Validation of an Automated Safety Surveillance System with Prospective, Randomized Trial Data

Michael E. Matheny, David A. Morrow, Lucila Ohno-Machado, Christopher P. Cannon, Marc S. Sabatine and Frederic S. Resnic

Med Decis Making 2009; 29; 247 originally published online Nov 17, 2008;
DOI: 10.1177/0272989X08327110

Validation of an Automated Safety Surveillance System with Prospective, Randomized Trial Data

Michael E. Matheny, MD, MS, MPH, David A. Morrow, MD, MPH, Lucila Ohno-Machado, MD, PhD, Christopher P. Cannon, MD, Marc S. Sabatine, MD, MPH, Frederic S. Resnic, MD, MS

Objective. *We sought to validate 3 methods for automated safety monitoring by evaluating clinical trials with elevated adverse events.* ***Methods.*** *An automated outcomes surveillance system was used to retrospectively analyze data from 2 randomized, TIMI multicenter trials. Trial A was stopped early due to elevated 30-day mortality rates in the intervention arm. Trial B was not stopped early, but there was transient concern regarding 30-day intracranial hemorrhage rates. We compared statistical process control (SPC), logistic regression risk adjusted SPC (LR-SPC), and Bayesian updating statistic (BUS) methods with a standard prospective 2-arm event rate analysis. Each method compares observed event rates to alerting boundaries established with previously collected data. In this evaluation, the control arms approximated prior data, and the intervention arms approximated the observed data.* ***Results.*** *Trial A experienced elevated 30-day mortality rates beginning 7 months after the start of the trial and continuing until termination at month 14. Trial B did not experience elevated major bleeding rates. Combining the alerting performance of each method across both trials resulted in sensitivities and specificities of 100% and 85% for SPC, 0% and 100% for BUS, and 100% and 93% for both LR-SPC models, respectively.* ***Conclusion.*** *Both SPC and LR-SPC methods correctly identified the majority of months during which the cumulative event rates were elevated in trial A but were susceptible to false positive alerts in trial B. The BUS method did not result in any alerts in either trial and requires revision.* ***Key words:*** *risk adjustment; risk stratification; decision support techniques; cardiology.* ***(Med Decis Making 2009;29:247–256)***

Recent product recalls of both medications and medical devices have highlighted the need for robust improvements in postmarketing surveillance methods.[1] While postmarketing clinical outcomes data have become increasingly available in the form of clinical registries and electronic health records, there is no consensus on which continuous monitoring methodologies would be most appropriate for these types of observational cohort data.

There are a number of statistical process control (SPC) techniques used for adverse event surveillance in industrial processes, such as Shewart control charts,[2] exponentially weighted moving-average (EWMA) charts, cumulative sum (CUSUM) charts, and sequential probability ratio tests (SPRTs).[3] Until recently, these methods had limited application in the medical environment because of clinical data heterogeneity, which can be found among demographic and clinical characteristics of patients, provider practice variation, nonstandard data collection, and missing data elements. While attempts have been made to apply basic SPC methods to medical outcomes surveillance,[4–7] these factors require sophisticated risk adjustment and protocols for establishing adverse event rate alerting boundaries that are unnecessary in industrial processes.

Received 14 December 2007 from the Decision Systems Group, Department of Radiology (MEM, LOM, FSR), the Division of General Medicine (MEM), the Division of Cardiology (DAM, CPC, MSS, FSR), and the TIMI Study Group, Brigham & Women's Hospital, Boston, MA (DAM, CPC, MSS); and the Division of Health Sciences & Technology, Massachusetts Institute of Technology, Cambridge, MA (LOM). This study was funded in part by grants T15-LM-007092, R01-LM-008142, and R01-LM-009520 from the National Library of Medicine of the National Institutes of Health. Revision accepted for publication 3 August 2008.

Address correspondence and reprint requests to Michael E. Matheny, MD, MS, MPH, 1310 24th Ave. S., GRECC, Room 4-B110, Nashville, TN 37212; telephone: (615)327-4751 x6821; fax: (615)327-5381; e-mail: michael.matheny@vanderbilt.edu.

DOI: 10.1177/0272989X08327110

Statistical advances in risk adjustment methods and subsequent incorporation of those techniques into some SPC frameworks have facilitated use of these tools in medical outcomes surveillance. There have been some evaluations of these methods among specific clinical domains, such as pediatric and adult cardiac surgery,[3,8,9] general surgery,[10] and interventional cardiology.[11] Most of these studies were performed on retrospective clinical registry data, but the results are encouraging for detecting unexpected or elevated adverse event rates among broader applications. However, the full utility of these methods can be found for prospective data monitoring.

We developed a computer application, Data Extraction and Longitudinal Time Analysis (DELTA), that provides both retrospective and prospective outcomes monitoring among new and established medical devices and medications.[12] A number of nonrisk-adjusted and risk-adjusted SPC-based methods were developed and adapted for use in this application, and pilot studies using the methods and system have been successfully conducted within a single-institution interventional cardiology clinical registry.[11–14]

However, our methods have not been fully validated. Randomized controlled trial (RCT) data provide a gold standard for comparison. Any attempt to validate these methods using observational cohort data faces limitations from potential unmeasured confounding. RCT data balance unmeasured confounding between control and intervention arms within the trial design, and using the control arm to provide the baseline or event rate expectations, address this limitation. In addition, RCT data provide meticulously adjudicated outcomes with independent review by a data safety monitoring board (DSMB) both at set time intervals and after the conclusion of a trial. In this idealized setting, results from an SPC-based method with and without risk adjustment should be approximately the same, which allows flaws in the risk adjustment unrelated to confounding to be discovered. This also compares baseline accuracy of both types of methods to standard trial statistical analysis and the DSMB findings.

In this study, we sought to validate 3 SPC-based methods imbedded in an automated monitoring application against a standard of statistical methods employed by DSMBs with RCT data. To provide a true positive signal, we selected a trial that was stopped early because of a high rate of adverse events. To provide a true negative signal with a reasonable chance of a false positive alarm, we selected a trial in which adverse event rates were of concern early on but never met the established DSMB stopping rules.

METHODS

Clinical Trial Data

Two Thrombolysis In Myocardial Infarction (TIMI) RCTs with DSMB monitoring and safety endpoint stopping rules were selected for use in this study.[15,16] Trial A was a multicenter, randomized trial that evaluated the efficacy of an oral platelet glycoprotein antagonist (GP IIb/IIIa inhibitor) versus placebo with regards to a composite primary endpoint of death, myocardial infarction, stroke, or recurrent ischemia at rest leading to rehospitalization or emergent revascularization at 30 days and 10 months postrandomization. A total of 10,288 patients were enrolled in the study within 3 arms: 3421 in the placebo arm, 3330 in the active treatment arm with sustained dosing, and 3537 in the active treatment arm with reduced dosing after the first 30 days. The primary safety endpoints for the trial were all-cause mortality or severe or life-threatening bleeding, defined as intracranial hemorrhage or bleeding associated with severe hemodynamic compromise, a drop in hematocrit by 15% or more, or requiring a blood transfusion. The trial was terminated by the DSMB before the goal of 12,000 patients was reached due to an increase in the 30-day mortality in the reduced dose treatment arm.[15]

Trial B was a multicenter, randomized trial that evaluated the efficacy of an oral antiplatelet agent versus placebo in combination with fibrinolytic therapy in ST elevation myocardial infarction with regards to a composite primary endpoint of an occluded infarct-related artery by angiography, or death or recurrent myocardial infarction in the absence of angiography. A total of 3491 patients were enrolled in the study: 1739 in the control arm and 1751 in the treatment arm. The primary safety endpoint was TIMI major bleeding, of which intracranial hemorrhage was a component.[17] The trial operations committee became concerned with the rates of intracranial bleeding and major bleeding early in the trial based upon aggregated blinded data, but neither safety outcome reached DSMB stopping criteria during the trial. This study was approved by the Brigham & Women's Institutional Review Board.

Statistical Methodologies

Three statistical methods were used to assess for deviation from acceptable safety benchmarks during this evaluation. These include both a nonrisk-adjusted and risk-adjusted SPC method based upon Shewart control charts. The third method is a nonrisk-adjusted

Bayesian adaptation of Shewart control charts. These methods are reviewed in brief below and are more fully described elsewhere.[12]

All of the methods use 4 values in each time period in order to calculate whether to generate an alert. These values are observed number of events, observed number of cases, expected number of events, and expected number of cases. A limitation of standard Shewart control charts is that the observed number of cases (sample size) is ignored because the observed value is only represented as a proportion or point estimate. This can result in alerting insensitivity, particularly in cumulative analyses where the values for each observed period are the sum of all the observed periods and the current period. We addressed this by adapting each of the methods to generate the alert threshold using the Wilson method for comparing independent proportions.

Statistical process control (SPC). This method is an adaptation (as described above) of standard nonrisk-adjusted Shewart control charts to provide cumulative event rate monitoring in which adverse events and total cases are aggregated and analyzed in predefined time periods. This method is most appropriate under unchanging conditions where deviations from an established norm need to be detected and is very reliable for these purposes.[2]

Logistic regression adjusted statistical process control (LR-SPC). This[18] is an experimental methodology that incorporates our Shewart control chart adaptation with logistic regression (LR), a modeling technique that estimates the probability of an outcome on a case-level basis. An LR model is developed from the available baseline data and validated by resampling methods or external data sets. The model is then applied to the observed data, and the model outcome probabilities are considered the expected (baseline) number of events. This incorporates risk adjustment into the method by allowing the expected event rate to change over time with the composition of the observed cases.

However, the limitations of this method are those inherent in LR modeling in general. The levels of discrimination and calibration of the model on the baseline data are not guaranteed to remain the same even on closely related subsequent populations. While discrimination is generally retained across different patient populations, calibration can vary, which may directly impact monitoring results.[19]

Bayesian updating statistics (BUS). This is an experimental methodology pioneered in nuclear power safety monitoring.[20] We developed this method by incorporating Bayesian statistics[21] into a traditional SPC framework by utilizing prior observed data to evolve the estimates of risk.[22] The incorporation of previously observed information into the expected data allows the method to be very sensitive for detecting reversal of trends and sudden, large changes in event rates. However, a slow drift (either elevation or depression) of the observed event rate can be missed.[13] This method partially addresses the limitation in changing conditions in SPC and is best suited to detecting changes after an incremental change (such as a medication or device) is introduced. However, this method should be considered nonrisk adjusted because individual patient conditions and exposures are not considered in establishing the alerting threshold.

Automated Monitoring Tool

We have previously described an automated real-time safety monitoring tool, DELTA, that is able to perform an arbitrary number of concurrent prospective analyses using statistical methodologies (SPC, BUS, LR-SPC) and alerting thresholds.[12] The system uses an SQL 2000 server (Microsoft Corp., Redmond, WA) to provide internal data storage and configuration information, as well as provide the capability to integrate with external databases. The user interface is displayed in a Web browser from a Microsoft IIS 5.0 Web Server.

The system is currently in operation within the Partners Healthcare intranet, a secure multihospital network. Security of patient data is further addressed by record de-identification steps[23] and user login access restrictions. DELTA is part of ongoing quality assessment and control measures within the institution.

Data Analysis

Both sets of trial data were imported into DELTA, and then SPC, LR-SPC, and BUS analyses were configured to evaluate the outcomes of interest in monthly intervals. These outcomes were the primary safety endpoints of the trials, which were 30-day mortality for trial A and major bleeding for trial B. The gold standard of whether the appropriate safety endpoint event rate in each trial was elevated in a particular month was determined using standard DSMB analysis methods. This was calculated using the Fisher exact test applied to the cumulative control and intervention data on a monthly basis for each

outcome of interest in both trials. The proportional difference method with 95% confidence intervals was used to establish alert thresholds for the same sets of cumulative data in order to compare performance with the Fisher exact method.

Each of the statistical methods used by DELTA requires a baseline event rate expectation to establish alerting thresholds, and these data are generally obtained from observational cohort data prior to the initiation of a new medication or device. In order to simulate this environment with RCT data, the control arms were used for this baseline measurement, and the intervention arms were used as the monitored prospective observational cohort. This resulted in all of the control arm data being available at the ''beginning'' of the intervention arm monitoring. The reduced dosage treatment arm in trial A was utilized for the intervention arm because that was the arm for which the DSMB stopped the trial.

The SPC alerting threshold was static, the LR-SPC alerting threshold was adjusted in each analyzed time period by the model predicted event rate of the observed (intervention) data, and the BUS alerting threshold was adjusted in each analyzed time period by the observed event rate of the observed (intervention) data. Monthly time intervals and 95% confidence intervals or posterior credible intervals were used for each analysis in this study.[24] Overall sensitivity and specificity of the methods can be ''tuned'' by adjusting the alerting threshold, but the emphasis in this evaluation was to determine relative performance between the methods for a standard threshold set point.

LR-SPC required the development of a logistic regression model in order to perform case-level risk adjustment. A literature search was conducted to identify risk factors for each of the outcomes of interest in the trials, and all such factors that were associated with the respective outcome of interest were included in the LR model development process. Model development was done in SAS (version 9.1, Cary, NC). The models were evaluated for discrimination with the area under the ROC curve (AUC) and calibration with the Hosmer-Lemeshow goodness-of-fit (HL-GOF) deciles test using 10-fold cross-validation.[25,26]

Accuracy of each method for detecting elevated event rates was calculated by comparing whether each method alerted or not compared to the standard trial analysis. There were a total of 14 months in trial A and 21 months in trial B, resulting in 35 values (alert/nonalert) for each method. The results in each month for both trials were aggregated together to determine overall sensitivity (defined as the true positives divided by the sum of true positives and false negatives) and specificity (defined as the true negatives divided by the sum of true negatives and false positives).

The cross-validation results for the LR models developed from the trial A control arm data were an AUC of 0.67 (0.59–0.75) and an HL-GOF of 8.82 ($P = 0.358$) for a variable selection threshold of 0.01 and an AUC of 0.70 (0.62–0.78) and an HL-GOF of 8.38 ($P = 0.397$) for a threshold of 0.20. The cross-validation results for the LR models developed from the trial B control arm data were an AUC of 0.79 (0.73–0.86) and an HL-GOF of 15.7 ($P = 0.047$) for a threshold of 0.01 and an AUC of 0.80 (0.74–0.87) and an HL-GOF of 10.5 ($P = 0.230$) for a threshold of 0.20.

RESULTS

Significant differences for the outcome of all-cause death at 30 days in trial A were noted between the control arm and the reduced treatment arm from month 7 until the trial's early termination at month 14. A summary of the event rates of each arm and *P* values by month is listed in Table 1. The SPC monitoring method also reported significant event rate elevations in months 7 through 14 (Figure 1A). The BUS method, however, did not report any intervention arm event rate elevations during monitoring. The LR-SPC method using a model building threshold of 0.01 reported elevations in months 7 through 14 (Figure 1B) and reported elevations in months 6 through 14 for a model building threshold of 0.20 (Figure 1C). A summary of the trial A proportional difference results for each of the monitoring methods by month is shown in Appendix 1.

No significant differences for the outcome of major bleeding at 30 days were noted between the control and intervention arms in trial B. Month 8 was the period in which the event rate of the intervention arm was the most elevated in relation to the control arm with a *P* value of 0.262. A summary of the event rates for each arm with *P* values by month is listed in Table 2. The SPC monitoring method did generate alerts for months 7 through 9 and 11 (Figure 2A). The BUS method did not generate any alerts. The LR-SPC method using a model building threshold of 0.01 reported elevations in months 7 and 14 (Figure 2B) and reported an elevation in month 7 for a model building threshold of 0.20 (Figure 2C). A summary of the trial B proportional difference results for each of

Table 1 Trial A Fisher Exact Test Method

	Control (Expected)			Intervention Arm B (Observed)			
Period	Events	Patients	Event Rate (%)	Events	Patients	Event Rate (%)	*P*
1	0	0	0.0	0	5	0.0	—
2	0	17	0.0	0	16	0.0	—
3	0	48	0.0	1	52	1.9	1.000
4	0	135	0.0	4	131	3.1	0.122
5	1	268	0.4	6	279	2.2	0.124
6	4	463	0.9	12	486	2.5	0.078
7	5	764	0.7	21	794	2.6	**0.003**
8	8	1089	0.7	30	1118	2.7	**<0.001**
9	11	1412	0.8	36	1459	2.5	**<0.001**
10	16	1805	0.9	50	1882	2.7	**<0.001**
11	21	2173	1.0	59	2277	2.6	**<0.001**
12	33	2701	1.2	66	2824	2.3	**0.003**
13	46	3360	1.4	81	3470	2.3	**0.005**
14	46	3421	1.3	81	3537	2.3	**0.005**

Note: Death at 30 days for intervention arm B. Cumulative enrollment and events. Bold = alert.

the monitoring methods by month is shown in Appendix 2.

Aggregating the results of both trials for each method as compared to the trial analysis standard resulted in sensitivities and specificities of 100% and 85% for SPC, 0% and 100% for BUS, and 100% and 93% for both LR-SPC models, respectively. A summary of the 2 × 2 table elements is listed in Table 3.

DISCUSSION

This study evaluated nonrisk-adjusted and risk-adjusted SPC methods for detecting elevated adverse event rates among RCT data. Both SPC and LR-SPC performed well and were comparable to each other. However, BUS was significantly overspecific and did not alert in any month in either trial.

The proportional difference test was validated by comparison with the Fisher exact test, and the results were concordant in both trials analyzed. The SPC method alerted properly in each of the months identified by the Fisher method for the trial A data and did alert in 4 months in the trial B data in which the Fisher method did not find a significant difference. The LR-SPC method alerted appropriately with the exception of one false positive in trial A (for the 0.20 model threshold) and one false positive alert in trial B (both thresholds). The BUS method did not alert any during either trial, resulting in false negatives in trial A.

Substantially different performance was found between SPC and BUS. The SPC method was consistently more sensitive, and the BUS method was more specific. The performance of both methods is sensitive to the data used in establishing the expected event rates and alerting thresholds. Theoretically, as the *n* (number of subjects) of the baseline data increases, SPC becomes more sensitive and BUS becomes less sensitive (and more specific) to event rate deviations in the monitored data. Conversely, BUS should more rapidly detect an event rate difference than SPC for sparse or low volumes of baseline data, and this has been shown in other monitoring applications.[27,28] The clinical trials evaluated here had large numbers of control patients in order to appropriately evaluate the primary outcomes, and this could have favored the SPC method in this analysis. A sensitivity analysis between the performance of SPC and BUS for large ranges of *n* is ongoing in order to determine relative performance between the methods, and we are currently evaluating an alternate BUS alerting threshold using the percentage overlap of the area under the probability density function.

LR-SPC, unlike SPC and BUS, is not directly sensitive to the *n* of the baseline data because the alerting threshold is generated from the predicted event rate of the observed data, which results in the *n* used to generate the alerting threshold being equal to *n* of the observed data. LR-SPC is sensitive to the performance of the LR model used, and such models are more robust when generated from larger data sets.

In this evaluation, LR-SPC performed in a comparable manner to the SPC method. However, this result should be interpreted as LR-SPC performing in a noninferior way to SPC in the absence of confounding in the evaluated data. This is a useful

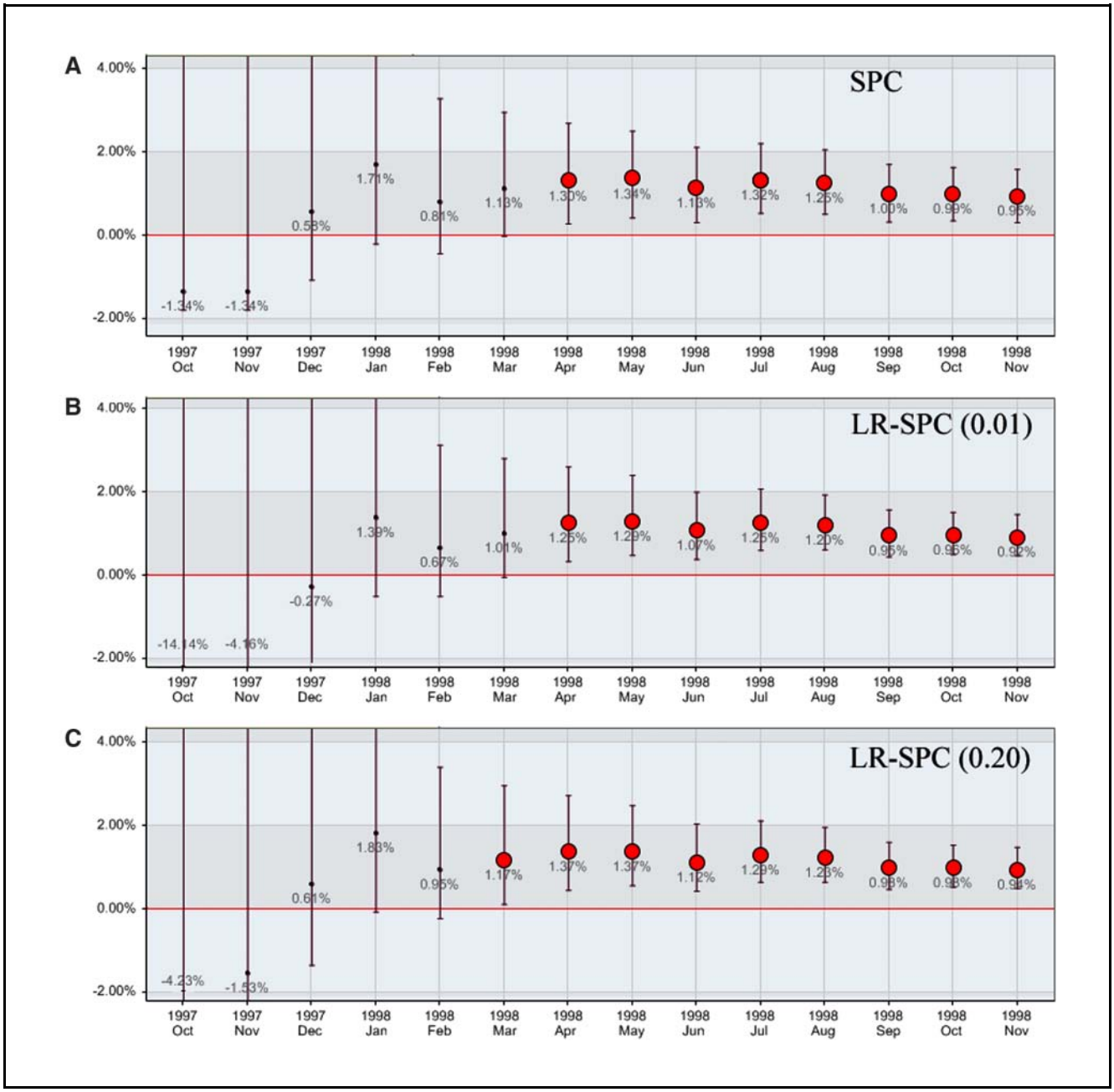

Figure 1 Trial A mortality for treatment arm B by method. (A) Statistical process control (SPC), (B) logistic regression adjusted statistical process control (0.01) (LR-SPC), and (C) logistic regression adjusted statistical process control (0.20) (LR-SPC).

finding because it supports the use of the methodology, but it does not provide an evaluation of the method's risk adjustment efficacy. Further work needs to be performed to establish the relative performance between SPC and LR-SPC using observational cohort data.

There are a number of limitations to this study. The control group data in both trials were accumulated concurrently and in a randomized fashion with the intervention data. However, in order to evaluate the system, the control data were assumed to be collected prior to the intervention data. This allows

Table 2 Trial B Fisher Exact Test Method

	Control			Intervention			
Period	Events	Patients	Event Rate (%)	Events	Patients	Event Rate (%)	*P*
1	0	4	0.0	0	1	0.0	—
2	0	13	0.0	0	12	0.0	—
3	1	27	3.7	0	23	0.0	1.000
4	2	40	5.0	1	49	2.0	0.586
5	2	73	2.7	2	83	2.4	1.000
6	3	116	2.6	5	126	4.0	0.724
7	4	168	2.4	7	173	4.0	0.548
8	4	214	1.9	9	226	4.0	0.262
9	7	276	2.5	10	284	3.5	0.624
10	8	337	2.4	12	361	3.3	0.502
11	10	424	2.4	15	450	3.3	0.423
12	11	536	2.1	16	554	2.9	0.438
13	13	639	2.0	18	649	2.8	0.468
14	17	776	2.2	22	780	2.8	0.517
15	17	892	1.9	23	926	2.5	0.427
16	18	1041	1.7	28	1058	2.6	0.180
17	20	1192	1.7	31	1195	2.6	0.156
18	24	1314	1.8	31	1327	2.3	0.414
19	25	1457	1.7	31	1459	2.1	0.500
20	26	1584	1.6	32	1606	2.0	0.509
21	30	1739	1.7	34	1751	1.9	0.706

Note: Major bleeding at 30 days. Cumulative enrollment and events. Shaded period was the closest to significance.

direct comparison of methods but may result in overoptimistic performance measurements when such methods are applied to a prospective patient cohort, which experiences shifts in patient case mix and provider behavior over time. In addition, all of the methods used perform serial evaluations of the data, which can increase the false positive alerting rate. However, these methods are intended for screening large numbers of outcomes for a wide variety of medications and medical devices within an automated application. Such surveillance emphasizes early detection and accepts lower sensitivity for additional specificity in this setting. Because of this, in-depth manual review of identified signals must then be performed in order to determine whether the signal is a true positive. Additional work will be required to satisfactorily adjust the sensitivity and specificity of the alerts to a manageable rate for manual review of the results from this application.

These methods are intended for use in prospective observational cohort surveillance within a health-care environment, whether it is one hospital or a network of hospitals and outpatient clinics. Once a surveillance methodology is validated and established, selection of the baseline or expected data becomes critical for risk-adjustment purposes and defines the nature of any resulting alerts. For example, a medical product just released into the market could use phase 3 trial data as a baseline, which would evaluate whether the observed population experienced safety outcomes in excess of that reference group. However, such trials are well known to recruit healthy patients, and sample sizes are generally low. Alternatively, outcome data from a closely related product with the same indication could be collected in the local environment for this purpose. This has the benefit of a larger sample size and could allow more granular data collection (because data elements in phase 3 trial data are expensive to collect) but might also suffer from missing data or collection, recall, or other biases. Further work must be done in this area to establish data selection hierarchies and protocols in order to inform such a process.

In conclusion, the SPC and LR-SPC methods performed well when evaluating RCT data for significant safety event rate elevations. For monitoring where large amounts of data are available to provide the expected event rate (and threshold), SPC and LR-SPC appear to outperform BUS monitoring. Further work is required to establish risk-adjustment

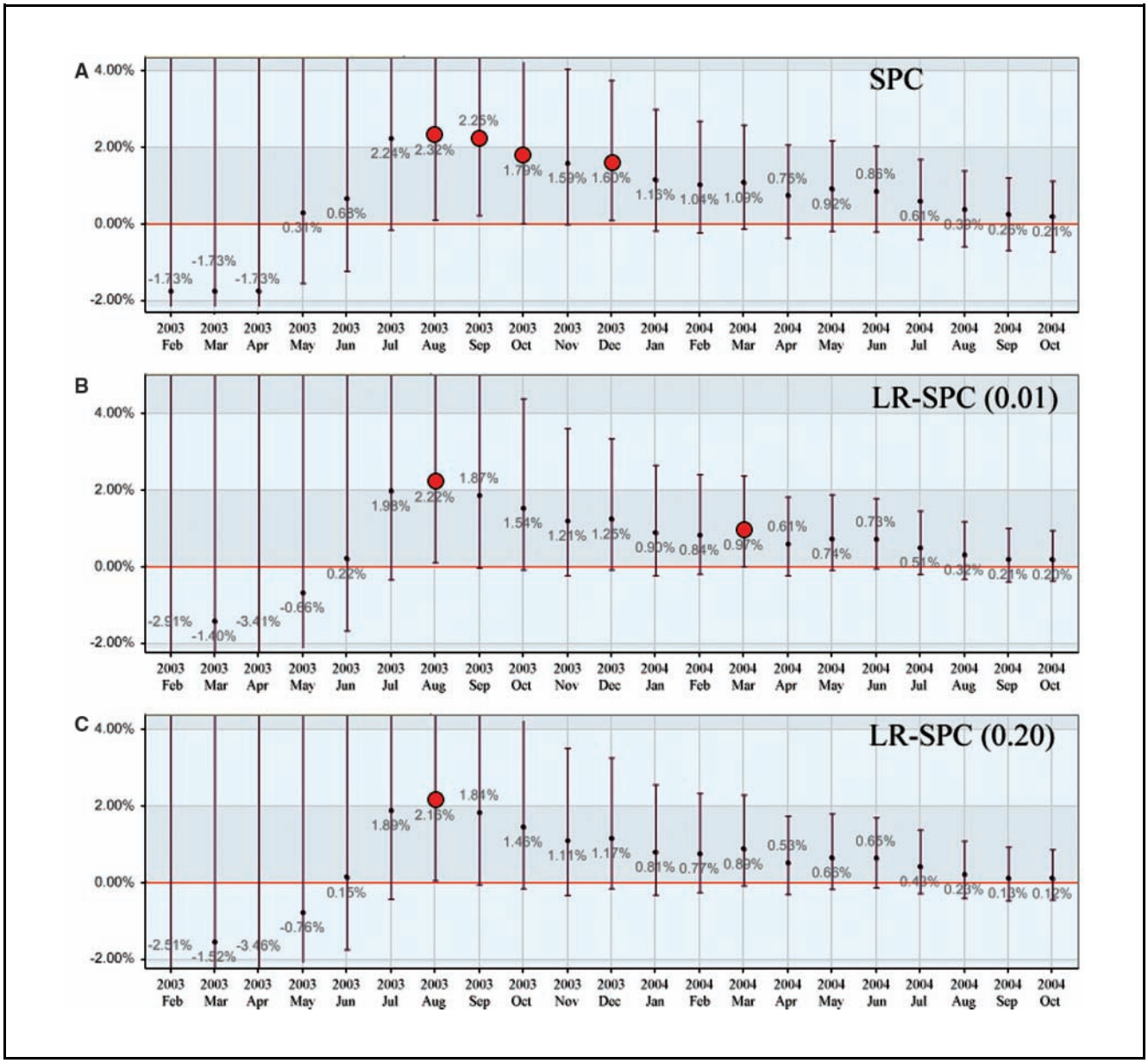

Figure 2 Trial B major bleeding for intervention arm by method. (A) Statistical process control (SPC), (B) logistic regression adjusted statistical process control (0.01) (LR-SPC), and (C) logistic regression adjusted statistical process control (0.20) (LR-SPC).

Table 3 Summary of Performance by Method for Both Trials

Method	TP	FP	FN	TN	Sensitivity	Specificity
SPC	8	4	0	23	100%	85%
BUS	0	0	8	27	0%	100%
LR-SPC (0.01)	8	2	0	25	100%	93%
LR-SPC (0.20)	8	2	0	25	100%	93%

Note: TP = true positive; FP = false positive; FN = false negative; TN = true negative; SPC = statistical process control; BUS = Bayesian updating statistics; LR-SPC = logistic regression adjusted SPC.

performance in the LR-SPC method and to establish BUS performance for event rate monitoring in conditions with sparse prior data or when highly variable trends in safety are present.

ACKNOWLEDGMENTS

The authors thank Anne Fladger and her library staff for their assistance.

Appendix 1 Proportional Difference Method Observed and Expected Trial A Death

Period	Prospective Alert % Difference (95% CI)	SPC Alert % Difference (95% CI)	BUS Alert % Difference (95% CI)	LR-SPC (0.01) Alert % Difference (95% CI)	LR-SPC (0.20) Alert % Difference (95% CI)
1	—	−1.3 (−1.8 to 54.8)	0.0 (−0.6 to 0.6)	−14.1 (−49.5 to 42.4)	−4.2 (−49.2 to 52.0)
2	0.0 (−18.4 to 11.4)	−1.3 (−1.8 to 18.0)	0.0 (−0.6 to 0.6)	−4.2 (−9.8 to 15.2)	−1.5 (−7.2 to 17.8)
3	1.9 (−18.4 to 19.4)	0.6 (−1.1 to 8.8)	0.0 (−0.5 to 0.6)	−0.3 (−2.3 to 7.9)	0.6 (−1.3 to 8.8)
4	3.1 (−0.3 to 7.6)	1.7 (−0.2 to 6.3)	0.1 (−0.5 to 0.6)	1.4 (−0.5 to 5.9)	1.8 (−0.1 to 6.4)
5	1.8 (−0.3 to 4.3)	0.8 (−0.4 to 3.3)	0.1 (−0.5 to 0.6)	0.7 (−0.5 to 3.1)	1.0 (−0.2 to 3.4)
6	1.6 (−0.1 to 3.5)	1.1 (−0.1 to 3.0)	0.1 (−0.4 to 0.7)	1.0 (−0.1 to 2.8)	**1.2 (0.1 to 3.0)**
7	**2.0 (0.7 to 3.4)**	**1.3 (0.3 to 2.7)**	0.3 (−0.3 to 0.8)	**1.3 (0.3 to 2.6)**	**1.4 (0.5 to 2.7)**
8	**2.0 (0.9 to 3.1)**	**1.3 (0.4 to 2.5)**	0.3 (−0.2 to 0.9)	**1.3 (0.5 to 2.4)**	**1.4 (0.6 to 2.5)**
9	**1.7 (0.8 to 2.7)**	**1.1 (0.3 to 2.1)**	0.3 (−0.2 to 0.9)	**1.1 (0.4 to 2.0)**	**1.1 (0.4 to 2.1)**
10	**1.8 (0.9 to 2.7)**	**1.3 (0.5 to 2.2)**	0.5 (−0.1 to 1.0)	**1.3 (0.6 to 2.1)**	**1.3 (0.7 to 2.1)**
11	**1.6 (0.9 to 2.4)**	**1.3 (0.5 to 2.1)**	0.5 (−0.1 to 1.0)	**1.2 (0.6 to 1.9)**	**1.2 (0.7 to 2.0)**
12	**1.1 (0.4 to 1.8)**	**1.0 (0.3 to 1.7)**	0.5 (−0.1 to 1.0)	**1.0 (0.5 to 1.6)**	**1.0 (0.5 to 1.6)**
13	**1.0 (0.3 to 1.6)**	**1.0 (0.4 to 1.6)**	0.5 (−0.1 to 1.0)	**1.0 (0.5 to 1.5)**	**1.0 (0.5 to 1.5)**
14	**1.0 (0.3 to 1.6)**	**1.0 (0.3 to 1.6)**	0.5 (−0.1 to 1.0)	**0.9 (0.5 to 1.5)**	**1.0 (0.5 to 1.5)**

Note: Bold = alert. SPC = statistical process control; BUS = Bayesian updating statistics; LR-SPC = logistic regression adjusted SPC; CI = confidence interval.

Appendix 2 Proportional Difference Method Observed and Expected Trial B Bleeding

Period	Prospective Alert % Difference (95% CI)	SPC Alert % Difference (95% CI)	BUS Alert % Difference (95% CI)	LR-SPC (0.01) Alert % Difference (95% CI)	LR-SPC (0.20) Alert % Difference (95% CI)
1	0.0 (−49.0 to 79.4)	−1.7 (−2.5 to 77.6)	−1.7 (−2.4 to 77.6)	−2.9 (−80.5 to 76.5)	−2.5 (−80.4 to 76.9)
2	0.0 (−22.8 to 24.3)	−1.7 (−2.5 to 22.5)	−1.7 (−2.4 to 22.5)	−1.4 (−9.7 to 22.9)	−1.5 (−9.8 to 22.7)
3	−3.7 (−18.3 to 10.9)	−1.7 (−2.5 to 12.6)	−1.7 (−2.4 to 12.6)	−3.4 (−7.2 to 10.9)	−3.5 (−7.2 to 10.9)
4	−3.0 (−14.6 to 6.4)	0.3 (−1.5 to 9.0)	0.3 (−1.5 to 9.0)	−0.7 (−2.8 to 8.0)	−0.8 (−2.9 to 7.9)
5	−0.3 (−7.3 to 6.0)	0.7 (−1.2 to 6.7)	0.7 (−1.2 to 6.6)	0.2 (−1.7 to 6.2)	0.2 (−1.7 to 6.1)
6	1.4 (−3.9 to 6.7)	2.2 (−0.1 to 7.3)	2.1 (−0.3 to 7.1)	2.0 (−0.3 to 7.0)	1.9 (−0.4 to 6.9)
7	1.7 (−2.5 to 6.0)	**2.3 (0.1 to 6.4)**	2.1 (−0.1 to 6.2)	**2.2 (0.1 to 6.3)**	**2.2 (0.1 to 6.2)**
8	2.1 (−1.3 to 5.7)	**2.3 (0.2 to 5.7)**	2.0 (0.0 to 5.5)	1.9 (0.0 to 5.3)	1.8 (0.0 to 5.3)
9	1.0 (−2.1 to 4.1)	**2.3 (0.2 to 5.7)**	1.5 (−0.2 to 4.4)	1.5 (−0.1 to 4.4)	1.5 (−0.1 to 4.3)
10	1.0 (−1.7 to 3.6)	1.6 (0.0 to 4.0)	1.3 (−0.3 to 3.8)	1.2 (−0.2 to 3.6)	1.1 (−0.3 to 3.5)
11	1.0 (−1.4 to 3.3)	**1.6 (0.1 to 3.8)**	1.3 (−0.2 to 3.4)	1.3 (−0.1 to 3.3)	1.2 (−0.1 to 3.3)
12	0.8 (−1.1 to 2.8)	1.2 (−0.2 to 3.0)	0.9 (−0.4 to 2.7)	0.9 (−0.2 to 2.7)	0.8 (−0.3 to 2.6)
13	0.7 (−1.0 to 2.5)	1.0 (−0.2 to 2.7)	0.8 (−0.4 to 2.4)	0.8 (−0.2 to 2.4)	0.8 (−0.2 to 2.3)
14	0.6 (−1.0 to 2.3)	1.1 (−0.1 to 2.6)	0.8 (−0.4 to 2.3)	**1.0 (0.0 to 2.4)**	0.9 (−0.1 to 2.3)
15	0.6 (−0.8 to 2.0)	0.8 (−0.4 to 2.1)	0.5 (−0.5 to 1.8)	0.6 (−0.2 to 1.8)	0.5 (−0.3 to 1.8)
16	0.9 (−0.4 to 2.2)	0.9 (−0.2 to 2.2)	0.6 (−0.4 to 1.8)	0.7 (−0.1 to 1.9)	0.7 (−0.2 to 1.8)
17	0.9 (−0.3 to 2.1)	0.9 (−0.2 to 2.0)	0.5 (−0.5 to 1.7)	0.7 (0.0 to 1.8)	0.7 (−0.1 to 1.7)
18	0.5 (−0.6 to 1.6)	0.6 (−0.4 to 1.7)	0.4 (−0.5 to 1.4)	0.5 (−0.2 to 1.5)	0.4 (−0.3 to 1.4)
19	0.4 (−0.6 to 1.4)	0.4 (−0.6 to 1.4)	0.2 (−0.6 to 1.2)	0.3 (−0.3 to 1.2)	0.2 (−0.4 to 1.1)
20	0.4 (−0.6 to 1.3)	0.3 (−0.7 to 1.2)	0.1 (−0.6 to 1.0)	0.2 (−0.4 to 1.0)	0.1 (−0.5 to 0.9)
21	0.2 (−0.7 to 1.1)	0.2 (−0.7 to 1.1)	0.1 (−0.6 to 1.0)	0.2 (−0.4 to 1.0)	0.1 (−0.4 to 0.9)

Note: Bold = alert. SPC = statistical process control; BUS = Bayesian updating statistics; LR-SPC = logistic regression adjusted SPC; CI = confidence interval.

REFERENCES

1. Fontanarosa PB, Rennie D, DeAngelis CD. Postmarketing surveillance: lack of vigilance, lack of trust. JAMA. 2004;292(21): 2647–50.

2. Wheeler DJ, Chambers DS. Understanding Statistical Process Control. 2nd ed. Knoxville (TN): SPC Press; 1992.

3. Spiegelhalter D, Grigg O, Kinsman R, Treasure T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. Int J Qual Health Care. 2003;15(1):7–13.

4. Shahian DM, Williamson WA, Svensson LG, Restuccia JD, D'Agostino RS. Applications of statistical quality control to cardiac surgery. Ann Thorac Surg. 1996;62(5):1351.

5. Mohammed MA, Cheng KK, Rouse A, Marshall T. Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. Lancet. 2001;357(9254):463.

6. Williamson GD, Hudson GW. A monitoring system for detecting aberrations in public health surveillance reports. Stat Med. 1999;18(23):3283–98.

7. Clark DE, Cushing BM, Bredenberg CE. Monitoring hospital trauma mortality using statistical process control methods. J Am Coll Surg. 1998;186(6):630.

8. Rogers CA, Ganesh JS, Banner NR, Bonser RS. Cumulative risk adjusted monitoring of 30-day mortality after cardiothoracic transplantation: UK experience. Eur J Cardiothorac Surg. 2005; 27(6):1022–9.

9. Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. Stat Methods Med Res. 2003;12(2):147–70.

10. Steiner SH, Cook RJ, Farewell VT, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. Biostatistics. 2000;1(4):441–52.

11. Matheny ME, Ohno-Machado L, Resnic FS. Risk-adjusted sequential probability ratio test control chart methods for monitoring operator and institutional mortality rates in interventional cardiology. Am Heart J. 2008;155(1):114–20. Epub 2007 Oct 17.

12. Matheny ME, Ohno-Machado L, Resnic FS. Monitoring device safety in interventional cardiology. J Am Med Inform Assoc. 2006;13(2):180–7. Epub 2005 Dec 15.

13. Matheny ME. Development of Statistical Methodologies and Risk Models to Perform Real-Time Safety Monitoring in Interventional Cardiology [master's thesis]. Cambridge: Health Sciences & Technology, Massachusetts Institute of Technology; 2006.

14. Matheny ME, Arora N, Ohno-Machado L, Resnic FS. Rare adverse event monitoring of medical devices with the use of an automated surveillance tool. AMIA Annu Symp Proc. 2007:518–22.

15. Cannon CP, McCabe CH, Wilcox RG, et al. Oral glycoprotein IIb/IIIa inhibition with orbofiban in patients with unstable coronary syndromes (OPUS-TIMI 16) trial. Circulation. 2000;102(2): 149–56.

16. Sabatine MS, Cannon CP, Gibson CM, et al. Addition of clopidogrel to aspirin and fibrinolytic therapy for myocardial infarction with ST-segment elevation. N Engl J Med. 2005;352(12): 1179–89. Epub 2005 Mar 9.

17. Bovill EG, Terrin ML, Stump DC, et al. Hemorrhagic events during therapy with recombinant tissue-type plasminogen activator, heparin, and aspirin for acute myocardial infarction: results of the Thrombolysis in Myocardial Infarction (TIMI), Phase II Trial. Ann Intern Med. 1991;115(4):256–65.

18. Hosmer D, Lemeshow S. Applied Logistic Regression. 2nd ed. San Francisco: Jossey-Bass; 2000.

19. Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. J Biomed Inform. 2005;38(5):367–75. Epub 2005 Mar 26.

20. Siu N, Apostolakis G. Modeling the detection rates of fires in nuclear plants: development and application of a methodology for treating imprecise evidence. Risk Anal. 1986;6(1):43–59.

21. Bayes T. Essay towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond. 1763;53:370–418.

22. 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. Med Decis Making. 2004;24(4):399–407.

23. Standards for Privacy of Individually Identifiable Health Information: Final Rule. Federal Register. 2002;67(157):53182–273.

24. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med. 1998;17(8):873–90.

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

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

27. Cauchemez S, Boelle PY, Donnelly CA, et al. Real-time estimates in early detection of SARS. Emerg Infect Dis. 2006;12(1): 110–3.

28. Zohar S, Chevret S. The continual reassessment method: comparison of Bayesian stopping rules for dose-ranging studies. Stat Med. 2001;20(19):2827–43.