Biostatistics (2014), **15**, 4, *pp.* 665–676
doi:10.1093/biostatistics/kxt057
Advance Access publication on December 29, 2013

Monitoring risk-adjusted medical outcomes allowing for changes over time

STEFAN H. STEINER*, R. JOCK MACKAY

Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Canada N2L 3G1

shsteine@uwaterloo.ca

Summary

We consider the problem of monitoring and comparing medical outcomes, such as surgical performance, over time. Performance is subject to change due to a variety of reasons including patient heterogeneity, learning, deteriorating skills due to aging, etc. For instance, we expect inexperienced surgeons to improve their skills with practice. We propose a graphical method to monitor surgical performance that incorporates risk adjustment to account for patient heterogeneity. The procedure gives more weight to recent outcomes and down-weights the influence of outcomes further in the past. The chart is clinically interpretable as it plots an estimate of the failure rate for a "standard" patient. The chart also includes a measure of uncertainty in this estimate. We can implement the method using historical data or start from scratch. As the monitoring proceeds, we can base the estimated failure rate on a known risk model or use the observed outcomes to update the risk model as time passes. We illustrate the proposed method with an example from cardiac surgery.

Keywords: Exponentially weighted moving average; Likelihood; Score function; Weighted estimating equation.

1. Introduction

There is growing interest in monitoring medical outcomes such as surgical performance to assist in allocation of resources and decision-making. Surgeons, hospital administrators, and other stakeholders want up-to-date estimates of performance to detect changes or to compare individual surgeons at a single or different surgical centers. With procedural, technological, and surgical team-related changes continuously in effect, the surgical failure rate is likely to vary over time. One example is the learning curve (Novick and Stitt, 1999) where we expect the underlying success rate to improve over time as the surgeon and surgical team increase their skills and familiarity with the procedures. While we hope for rapid improvements in performance, there are also many reasons why surgical performance could worsen over time including deteriorating skills for an aging surgeon, poor dynamics in a surgical team, inappropriate changes to the surgical technique, etc. It can be difficult to detect changes over time since they may be relatively small, there is a lot of noise due to patient mix and each individual patient provides little information, especially with a binary outcome.

*To whom correspondence should be addressed.

While process monitoring methods have been used in industrial contexts for decades (Montgomery, 1996), they have only recently gained traction in medical and surgical applications. One complication in the medical context is that to make results from heterogeneous patients comparable, it is necessary to use risk adjustment. Lovegrove *and others* (1997) and Poloniecki *and others* (1998) proposed monitoring the observed minus expected outcomes where expected outcomes are determined from a risk model relating the chance of a successful outcome to patient covariates. This model is fit using data available prior to the monitoring. The expected outcome is calculated assuming this model is known and does not change over time. This approach gives equal weight to all patients, even those from long ago.

Steiner *and others* (2000) give a formal procedure using a cumulative sum (CUSUM) control chart to monitor 30-day mortality after surgery. They incorporated risk adjustment through a logistic regression model and based the CUSUM on likelihood ratio scores defined in terms of a null hypothesis (current failure rate) and an alternative hypothesis based on a clinically important change in the failure rate quantified in terms of an odds ratio. While the risk-adjusted CUSUM has optimality properties because of the use of the likelihood ratio and allows an assessment of uncertainty, it lacks a clear clinical interpretation which has likely limited the use of the method in practice. The CUSUM gives equal weight to all patients following a chart reset after a signal.

Outcomes from recent patients are more relevant to estimating the current failure rate when performance is changing over time. The goal of this article is to provide an intuitive, clinically interpretable tool to monitor surgical performance that incorporates risk adjustment and allows for changes in underlying performance over time. Such a monitoring scheme is valuable in the context of a learning curve or any other situation where we expect the overall surgical performance to change slowly over time. Our proposed method is relatively easy to understand and can include pointwise confidence bands that quantify the estimation uncertainty. We illustrate the methodology with the binary outcome 30-day mortality after cardiac surgery. In the discussion section, we show that the approach is easily adapted to handle other types of outcomes including continuous performance measures or counts.

In the next section, we describe a graphical method based on a weighted estimating equation (WEE) that provides up-to-date estimates of the failure rate for a "standard" patient. We show how to calculate the estimate and corresponding standard errors and discuss implementation with or without historical data. The monitoring approach can be implemented with the assumption that the risk adjustment parameters are known or alternately by updating the estimates of the risk adjustment parameters as data accumulate. Then, in Section 3, we illustrate the method by applying the proposed WEE chart in a cardiac surgery example. Section 4 assesses performance of the proposed chart using a simulation study. In addition, we discuss the setup of the chart and compare it to other proposed monitoring methods. Finally, we provide a discussion and conclusions.

2. WEE MONITORING CHART

To monitor performance, we plot a graph (see Figure 1) of the current estimate of the probability of failure for a preselected "standard" patient versus time t. After the result of each surgery is known, we update the chart. The "standard" patient can be chosen to have an important clinical interpretation. For example, we could select a common or average patient in terms of risk characteristics and procedure type.

We employ a risk adjustment mechanism to make the results of heterogeneous patients comparable. Let

$$y_i = \begin{cases} 1 & \text{if surgery is a failure for patient } i, \\ 0 & \text{otherwise,} \end{cases}$$

and $p_i = \Pr(Y_i = 1)$ be the preoperative risk for patient i from some risk model determined either from historical data or based on surgery-specific risk models published in the literature. The model for p_i

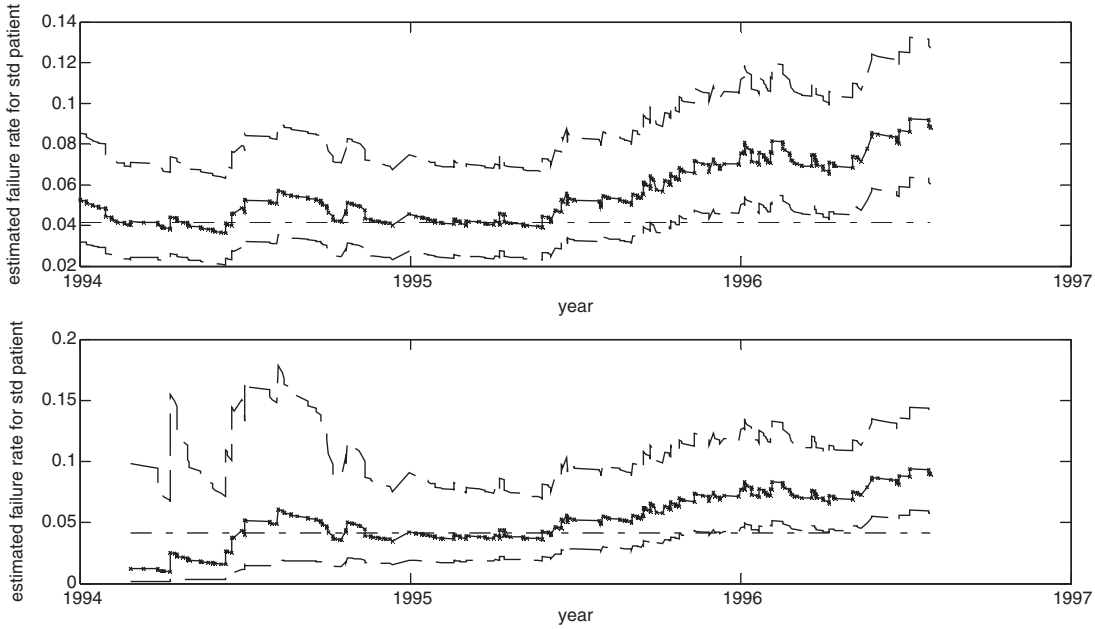

Fig. 1. WEE chart of estimated failure rate for the standard patient with $x_0 = 7$, $\lambda = 0.01$. Surgeon 2 (330 patients), top, use 1992–1993 data in WEE; bottom, start from scratch. Solid thick line, WEE; dashed lines, 95% pointwise confidence bands; horizontal line at 0.041, performance standard.

incorporates the important covariates for patient i such as age, gender, underlying health, etc. that explain the variation in risk. Many different forms of the risk model are possible. We use a logistic regression model $\log(p_i/(1-p_i)) = \alpha + \beta^{\mathrm{T}}(x_i - x_0)$ and so

$$p_i = \frac{\exp(\alpha + \beta^{\mathrm{T}}\{x_i - x_0\})}{1 + \exp(\alpha + \beta^{\mathrm{T}}\{x_i - x_0\})}, \tag{2.1}$$

where x_i is a vector of covariate values for the ith patient and x_0 is the vector of covariate values for the standard patient. The vector β represents the regression coefficients corresponding to the covariates and α represents the log odds of failure for a standard patient with $x_i = x_0$. We discuss the choice of how to define the standard patient in Section 4. To use risk adjustment, we consider two cases:

- assume that β is known with negligible error;
- estimate β using both historical and ongoing monitoring data.

In the second case, we reestimate β each time a new surgical outcome is observed. We assume that the underlying vector β is constant over time. We look at the case where β is assumed known in Subsection 2.1 and the case where we estimate β as time passes in Subsection 2.2.

In either case, we refit a model like (2.1) after each outcome that allows the intercept term α to vary with time. We denote the time dependence by replacing α in (2.1) by α_t. To make the estimates of α_t reflect recent performance and not be unduly influenced by outcomes from long ago, we specify an estimating equation with declining weights for patients further in the past. For simplicity, we use exponentially decaying weights based on patient order. That is, when we have observed a total of t patients, we let the weight for patient i,

$i = 1, 2, \ldots, t$, be

$$w_i = \frac{t\lambda(1-\lambda)^{t-i}}{[1-(1-\lambda)^t]}, \tag{2.2}$$

where λ, $0 < \lambda \leqslant 1$, is a smoothing constant. With (2.2), the weight for the most recent patient is proportional to λ, the patient just before that has a weight proportional to $\lambda(1-\lambda)$, the one before that $\lambda(1-\lambda)^2$, and so on. Thus, using (2.2), the weight for patient i decreases as each new outcome becomes available, i.e. as t increases. In (2.2), the denominator $1-(1-\lambda)^t$ is used to standardize the weights so that $\sum_{i=1}^{t} w_i = t$. As λ approaches zero, the weights for each patient approach 1.

To estimate the model parameters, we use estimating equations based on the score function. With binary outcomes, the log-likelihood for patient i is

$$l = y_i \log(p_i) + (1-y_i)\log(1-p_i). \tag{2.3}$$

Suppose we assume that β is known. Then, the score for patient i is $\partial l/\partial\alpha = (y_i/p - (1-y_i)/(1-p_i))\partial p_i/\partial\alpha = y_i - p_i$ since $\partial p_i/\partial\alpha = p_i(1-p_i)$. Because we believe performance might be changing over time, we estimate α_t, the value of α in (2.1) at time t, using a weighted sum of the scores to get the WEE

$$Q(\hat{\alpha}_t) = \sum_{i=1}^{t} w_i(y_i - p_i) = \sum_{i=1}^{t} w_i(y_i - \{1+\exp[-\hat{\alpha}_t - \hat{\beta}(x_i - x_0)]\}^{-1}) = 0. \tag{2.4}$$

If β is unknown but constant, then using the scores $\partial l/\partial\beta = (y_i - p_i)(x_i - x_0)$, we build a set of standard (un-weighted) estimating equations, as given in (2.5), to reestimate β after each new outcome is observed.

$$Q(\hat{\beta}) = \sum_{i=1}^{t} (y_i - p_i)(x_i - x_0) = \sum_{i=1}^{t} (y_i - [1+\exp\{-\hat{\alpha}_i - \hat{\beta}^{\mathrm{T}}(x_i - x_0)\}]^{-1})(x_i - x_0) = 0. \tag{2.5}$$

The notation hides the fact that (2.5) is a set of equation corresponding to each of the covariates in the risk model (2.1). Because we assume that β is constant over time, the terms of the estimating equations for β are not weighted. Note that (2.5) depends on $\hat{\alpha}_i$, $i = 1, 2, \ldots, t$, and not just the final estimate $\hat{\alpha}_t$ as in (2.4). In (2.5), we use $\hat{\alpha}_i$, $i = 1, 2, \ldots, t$, for two reasons. First, we believe that α may change over time, and thus the final estimate may not well represent the situation earlier in the series. For the same reason, we rely on the weights in (2.4) to accommodate slowly changing values of α. Secondly, we found that the implementation using only $\hat{\alpha}_t$ instead of $\hat{\alpha}_i$, $i = 1, 2, \ldots, t$, was unstable in cases where there were very few recent deaths and especially when any deaths that did occur were for high-risk patients. The instability occasionally resulted in wild values of $\hat{\alpha}_t$ and $\hat{\beta}$. Using (2.5) rather than the alternative with $\hat{\alpha}_i$, $i = 1, 2, \ldots, t$, replaced by $\hat{\alpha}_t$ does not allow $\hat{\beta}$ to change much when a new observation becomes available since the sum of the first $t-1$ terms in (2.5) will be zero with the estimates from the previous time period.

In the case where we assume that β is known, we iteratively solve (2.4) as each new patient outcome is available to provide a current estimate for α_t that is based mostly on recently observed data. When β is also unknown, we solve (2.4) and (2.5) simultaneously. This estimation procedure can be easily programmed, for instance, using the Solver AddIn to Microsoft Excel®.

Note that rescaling all the patient weights by some constant (as done in (2.2)) will factor out of (2.4) and have no effect on the derived estimates. However, the constant is needed to derive approximate standard errors since we wish to recover the usual standard errors with equal patient weights, i.e. as λ approaches zero. Also, we found the rescaling helpful to avoid numerical instability in solving (2.4) for small values of λ.

Using the results in Lipsitz *and others* (1999) or Hu and Kalbfleisch (2000), the variance–covariance matrix for the estimators from (2.4) and (2.5) with α_t rather than α_i, $i = 1, 2, \ldots, t$, has the sandwich form

$$\Sigma = W^{-1} V (W^{\mathrm{T}})^{-1}, \tag{2.6}$$

where W is the expected value of the matrix of derivatives and V is the variance–covariance matrix of the two sets of score equations given by (2.4) and (2.5). In this derivation, we assume that both α and β are constant and hope that it is approximately correct in cases where α_t changes slowly over time. Simulation results presented later in Section 4 suggest the approximation is reasonably good. If there is only a single covariate, i.e. β is a scalar (as in the example in the next section), we get

$$W = -\begin{bmatrix} \sum_{i=1}^{t} w_i p_i (1 - p_i) & \sum_{i=1}^{t} w_i p_i (1 - p_i)(x_i - x_0) \\ \sum_{i=1}^{t} p_i (1 - p_i)(x_i - x_0) & \sum_{i=1}^{t} p_i (1 - p_i)(x_i - x_0)^2 \end{bmatrix},$$

$$V = \begin{bmatrix} \sum_{i=1}^{t} w_i^2 p_i (1 - p_i) & \sum_{i=1}^{t} w_i p_i (1 - p_i)(x_i - x_0) \\ \sum_{i=1}^{t} w_i p_i (1 - p_i)(x_i - x_0) & \sum_{i=1}^{t} p_i (1 - p_i)(x_i - x_0)^2 \end{bmatrix},$$

since $\mathrm{Var}(y_i) = p_i(1 - p_i)$. We then can approximate the standard error of $\hat{\alpha}_t$ using the square root of the first diagonal element of Σ when we substitute the estimates $\hat{\alpha}_t$ and $\hat{\beta}$ into (2.1) to determine the estimated failure probabilities $\hat{p}_i = (1 + \exp\{-\hat{\alpha}_t - \hat{\beta}[x_i - x_0]\})^{-1}$, $i = 1, 2, \ldots, t$. In the determination of the standard error, we use only the final (current) estimates of the model parameters.

If we are willing to assume that the risk model parameters β that quantify the effect of the patient covariates are known, the standard error simplifies to

$$\sqrt{\frac{\sum_{i=1}^{t} w_i^2 \hat{p}_i (1 - \hat{p}_i)}{[\sum_{i=1}^{t} w_i \hat{p}_i (1 - \hat{p}_i)]^2}}. \tag{2.7}$$

Note that had we not used weighting, i.e. if we gave equal weight to all previous patients, i.e. $w_i = 1$ for all i, (2.7) simplifies to the usual expression for the standard error of an estimate for a proportion given by $1/\sqrt{\sum_{i=1}^{t} \hat{p}_i (1 - \hat{p}_i)}$. Given the standard error from (2.7) or, in the general case, the square root of the first diagonal element of Σ, approximate 95% pointwise confidence bands for α_t are given by

$$\hat{\alpha}_t \pm 1.96 \mathrm{SE}(\hat{\alpha}_t). \tag{2.8}$$

To show the results graphically over time, we plot $\hat{p}_{0,t}$, the estimate at time t of the failure rate for a standard patient with covariate value(s) $x_i = x_0$. We calculate $\hat{p}_{0,t}$ from (2.1) by plugging in the estimate $\hat{\alpha}_t$. Note that since $x_i = x_0$, $\hat{p}_{0,t}$ does not depend on β. However, changes in the value of β will affect the estimate of α_t we get from solving (2.4) and (2.5). Also using (2.1), we can translate the endpoints from (2.8) into approximate 95% pointwise confidence bands for $p_{0,t}$. Since the confidence bands as proposed are pointwise, we will need to take care in the interpretation since clearly the series given by $\hat{p}_{0,t}$ will be correlated over time.

3. EXAMPLE

We illustrate the use of the proposed WEE chart with data from a UK center for cardiac surgery (Steiner *and others*, 2000). The data set consists of the outcomes of 6994 operations by seven surgeons in a single surgical center over the period 1992–1998. For each patient, we have the surgery date, surgeon, type of procedure and preoperative covariates including age, gender, presence of hypertension, diabetic status, renal function, and left ventricular mass. These covariates determine the Parsonnet score (Parsonnet *and others*, 1989), a common risk scoring approach for cardiac surgery. To illustrate the methodology, we define the response y_i using 30-day postoperative mortality. In the data, 461 deaths occurred within 30 days of surgery, giving an overall mortality rate of 6.6%. Steiner *and others* (2000) used the available data from all surgeons in the first 2 years (1992 and 1993) to fit a risk adjustment model (2.1) with the Parsonnet score as the only covariate. We define the standard patient as having the median Parsonnet score $x_0 = 7$. The estimates of the parameters in model (2.1) are $\hat{\alpha} = -3.14$ and $\hat{\beta} = 0.077$ with corresponding standard errors of 0.11 and 0.006. With these parameter estimates, the standard patient has a predicted preoperative 30-day mortality risk of 4.1%. The choice of standard patient can be changed if desired.

Suppose that at the beginning of 1994, we had decided to start monitoring the surgical results. We stratify by surgeon and focus here on the results for Surgeon 2 since they provide the most interesting series of outcomes. As a result, we only show monitoring of the surgical outcomes for the roughly 2.5 year period starting in January 1994 since Surgeon 2 left the cardiac center in August 1996. In addition to the 330 surgeries performed by Surgeon 2 in this period, we also have "historical" results (from 1992 to 1993) for Surgeon 2 consisting of 287 patients. In this application, since the data are available, it makes sense to incorporate the 1992–1993 data in the monitoring procedure for Surgeon 2 as we expect only gradual changes in performance. In other words, the historical data are used in (2.4) and, if desired, (2.5) with appropriate weights. For each new patient starting in 1994, we iteratively find new estimates for α_t (and β). Note that here we demonstrate applying the WEE chart retrospectively (since the data are already given). Ideally the method would be employed prospectively as the data arise. In this way, it would be possible to promptly react to any identified problems or concerns.

3.1 *WEE chart with β assumed known*

We first consider the case where we assume that the effect of the Parsonnet scores on the chance of mortality is known, i.e. we assume that β is known. With β known, there is only a single WEE (2.4). The top panel of Figure 1 shows the resulting WEE chart for Surgeon 2 starting in 1994. The crosses in the chart (connected by solid lines) give the estimated failure rate for the standard patient derived with the smoothing constant $\lambda = 0.01$ as suggested by Cook (2003) for binary outcomes. We consider the choice of λ in Section 4. In Figure 1, the two lighter lines give the 95% pointwise confidence bands for the estimated failure rate of the standard patient. We include, for the sake of comparison, a horizontal line at 0.041, the expected failure rate for a standard patient from the Steiner *and others* (2000) model.

The plotted estimates incorporate the results from 1992 to 1993. For instance, when we obtain the result for the first patient operated on by Surgeon 2 in 1994, we estimate $\hat{\alpha}_t$ based on 288 patient results (287 from the period 1992–1993 and the first patient in 1994). In that case the most recent patient has a weight of 3.05 in (2.4) while the first patient from 1992 has a weight of only 0.17. From Figure 1, we see that the predicted standard patient failure rate for Surgeon 2 increases after about the middle of 1995.

We can also apply the WEE chart when we do not have historical data, for instance, with a new surgeon. We assume that the estimated risk model applies but we have no other data for the new surgeon. While not appropriate in this example, for illustration the bottom panel of Figure 1 shows the resulting WEE chart

that ignores the 1992–1993 data for Surgeon 2. With no historical data, we start the plot only after we have observed at least one failure (and one success). Otherwise, the estimate of $p_{0,t}$ is zero and we cannot estimate the corresponding standard error. In this example, the first death that occurred was of patient 22 (February 14, 1994). Comparing the top and bottom panels of Figure 1 shows that, for the initial period of monitoring, we get quite different estimates for the failure rate for the standard patient. Note also that, in the bottom panel of Figure 1, the pointwise confidence bands are very wide at the start since there is little information due to the small number of patients included in the estimation. As the number of observed outcomes increases, the pointwise confidence bands narrow but will not shrink to a width of zero due to the exponential weighting of the patient results. By the end of the series, the two WEE charts in Figure 1 are virtually identical since at that point, due to the exponential weighting, the 287 surgical results from 1992 to 1993 have little effect on the top panel.

The WEE charts in Figure 1 provide a useful interpretable summary of performance that incorporates risk adjustment and allows for a changing underlying rate over time. If we wish to determine whether the observed performance (rescaled to the standard patient) matches some established performance standard, we can check if the performance standard lies within the pointwise confidence bands or not. If not, we conclude there is evidence that the observed performance is better (i.e. if the expected failure rate for the standard patient lies below the pointwise confidence bands) or worse than the performance standard. Compared with the expected failure rate for a standard patient, both WEE plots in Figure 1 suggest that Surgeon 2's performance worsens at the end and according to the pointwise confidence bands is significantly worse than the expected performance. In the top panel of Figure 1, the first time the approximate 95% pointwise confidence bands for $p_{0,t}$ lie completely above the expected rate is after the death of patient 218 in September 1995. Had the method been applied prospectively, the observed increased failure rate for Surgeon 2 could have triggered an investigation into the cause(s) and possible responses.

3.2 *WEE chart with β reestimated after each outcome*

To create Figure 1, we assumed the risk adjustment parameter β was known. This is a reasonable assumption if the risk model was estimated from a large data set and β does not change. However, a number of authors (Jones *and others*, 2004; Jones and Steiner, 2011; Gandy and Kvaloy, 2013) have highlighted the potential large negative effect of estimation error on the performance of control charts generally and risk-adjusted charts in particular. For this reason, we also present an implementation of the WEE chart where we update the estimate of the risk parameter(s) as more data become available. Estimating both model parameters simultaneously is difficult without a reasonable amount of data, especially if the risk model includes many covariates. As a result, the WEE chart with an updating risk model is only feasible when we start with some historical data. Figure 2 gives the resulting WEE chart using an updated risk model when we use the 287 patients from 1992 to 1993 for Surgeon 2 as a starting point. To obtain these results, we simultaneously solve the estimating equations (2.4) and (2.5) and translate the obtained $\hat{\alpha}_t$ to $\hat{p}_{0,t}$ through (2.1). In Figure 2, for comparison, we include a horizontal line at 0.041, the risk estimate from the established risk model for the standard patient. The WEE chart in Figure 2 is different from the top panel of Figure 1 but is qualitatively similar. The biggest difference is that, in Figure 2, where we update the risk model as time passes, the pointwise confidence bands are wider due to the additional uncertainty from estimating β. In addition, the estimated risk $\hat{p}_{0,t}$ is larger at the end of the series in Figure 2 with $\hat{\alpha}_{287+330} = -2.175$, $\hat{\beta} = 0.061$, and $\hat{p}_{0,287+330} = 0.102$. The values of $\hat{\beta}$ across the 330 patients from 1994 to 1996 vary between 0.0561 and 0.0638 and thus are smaller than the previously assumed value derived by fitting a model to 1992–1993 data from all surgeons. With the larger risk estimates and the wider pointwise confidence bands in Figure 2, the first time the whole confidence band exceeds the expected value 0.041 is for patient 235 in November 1995, a little later than in the top panel of Figure 1.

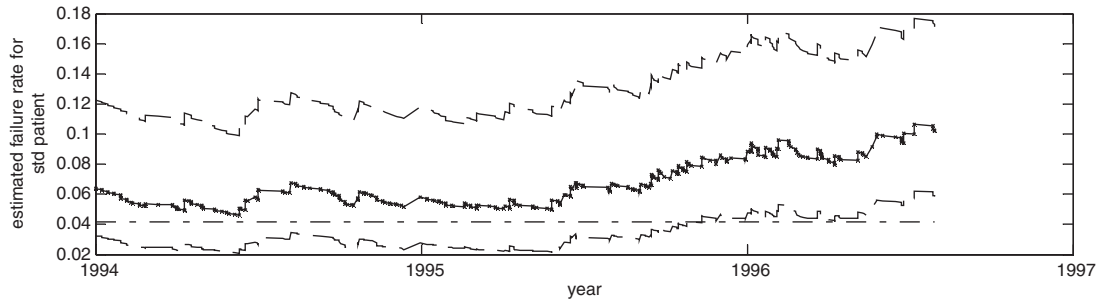

Fig. 2. WEE chart with updating risk model for Surgeon 2 (330 patients) from (2.4) and (2.5) for the standard patient with $x_0 = 7$, $\lambda = 0.01$.

4. PERFORMANCE ASSESSMENT, CHART SETUP, AND METHOD COMPARISON

In the proposed WEE chart, the choice of how to define the standard patient with score x_0 can be important. The best choice is some clinically important patient type with risk that is not too extreme relative to the patient mix. The value of x_0 impacts the level of chart and in the case where we update the estimate only for α_t and assume that β is known, any value of x_0 will yield the same $\hat{p}_{0,t}$ if we use (2.1) to translate the estimated failure rate. There is, however, more uncertainty in the estimates the closer the values are to 0.5 as is the case with a standard binomial model. When we use both (2.4) and (2.5) to estimate α_t and β, the choice of x_0 is more important. Now, because estimating equation (2.4) is weighted and (2.5) is not (and also because we use α_i, $i = 1, 2, \ldots, t$, rather than just α_t in (2.5)), the translated estimate of $p_{0,t}$ depends somewhat on x_0. Values of x_0 in the middle of the distribution of Parsonnet scores yield roughly the same results and we show in the example that using $x_0 = 7$ yields very close to unbiased estimates when α_t does not change. To avoid bias, we recommend not defining the standard patient to be one with an extreme Parsonnet score.

To investigate how well the proposed methodology can capture the uncertainty in the estimate of $p_{0,t}$, we conducted some simulation studies. We generated data from an assumed model (2.1) with a single covariate Parsonnet score distributed according to an exponential distribution with a mean of 8.9 as suggested for this data set by Sego *and others* (2009). We set $\beta = 0.077$ and a known time pattern for α_t. In one case, we used a constant $\alpha = -3.141$ as in Steiner *and others* (2000). In a second case, we allow α_i to change according to the model $\alpha_i = 0.737 \exp(-5i/t) - 3.141$ for $i = 1, 2, \ldots, t$. With this model, over the course of the series, α_i changes from -2.404 to -3.141 and the corresponding $p_{0,t}$ declines from 0.08 to 0.04, in other words is cut in half. In a third case, we allowed α_i to change linearly from -2.404 to -3.141. We looked at both cases where we assumed that β is known and when we reestimated β as time passed. In each simulation, we determined $\hat{p}_{0,t}$ and found the approximate 95% pointwise confidence bands for $p_{0,t}$ using (2.8) and (2.1) based on the final $\hat{\alpha}_t$ derived from (2.4) (and (2.5) when we did not assume β was known) at the end of the series. When we assumed that β was known, each simulation consisted of 50 000 runs and we looked at all possible series lengths t between 40 and 150 in steps of 5. When we used both estimating equations, each simulation consisted of 10 000 runs and we looked at all possible series lengths between 40 and 150 in steps of 10. Note that it does not matter if we consider the data arising prospectively or retroactively since we assess only the final estimates at the end of the series. In the simulation, we determined the bias of $\hat{p}_{0,t}$ and estimated the coverage of the pointwise confidence bands, i.e. the probability that the pointwise confidence bands contain the true value. In any simulation run with no deaths, we set $\hat{p}_{0,t}$ equal to zero and concluded that the pointwise confidence bands did not include the true value. This happened rarely and only for short series. Figures 3 and 4 show the results when we assumed

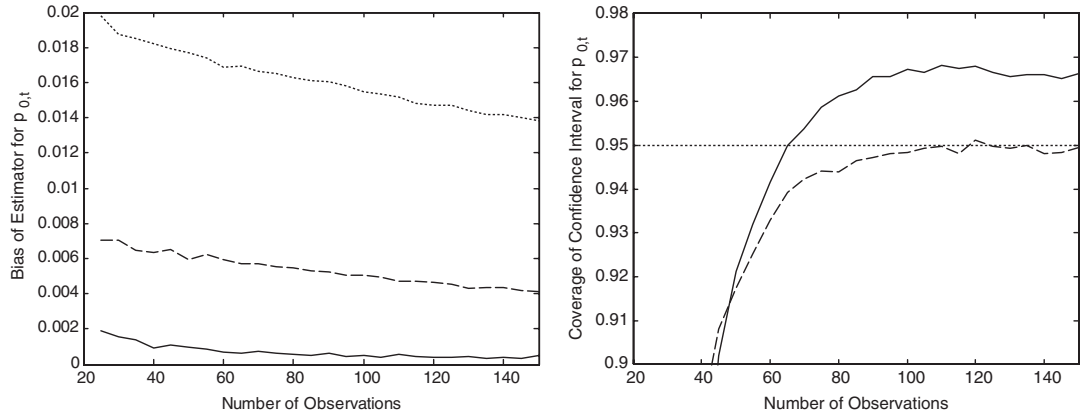

Fig. 3. Simulated bias of $\hat{p}_{0,t}$ and coverage of the pointwise confidence bands for $p_{0,t}$, $\beta = 0.077$ assumed known, and $x_0 = 7$, $p_{0,t} = 0.043$. Solid lines, $\alpha_i = 0.041$; dashed lines,$\alpha_i = 0.737\exp(-5i/t) - 3.141$; dotted lines, $\alpha_i = -2.404 - 0.737i/t$ (not shown in the right panel).

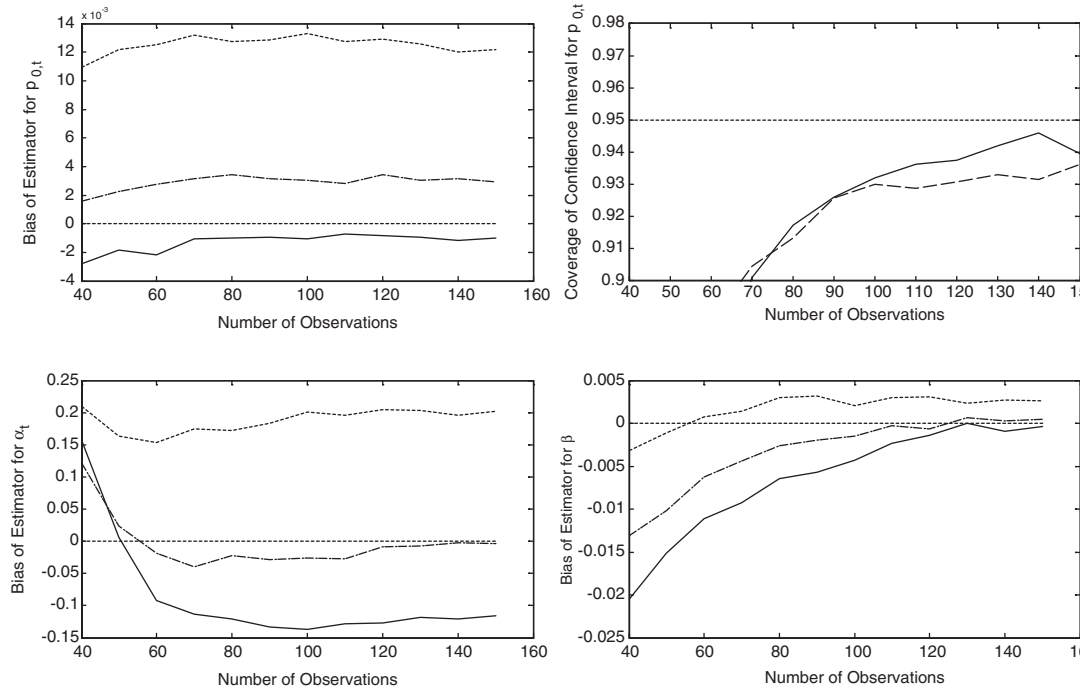

Fig. 4. Simulated bias of $\hat{\alpha}_t$, $\hat{\beta}$, and $\hat{p}_{0,t}$ and coverage of the pointwise confidence bands for $p_{0,t}$ with $\hat{\alpha}_t$, $\hat{\beta}$ estimated using (2.4) and (2.5) and $\hat{p}_{0,t}$ from (2.1). In all cases $\beta = 0.077$, $x_0 = 7$, and $p_{0,t} = 0.043$, solid lines, $\alpha_i = 0.041$; dashed lines, $\alpha_i = 0.737\exp(-5i/t) - 3.141$; dotted lines, $\alpha_i = -2.404 - 0.737i/t$ (not shown in the top right panel due to excessive bias).

that β was known and when we estimated β, respectively. Not surprisingly, the simulations suggest that the estimator for $p_{0,t}$ is close to unbiased when α_i does not change over time, but is biased when α_i (and thus $p_{0,i}$) does change. Figure 4 shows that when β is estimated, its estimator is close to unbiased and any

bias in estimating $p_{0,t}$ comes mostly through the estimate for α_t. In the cases where α_i changes over time, the bias with our approach will be smaller than what we would see with the usual approach of giving equal weight to all patients. In the right panel of Figure 3 and top right panel of Figure 4, we do not show the coverage in the case where α_i changed linearly due to the large bias in that case. We see that the coverage of the pointwise confidence bands is close to the desired 0.95, except when the length of the series is small, say less than 60 when β is known and 100 when β is estimated. Not surprisingly when we run the WEE chart with the risk model being updated, we need a longer series for the pointwise confidence bands to have good coverage properties. Note that, for long series, the approach results in pointwise confidence bands that are slightly conservative (liberal) since their coverage is slightly larger (smaller) than 0.95 when β is known (estimated).

The simulation demonstrates that the WEE (2.4) leads to slightly biased estimates of $p_{0,t}$ when α_i is changing slowly. We are trading increased precision for this bias by down-weighting outcomes from the past. As λ increases, the standard errors will increase while the bias is reduced. We cannot avoid this trade-off unless we are willing to assume an underlying model such as a learning curve to specify how α_t changes over time.

The WEE chart of the weighted score estimates proposed here is similar to the exponentially weighted moving average chart of risk-adjusted outcomes suggested by Grigg and Spiegelhalter (2007). Their approach also produces a trace of an estimate of the risk for a standard patient by down-weighting past observations to allow for the possibility of change. As such, just like the WEE chart their method involves a trade-off of bias for precision and depends on the choice of standard patient. Our proposal differs from the Grigg and Spiegelhalter (2007) approach in a number of important ways. First, with our approach, we can obtain simple pointwise confidence intervals without resorting to a complicated Bayesian dynamic model. Secondly, our approach allows automatic updating of the estimate for β (if desired) as more data become available. When β is estimated, the pointwise confidence intervals also take into account the estimation error. Finally, our approach seems simpler to understand.

Our approach is also similar to that suggested by Cook *and others* (2011) who proposed an exponentially weighted moving average chart of the observed minus expected outcome. We feel that plotting an estimate of the failure rate for a standard patient as in the WEE chart is easier to interpret. In addition, with the Cook *and others* (2011) methodology the expected outcome is always calculated using the fixed-risk model parameter estimates derived before the monitoring began. As a result, our proposed WEE chart will generally be more sensitive to changes in performance since it adapts the risk model as time passes.

As formulated, the proposed WEE charts, illustrated in Figures 1 and 2, are not control charts in the traditional sense since they are not decision-focused and do not have control limits. As a result, they are not directly comparable with risk-adjusted CUSUM charts (Steiner *and others*, 2000). As suggested, if we wish to incorporate a decision limit, we could decide to trigger action if the 95% pointwise confidence bands do not include the expected/desired failure rate. Further evaluation of this use of pointwise confidence bands to make decisions is warranted.

5. SUMMARY AND DISCUSSION

The proposed WEE chart monitoring approach allows for risk-adjusted monitoring of a process where we believe the underlying failure rate may be slowly changing over time. The chart provides an up-to-date estimate of the expected failure rate for a standard patient that gives more weight to recent results. By using the exponentially decaying weights, the monitoring will be sensitive to changes in the underlying failure rate whenever the change might occur.

The estimated failure rate for the standard patient is biased if the rate is changing over time. However, if the change of rate is slow, the bias will be small. Note also that the standard error of the estimated failure

rate is calculated assuming there is no change of rate. Again, if this change is slow, the standard error based on (2.7) will provide a good approximation.

In any application, we must select the smoothing constant λ to define the patient weights given by (2.2). Choosing a value close to unity gives large weights only to very recent surgical outcomes, whereas, in the limit as λ approaches zero, we give equal weight to all observed patients (even those a long time in the past). The best value of λ depends on how quickly performance changes. With binary outcomes, little information is provided by each individual patient, which suggests avoiding large smoothing constants. However, since we believe the underlying failure rate may change over time, using equal weights for all patients is not recommended. In effect, under the assumption that the failure rate is changing over time the choice of λ involves a trade-off between variability and bias in the estimation of α_t and thus $p_{0,t}$.

We can also use the proposed WEE approach to also compare surgeons. One informal option is to simply stratify the data by surgeon and plot the resulting $p_{0,t}$ estimates for each surgeon separately. It is also possible to formally compare surgeons using a hypothesis test.

The WEE methodology is easily adapted to other types of outcomes. For example with a continuous outcome, such as hospital length of stay, suppose that the risk adjustment model is given by the regression model $Y = \alpha + \beta^{\mathrm{T}}(x_i - x_0) + R$. Then, the weighted score equation is $W(\alpha) = \sum w_i(y_i - \mu_i(\alpha))$, where we have denoted the expected outcome for patient i as $\mu_i(\alpha)$ and the parameter α captures the possibly changing average outcome after risk adjustment for the standard patient given by covariate vector x_0. Here, unlike the logistic regression model (2.1), there is a separate parameter that captures the variability of the outcome. If we denote the standard deviation of the model error term R as σ (constant for all combinations of the risk factors), then, assuming the variability does not change, we obtain $\mathrm{SE}(\hat{\alpha}) = \sigma\sqrt{\sum_{i=1}^{t} w_i^2/(\sum_{i=1}^{t} w_i)^2}$, which is the usual result $\sigma/\sqrt{t}$ if we use equal weights for all patients. For a continuous outcome, the results given earlier carry over easily but now it makes sense to use a larger value for the smoothing constant λ in the range 0.05–0.20.

Acknowledgments

The authors would like to thank an anonymous referee and associate editor for substantially improving this paper. In addition, we thank Dr Michael Chu of the London Cardiac Institute, London, Ontario, for fruitful discussions. *Conflict of Interest:* None declared.

Funding

This research was supported, in part, by a Natural Sciences and Engineering Research (NSERC) Discovery Grant.

References

Cook, D. A. (2003). The development of risk-adjusted control charts and machine learning models to monitor the mortality rate of intensive care unit patients, [PhD. Thesis]. School of Information Technology and Electrical Engineering, The University of Queensland.

Cook, D. A., Coory, M. and Webster, R. A. (2011). Exponentially weighted moving average charts to compare observed and expected values for monitoring risk-adjusted hospital indicators. *BMJ Quality and Safety* **20**, 469–474.

Gandy, A. and Kvaloy, J. T. (2013). Guaranteed conditional performance of control charts via bootstrap methods. *Scandinavian Journal of Statistics*, doi:10.1002/sjos.12006.

GRIGG, O. AND SPIEGELHALTER, D. A. (2007). Simple risk-adjusted exponentially weighted moving average. *Journal of the American Statistical Association* **102**, 140–152.

HU, F. AND KALBFLEISCH, J. D. (2000). The estimating function bootstrap. *The Canadian Journal of Statistics* **28**, 449–499.

JONES, L. A., CHAMP, C. AND RIGDON, S. E. (2004). The run length distribution of the CUSUM with estimated parameters. *Journal of Quality Technology* **36**, 95–108.

JONES, M. AND STEINER, S. H. (2011). Assessing the effect of estimation error on risk-adjusted CUSUM chart performance. *International Journal for Quality in Health Care* **24**, 176–181.

LIPSITZ, S. R., IBRAHIM, J. G. AND ZHAO, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association* **94**, 1147–1160.

LOVEGROVE, J., VALENCIA, O., TREASURE, T., SHERLAW-JOHNSON, C. AND GALLIVAN, S. (1997). Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **18**, 1128–1130.

MONTGOMERY, D. C. (1996). *Introduction to Statistical Quality Control*, 3rd edition. New York: John Wiley & Sons.

NOVICK, R. J. AND STITT, L. W. (1999). The learning curve of an academic cardiac surgeon: use of the CUSUM method. *Journal of Cardiac Surgery* **14**, 312–320.

PARSONNET, V., DEAN, D. AND BERNSTEIN, A. D. (1989). A method of uniform stratification of risks for evaluating the results of surgery in acquired adult heart disease. *Circulation* **779** (Suppl. 1), 1–12.

POLONIECKI, J., VALENCIA, O. AND LITTLEJOHNS P. (1998). Cumulative risk-adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *British Medical Journal* **316**, 1697–1700.

SEGO, L. H., REYNOLDS, JR, M. R. AND WOODALL, W. H. (2009). Risk-adjusted monitoring of survival times. *Statistics in Medicine* **28**, 1386–1401.

STEINER, S. H., COOK, R. J., FAREWELL, V. T. AND TREASURE, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* **1**, 441–452.

[*Received June 25, 2013; revised November 14, 2013; accepted for publication November 17, 2013*]