

Study Setting

Brigham & Women's Hospital (BWH) is a 720 bed academic teaching hospital in Boston, Massachusetts. BWH's cardiac catheterization laboratory has maintained a detailed clinical outcomes database since 1997 for all patients undergoing percutaneous coronary intervention (PCI). The database is compliant with the domain's national data element standard, based on the American College of Cardiology National Cardiovascular Data Repository (ACC-NCDR) guidelines, and provides detailed mandatory quarterly reports to the state (MASS-DAQ).

Subjects

All angioplasty procedures performed between January 01, 2002 and December 31, 2004 were selected for inclusion in this study. The outcome of interest was a major adverse cardiac event (MACE). This is an aggregate outcome consisting of death, post-procedural myocardial infarction, or a repeat vascularization (PCI or bypass surgery). The actual overall outcome event rate for the sample was 6.5% (403/6175). This study was approved by the Partners Institutional Review Board.

Alerting Thresholds

The Wilson method[70] for the 95% confidence interval of a proportion was selected to generate the appropriate alerting thresholds. The equation is shown here:

$$\frac{2np + z^2 \pm z\sqrt{(z^2 + 4npq)}}{2(n + z^2)}$$

where p is the proportion, q is $1 - p$, z is Standard Normal deviate associated with the two-tailed probability α, and n is the sample size. The SPC alerting method uses the Wilson method to generate a static alerting threshold based on the expected number of events and cases.

The BUS method, however, incorporates past observed events and number of cases into the baseline prior expected events and cases to generate a posterior alerting threshold for each new time period. This updating process uses the beta inverse function on the updated baseline to find the 97.5% tail of the central posterior distribution.[77]

Observed and Expected Event Rate Simulation

A range of identical baseline alerting thresholds were generated for both SPC and BUS methods. These expected event rates were constructed by varying both the overall event rate and the number of cases (denominator). The overall event rates were manually set to be 0.005, 0.05, or 0.5, and a range of magnitudes (statistical power) for each of these event rates was established by varying the denominator by starting with 10 an doubling up to 1,000,000. In addition, 100, 1000, 10000, 100000, and 1000000 were evaluated to provide "rounded" denominators for method comparisons. Using this method, 21 levels of denominators were generated for each overall event rate (10, 21, 44, etc.).

A variety of overall sample observed event rates were needed to perform the evaluation. In order to preserve the periodic outcome event rate fluctuations, a risk model to predict the likelihood of MACE in the clinical data was developed. All variables that were significantly associated with the outcome from prior studies were selected for inclusion. A logistic regression model was developed using a backwards stepwise technique (threshold 0.15) with SAS 9.1 (Cary, NC).[71] Discrimination of the model was measured by the Area Under the Receiving Operating Characteristic (AUC)[51] and was 0.662. This is somewhat expected because of the composite end-point, and a prior LR model for MACE at this institution had an ROC of 0.74.[56] Calibration of the model was adequate using the Hosmer-Lemeshow goodness-of-fit test.[78] The Chi2 was 5.662 with 8 degrees of freedom, and p=0.6851. External validity was not assessed because the model was intended for use only with the development data. The model is shown in Table 1.

The probability of MACE for each case in the data set was generated with the resulting LR model. Overall data set event rates were set by applying a probability cutoff to enforce the appropriate proportion of cases to be positive outcomes. All cases with a probability above this cutoff received a simulated outcome of 1, and all those below received an outcome of 0.

Two sets of observed event rates were generated for each of 21 denominator levels of three expected event rates (0.5, 0.05, and 0.005). The "fine" variation in observed event rates varied from +/- 10% by 1% increments of the expected event rate. For example, for the 50% (0.5) event rate, the "fine" observed event rate variation ranged from 0.4 to 0.6 in increments of 0.01 (0.4, 0.41, 0.42, etc.). The "coarse" variation in observed event rates varied from +/- 90% by 10% increments of the expected event rate. For example, in the 5% (0.05) expected event rate, the observed event rates varied from 0.005 to 0.095 (0.05, 0.1, 0.15, etc.).

Parameter	β Coefficient	Standard Error	*p*
Intercept	-5.4657	0.4769	<0.001
IABP	0.6362	0.1080	<0.001
Age (in years)	0.0183	0.0046	<0.001
PCI Case	1.1344	0.3677	0.002
Shock on Presentation	0.9055	0.2647	0.001
Diabetes	0.2724	0.1133	0.016
Chronic Renal Insufficiency	0.3845	0.1990	0.053
Left Anterior Descending 70% Block	0.3043	0.1120	0.007
Emergent Case	0.5709	0.1422	<0.001
Salvage Case	1.8571	0.4160	<0.001
Prior Myocardial Infarction	-0.2403	0.1250	0.055

Table 1: Logistic Regression model for the outcome of MACE in the clinical data.

Statistical Analysis

The simulation protocol produced a total of 120 data sets (19 "coarse" and 21 "fine" for each expected event rate). The data was analyzed using both the SPC and the BUS methods to generate a number of alerts fired by each method over the first 24 of 36 months. The final cumulative 36th month data was used to determine whether or not the alarm should have fired (gold standard). No assumption was made that one method was superior to the other, so the number of alarms of each method was measured with respect to the final 36th month alarm status of each method. Analysis was performed with Analyze-It (Version 1.73, Leeds, England) which uses the non-parametric method of ROC calculation described by Beck and colleagues.[79]

Results

The results of the "coarse" analysis of SPC and BUS alerting methodologies for the 0.5% expected event rate showed an AUC of 0.999 and 1.000 for SPC and BUS, respectively. These results are shown in Table 2. Cross-analysis also showed AUCs of 0.999 for both methods. The excellent AUCs were retained for each method in the 5% expected event rate with 1.000 and 0.999 for SPC and BUS, respectively. The BUS alerts and the SPC final alert showed good discrimination at 0.997. While the difference was not statistically significant, the SPC alerts compared to the BUS final alerts showed a relative decrease in AUC with 0.978. Finally, for the 50% expected event rate, significantly decreased AUCs are noted for all analyses. The SPC and BUS methods showed AUCs of 0.950 and 0.954, respectively. The SPC alerts with the BUS final alerts showed an AUC of 0.881, and the BUS alerts with the SPC final alerts showed an AUC of 0.939. The performance of the SPC with the BUS final alert was significantly worse than the BUS with the SPC final alert and confirmed a trend suggested in the 5% expected event rate.

Alarm Type	True Alert Type	Base Rate (%)	AUC
SPC	SPC	50	0.950 (0.916 - 0.984)
SPC	BUS	50	0.881 (0.833 - 0.928)
BUS	BUS	50	0.954 (0.924 - 0.984)
BUS	SPC	50	0.939 (0.902 - 0.975)
SPC	SPC	5	1.000 (1.000 - 1.000)
SPC	BUS	5	0.978 (0.957 - 0.998)
BUS	BUS	5	0.999 (0.998 - 1.000)
BUS	SPC	5	0.997 (0.992 - 1.000)
SPC	SPC	0.5	0.999 (0.996 - 1.000)
SPC	BUS	0.5	0.999 (0.997 - 1.000)
BUS	BUS	0.5	1.000 (1.000 - 1.000)
BUS	SPC	0.5	0.999 (0.997 - 1.000)

Table 2: Comparison of AUCs for each of the three baseline event rates (0.5%, 5%, and 50%) and "broadly" varied observed rates. between the number of alerts for SPC and BUS (Alarm Type) and the final 36 month alarm status for SPC and BUS (True Alert Type).

When both methods were evaluated in an identical manner with the "fine" data sets to determine performance and differences within smaller magnitude event rate deviations, the trends of the "coarse" data were preserved, and are shown in Table 3. All analyses and cross-analyses for the SPC and BUS methods in the 0.5% and 5% expected event rates ranged from 0.997 to 1.000 and were not significantly different from one another. However, the 50% expected event rate, similar to the "coarse" data sets, showed significant decreases in AUCs. The SPC and BUS AUCs were 0.957 and 0.960, respectively. Cross-analysis of the SPC and BUS methods showed AUCs of 0.916 and 0.959, respectively. The SUC to BUS AUC was significantly lower than any of the others, and was consistent with the "coarse" data analysis.

Alarm Type	True Alert Type	Base Rate (%)	AUC
SPC	SPC	50	0.957 (0.938 - 0.976)
SPC	BUS	50	0.916 (0.891 - 0.942)
BUS	BUS	50	0.960 (0.943 - 0.978)
BUS	SPC	50	0.959 (0.941 - 0.977)
SPC	SPC	5	1.000 (1.000 - 1.000)
SPC	BUS	5	0.999 (0.998 - 1.000)
BUS	BUS	5	1.000 (0.999 - 1.000)
BUS	SPC	5	0.998 (0.996 - 1.000)
SPC	SPC	0.5	0.994 (0.990 - 0.999)
SPC	BUS	0.5	0.997 (0.993 - 1.000)
BUS	BUS	0.5	0.997 (0.993 - 1.000)
BUS	SPC	0.5	0.997 (0.993 - 1.000)

Table 3: Comparison of AUCs for each of the three baseline event rates (0.5%, 5%, and 50%) and "narrowly" varied observed rates. between the number of alerts for SPC and BUS (Alarm Type) and the final 36 month alarm status for SPC and BUS (True Alert Type).

In order to evaluate the effect the denominator (sample size) has on the alerting thresholds of each method, the sample size for the expected event rate of 50% was varied from 44 to 1787 for SPC and BUS (Figure 1). This shows that the denominator size plays a large role in determining the number of alerts that should fire for any particular variation between the expected and observed event rates. This is expected, as the size of the expected data improves the confidence in that event rate value. An important trend to note is that for any given denominator, the SPC method fires more frequently than the BUS method, but this difference is reduced as the denominator sizes increase.

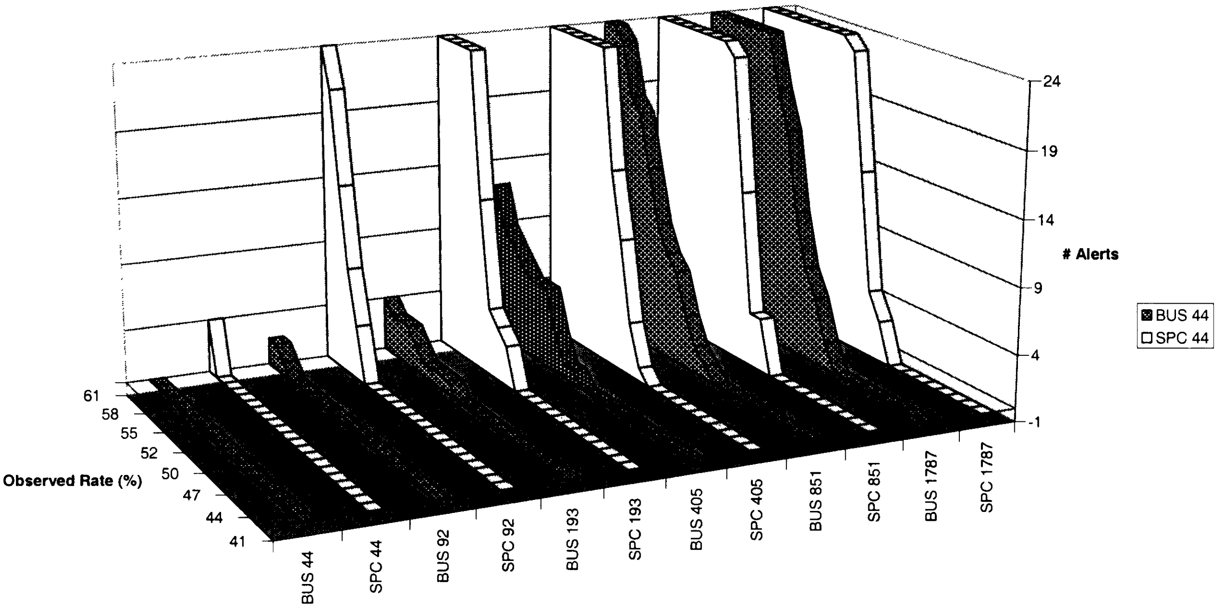


Figure 1: Illustration of Alert Firing for 50% Expected Event Rate of SPC and BUS from Expected sample sizes of 44 to 1878.

Discussion

There are a number of notable trends in the SPC and BUS alerting methodologies when applied to this clinical data. First, both SPC and BUS are able to excellently discriminate a 'true' alert in the first 24 months of data where the 36th month is the gold standard. This suggests that it is functioning within the statistical framework as an early detection system.

There is also a trend of decreasing performance as the expected event rate decreases. This is most notable in the 50% expected event rate category. This degradation of performance suggests that increasing event rates likely create greater variance in the monthly event rate which could generate more alerts without the setting of a 'true alert' (cumulative month 36).

In addition, when each method is compared to the other, the SPC number of alerts shows greater disagreement with the BUS 36th month alert than the BUS number of alerts and the SPC 36th month alert. Incorporation of prior month data into the BUS estimate creates a moving baseline which accounts for the differences, but further exploration of this is necessary. Both cross-analyses show generally lower AUC which is expected.

The primary limitation of this analysis is the LR modeling required to generate a probability of outcome to allow observed event rate scaling. The ROC was relatively low, although the calibration was adequate. This could impact the monthly event rate to be lower or higher than observed, and could subsequently have an effect on the number of alerts fired for a particular data set. As noted above, this is primarily due to the composite outcome (MACE) that was chosen for the evaluation. The other limitation of this work is that there is no true gold standard for whether or not a given data set should have been considered 'of concern.' This was approximated by truncating the number of months in the analysis to 24 and setting the gold standard to the result at 36 months, but this will generally inflate the ROC of each of the methods.

Further work in this area will be done to evaluate the absolute and relative performance between these methodologies for data with event rate trends. As noted above, the relative performance between the methods was relatively similar with the noted exceptions. However, greater variation between the methods might be observed for event rates with different event rate trends (shown in Figure 2). These trends will be modeled and the methods will be evaluated, such as: ascending event rates, descending event rates, "V" or descending then ascending event

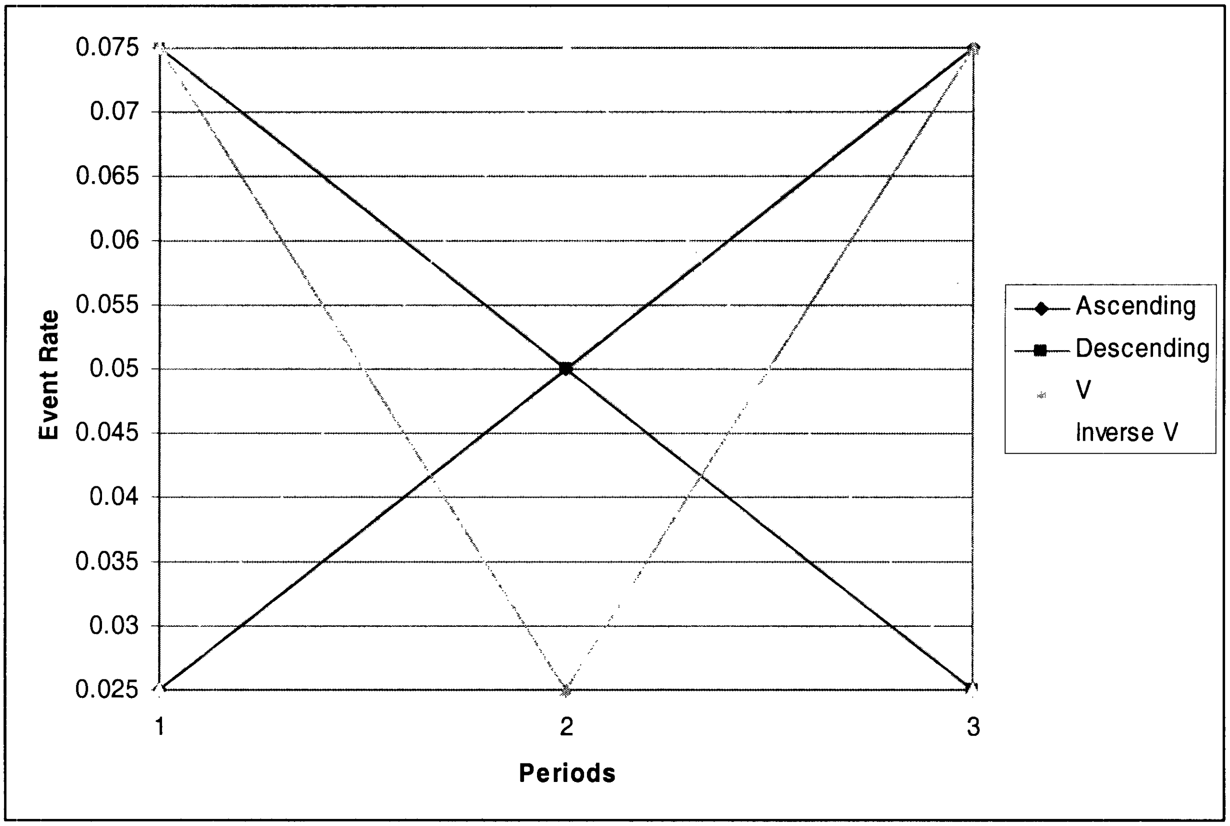


Figure 2: Illustration of possible event rate trends in a data set, such as ascending, descending, V, and inverse V event rate trends.

rates, and "Inverse V" or ascending then descending event rates. We suspect that each of the methods will have strengths and weaknesses relating to the data event rate trends.

Overall, this sensitivity analysis shows that relative agreement between SPC and BUS methods for low event rates (0.5% and 5%), and disagreement between the methods at high event rates (50%). In addition, using the number of alerts by month appears to be a valid way to determine when an outcome event rate is becoming a concern based on established baseline data. Further work to establish a 'number of alerts' threshold for the purposes of sensitivity and specificity will be required, and is underway.

Thesis Summary

Risk prediction models have been widely used for prognostication and retrospective comparisons. However, there are opportunities to adapt these methodologies to perform prospective monitoring of clinical outcomes. Incorporation of these methods into automated tools could help support the evolving field of medical post-marketing surveillance.

The performance of the most well-known mortality risk models for the outcome of death in percutaneous coronary intervention were evaluated on Brigham & Women's Hospital clinical data. All of the models retained good discrimination. While this was promising, all of the models except the recent locally developed one showed poor calibration, suggesting that changes over time as well as regional patient demographics and medical care delivered may play an important role in the generalizability of a model. The preservation of discrimination supports the use of these models for generic risk stratification, but poor calibration indicates that use of the models for individual prognostication should be done with caution, and indicates the need for further study.

An exploration of support vector machines as an alternative to logistic regression risk modeling for individual prognostication was performed. While both modeling techniques had excellent and relatively similar discrimination, the support vector machine models were consistently superior in terms of calibration. This suggests that this method could provide superior risk stratification and prognostication for use in the real-time safety monitoring application.

A real-time safety monitoring system, DELTA, was designed with a SQL database for back-end data storage and a web-based graphical user interface for user operation. The software satisfied all pre-specified design requirements, which were 1) the ability to accept a generic flat data table, 2) perform both retrospective and prospective analyses, 3) incorporate a complimentary set of statistical methods for event rate monitoring, 4) provide user-configurable alerting thresholds, and automated alert notification, and 5) support an unlimited number of simultaneous datasets with nested analyses.

A pilot investigation evaluating the in-hospital mortality following implantation of drug-eluting coronary artery stents was then performed with the system. The system performed as expected, and event rates compared to University of Michigan bare-metal stent data were acceptably low using adapted statistical techniques including Statistical Process Control, Bayesian Updating Statistics, and Logistic Regression.

Subsequently, sensitivity analyses between the SPC and BUS alerting methods were performed to evaluate individual and relative performance of these measures on scaled event rates for BWH in-hospital major adverse coronary events. A variety of baseline event rates were evaluated. The analysis revealed relative agreement between SPC and BUS methods for low event rates, but noted an increasing disagreement between the methods as observed event rates escalated. In addition, the SPC method generated more alerts for a given analysis configuration than BUS. Finally, using the number of alerts by month appeared to be a valid way to determine a 'significant event rate.' Further work to establish specificity and sensitivity thresholds will be required as gold standard data becomes available.

These research efforts have shown highlighted some of the limitations relating to calibration in current risk modeling methods. Machine learning modeling methods may provide improvements in this area. In addition, the design feasibility of a real-time automated monitoring system was confirmed, some potentially useful statistical monitoring methodologies were established and evaluated, and the utility and applicability of both logistic regression and support vector machine modeling were explored for use in the application. Further work remains to validate and scale up the application for use in multiple clinical environments.

Acknowledgements

The author would like to thank Dr. Frederic Resnic and Dr. Lucila Ohno-Machado for their excellent mentorship and guidance, without which this work would not have been possible. In addition, the author would like to thank Richard Cope and Barry Coflan for their excellent programming work on the DELTA application, Dr. Robert Greenes for guidance and editing assistance, and Anne Fladger and her staff for literature procurement assistance.

This research was supported in part by Grants T15-LM-07092 and R01-LM-08142 from the National Library of Medicine of the National Institutes of Health.

REFERENCES

1. Munsey RR. Trends and events in FDA regulation of medical devices over the last fifty years. *Food Drug Law J.* 1995;50 Spec:163-177.
2. Medical Device Amendments of 1976 to the Federal Food, Drug, and Cosmetic Act. *Pub L No. 94-295.* 1976:90 Stat 539.
3. Pritchard WF, Jr., Carey RF. U.S. Food and Drug Administration and regulation of medical devices in radiology. *Radiology.* Oct 1997;205(1):27-36.
4. Safe Medical Devices Act of 1990. *Pub L No. 101-629.* 1990:104 Stat 4511.
5. Merrill RA. Modernizing the FDA: an incremental revolution. *Health Aff (Millwood).* Mar-Apr 1999;18(2):96-111.
6. Kessler L, Richter K. Technology assessment of medical devices at the Center for Devices and Radiological Health. *American Journal of Managed Care.* Sep 25 1998;4 Spec No:SP129-135.
7. O'Neill WW, Chandler JG, Gordon RE, et al. Radiographic detection of strut separations in Bjork-Shiley convexo-concave mitral valves.[see comment]. *New England Journal of Medicine.* Aug 17 1995;333(7):414-419.
8. Brown SL, Morrison AE, Parmentier CM, Woo EK, Vishnuvajjala RL. Infusion pump adverse events: experience from medical device reports. *Journal of Intravenous Nursing.* Jan-Feb 1997;20(1):41-49.
9. Fuller J, Parmentier C. Dental device-associated problems: an analysis of FDA postmarket surveillance data. *Journal of the American Dental Association.* Nov 2001;132(11):1540-1548.
10. White GG, Weick-Brady MD, Goldman SA, et al. Improving patient care by reporting problems with medical devices. *CRNA.* Nov 1998;9(4):139-156.
11. Dwyer D. Medical device adverse events and the temporary invasive cardiac pacemaker. *International Journal of Trauma Nursing.* Apr-Jun 2001;7(2):70-73.
12. Dillard SF, Hefflin B, Kaczmarek RG, Petsonk EL, Gross TP. Health effects associated with medical glove use. *AORN Journal.* Jul 2002;76(1):88-96.
13. Brown SL, Duggirala HJ, Pennello G. An association of silicone-gel breast implant rupture and fibromyalgia. *Current Rheumatology Reports.* Aug 2002;4(4):293-298.
14. *Managing Risks from Medical Product Use: Creating a Risk Management Framework. Report to the FDA Commissioner from the Task Force on Risk Management.* Rockville, MD: U.S. Department of Health and Human Services, Food and Drug Administration; 1999.
15. Monsein LH. Primer on medical device regulation. Part II. Regulation of medical devices by the U.S. Food and Drug Administration. *Radiology.* Oct 1997;205(1):10-18.
16. Maisel WH. Medical device regulation: an introduction for the practicing physician. *Annals of Internal Medicine.* Feb 17 2004;140(4):296-302.
17. Monsein LH. Primer on medical device regulation. Part I. History and background. *Radiology.* Oct 1997;205(1):1-9.
18. Medical device and user facility and manufacturer reporting, certification and registration; delegations of authority; medical device reporting procedures; final rules. *Fed Regist.* 1995;60:63577-63606.

19. *Improving patient care by reporting problems with medical devices. Med Watch.* Rockville, MD: Department of Health and Human Services, U.S. Food and Drug Administration, HF-2; 1997.

20. Feigal DW, Gardner SN, McClellan M. Ensuring safe and effective medical devices. *New England Journal of Medicine.* Jan 16 2003;348(3):191-192.

21. Postmarket surveillance. Final Rule. *Fed Regist.* 2002;67:5943-5942.

22. Center for Devices and Radiologic Health Annual Report Fiscal Year 2000. www.fda.gov/cdrh/annual/fy2000/annualreport-2000-5.html. Accessed January 18, 2005.

23. Cannon CP, Battler A, Brindis RG, et al. American College of Cardiology key data elements and definitions for measuring the clinical management and outcomes of patients with acute coronary syndromes. *J Am Coll Cardiol.* Dec 2001;38(7):2114-2130.

24. Doble M. Six Sigma and chemical process safety. *Int. J. Six Sigma & Compet. Adv.* 2005;1(2):229-244.

25. Randolph AG, Guyatt GH, Carlet J. Understanding articles comparing outcomes among intensive care units to rate quality of care. Evidence Based Medicine in Critical Care Group. *Critical Care Medicine.* Apr 1998;26(4):773-781.

26. Topol EJ, Block PC, Holmes DR, Klinke WP, Brinker JA. Readiness for the scorecard era in cardiovascular medicine. *Am J Cardiol.* Jun 1 1995;75(16):1170-1173.

27. Hunt JP, Meyer AA. Predicting survival in the intensive care unit. *Current Problems in Surgery.* Jul 1997;34(7):527-599.

28. Knaus WA, Wagner DP, Draper EA. The value of measuring severity of disease in clinical research on acutely ill patients. *Journal of Chronic Diseases.* 1984;37(6):455-463.

29. Mendez-Tellez PA, Dorman T. Predicting patient outcomes, futility, and resource utilization in the intensive care unit: the role of severity scoring systems and general outcome prediction models. *Mayo Clin Proc.* Feb 2005;80(2):161-163.

30. O'Connor GT, Malenka DJ, Quinton H, et al. Multivariate prediction of in-hospital mortality after percutaneous coronary interventions in 1994-1996. *J Am Coll Cardiol.* Sep 1999;34(3):681-691.

31. Hannan EL, Arani DT, Johnson LW, Kemp HG, Jr., Lukacik G. Percutaneous transluminal coronary angioplasty in New York State. Risk factors and outcomes. *JAMA.* Dec 2 1992;268(21):3092-3097.

32. Hannan EL, Racz M, Ryan TJ, et al. Coronary angioplasty volume-outcome relationships for hospitals and cardiologists. *JAMA.* Mar 19 1997;277(11):892-898.

33. Moscucci M, Kline-Rogers E, Share D, et al. Simple bedside additive tool for prediction of in-hospital mortality after percutaneous coronary interventions. *Circulation.* Jul 17 2001;104(3):263-268.

34. Shaw RE, Anderson HV, Brindis RG, et al. 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.* Apr 3 2002;39(7):1104-1112.

35. Ellis SG, Weintraub W, Holmes D, Shaw R, Block PC, King SB, 3rd. Relation of operator volume and experience to procedural outcome of percutaneous coronary revascularization at hospitals with high interventional volumes. *Circulation.* Jun 3 1997;95(11):2479-2484.

36. Holmes DR, Selzer F, Johnston JM, et al. Modeling and risk prediction in the current era of interventional cardiology: a report from the National Heart, Lung, and Blood Institute Dynamic Registry. *Circulation.* Apr 15 2003;107(14):1871-1876.

37. Holmes DR, Jr., Berger PB, Garratt KN, et al. Application of the New York State PTCA mortality model in patients undergoing stent implantation. *Circulation.* Aug 1 2000;102(5):517-522.

38. Moscucci M, O'Connor GT, Ellis SG, et al. Validation of risk adjustment models for in-hospital percutaneous transluminal coronary angioplasty mortality on an independent data set. *Journal of the American College of Cardiology.* Sep 1999;34(3):692-697.

39. Rihal CS, Grill DE, Bell MR, Berger PB, Garratt KN, Holmes DR, Jr. Prediction of death after percutaneous coronary interventional procedures. *American Heart Journal.* Jun 2000;139(6):1032-1038.

40. Singh M, Rihal CS, Selzer F, Kip KE, Detre K, Holmes DR. Validation of Mayo Clinic risk adjustment model for in-hospital complications after percutaneous coronary interventions, using the National Heart, Lung, and Blood Institute dynamic registry.[see comment]. *Journal of the American College of Cardiology.* Nov 19 2003;42(10):1722-1728.

41. Kizer JR, Berlin JA, Laskey WK, et al. Limitations of current risk-adjustment models in the era of coronary stenting. *American Heart Journal.* Apr 2003;145(4):683-692.

42. Hannan EL, Wu C. Assessing quality and outcomes for percutaneous coronary intervention: choosing statistical models, outcomes, time periods, and patient populations. *American Heart Journal.* Apr 2003;145(4):571-574.

43. deDombal FT. Computer-aided diagnosis and decision-making in the acute abdomen. *Journal of the Royal College of Physicians of London.* Apr 1975;9(3):211-218.

44. Oye RK, Bellamy PE. Patterns of resource consumption in medical intensive care. *Chest.* Mar 1991;99(3):685-689.

45. Williams DO, Holubkov R, Yeh W, et al. Percutaneous coronary intervention in the current era compared with 1985-1986: the National Heart, Lung, and Blood Institute Registries.[see comment]. *Circulation.* Dec 12 2000;102(24):2945-2951.

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

47. Poses RM, Smith WR, McClish DK, et al. Physicians' survival predictions for patients with acute congestive heart failure. *Archives of Internal Medicine.* May 12 1997;157(9):1001-1007.

48. Perkins HS, Jonsen AR, Epstein WV. Providers as predictors: using outcome predictions in intensive care. *Crit Care Med.* Feb 1986;14(2):105-110.

49. Shaw RE, Anderson HV, Brindis RG, et al. Updated risk adjustment mortality model using the complete 1.1 dataset from the American College of Cardiology National Cardiovascular Data Registry (ACC-NCDR). *Journal of Invasive Cardiology.* Oct 2003;15(10):578-580.

50. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine.* Feb 29 2000;19(4):453-473.

51. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* Apr 1982;143(1):29-36.

52. Swets JA. Measuring the accuracy of diagnostic systems. *Science.* Jun 3 1988;240(4857):1285-1293.

53. Lemeshow S, Hosmer DW, Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology.* Jan 1982;115(1):92-106.

54. Margolis DJ, Bilker W, Boston R, Localio R, Berlin JA. Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *Journal of Clinical Epidemiology.* May 2002;55(5):518-524.

55. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine.* May 15 2000;19(9):1141-1164.

56. 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.* Jul 1 2001;88(1):5-9.

57. Hennekens CH, Buring JE. *Epidemiology in Medicine.* Boston: Little & Brown; 1988.

58. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* 1986;1:54-77.

59. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* Sep 1983;148(3):839-843.

60. McGrath PD, Malenka DJ, Wennberg DE, et al. Changing outcomes in percutaneous coronary interventions: a study of 34,752 procedures in northern New England, 1990 to 1997. Northern New England Cardiovascular Disease Study Group. *Journal of the American College of Cardiology.* Sep 1999;34(3):674-680.

61. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine.* Dec 15 1997;16(23):2645-2664.

62. Peterson ED, DeLong ER, Muhlbaier LH, et al. Challenges in comparing risk-adjusted bypass surgery mortality results: results from the Cooperative Cardiovascular Project. *Journal of the American College of Cardiology.* Dec 2000;36(7):2174-2184.

63. Moss AJ, Hamburger S, Moore RM, Jeng LL, Howie LJ. *Use of selected medical device implants in the United States.* Vol 191. Rockville, MD: National Center For Health Statistics; 1988.

64. New York Public Health Law §2805-1 Incident Reporting (Added L. 1986, c.266).

65. Department of Public Health. *105 Code of Massachusetts Regulations.* 2001:130.1201-1130.1130.

66. Cook DF. Statistical Process Control for continuous forest products manufacturing operations. *Forest Products Journal.* 1992;42:47-53.

67. Grigg N, Walls L. The use of statistical process control in food packaging: Preliminary findings and future research agenda. *British Food Journal.* 1999;101:763-784.

68. Developed Wheel and Axle Assembly Monitoring System for Improved Passenger Train Safety. *US DOT Federal Railroad Administration.* March 2000;RR00-02.

69. Standards for Privacy of Individually Identifiable Health Information: Final Rule. *Fed. Regist.* 2002;67:53182-53273.

70. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine.* Apr 30 1998;17(8):857-872.

71. Hosmer DW, Lemeshow S. Applied Logistic Regression, 2nd Ed. 2000.

72. Siu N, Apostolakis G. Modeling the detection rates of fires in nuclear plants. *Risk Anal.* 1986;6:43-59.

73. Bayes T. Essay towards solving a problem in the doctrine of chanes. *Philos. Trans. R. Soc. Lond.* 1763;53:370-418.

74. Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. *J Biomed Inform.* Oct 2005;38(5):367-375.

75. Matheny ME, Ohno-Machado L, Resnic FS. Monitoring Device Safety in Interventional Cardiology. *J Am Med Inform Assoc.* 2006;13(2):180-187.

76. Oakland JS. *Statistical Process Control.* 5 ed. Jordan Hill, Oxford, UK: Butterworth-Heinemann; 2003.

77. Resnic FS, Zou KH, Do DV, Apostolakis G, Ohno-Machado L. Exploration of a bayesian updating methodology to monitor the safety of interventional cardiovascular procedures. *Medical Decision Making.* Jul-Aug 2004;24(4):399-407.

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

79. Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Arch Pathol Lab Med.* Jan 1986;110(1):13-20.

80. Hosmer D, Lemeshow S. *Applied Logistic Regression.* New York, NY: Wiley & Sons; 1989.

81. Scholkopf B, Sung K, Burges C, et al. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans Sig Proc.* 1997;45:2758-2765.

82. Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola AJ, Bartlett P, Schoelkopf B, Schuurmans D, eds. *Advances in Large Margin Classiers.* Cambridge, MA: MIT Press; 1999.

83. Lin H-T, Lin C-J, Weng RC. A note on Platt's probabilistic outputs for support vector machines. http://www.csie.ntu.edu.tw/~cjin/papers/plattprob.ps. Last Accessed: 03/08/2006.

84. Platt J. Fast training of support vector machines using sequential minimal optimization. In: Scholkopf B, Burges C, Smola A, eds. *Advances in Kernel Methods - Support Vector Learning*; 1998.

85. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics.* Mar 1 2005;21(5):631-643.