Methodol Comput Appl Probab (2017) 19:1169–1190
DOI 10.1007/s11009-017-9563-6

Performance Monitoring and Competence Assessment in Health Services

Sotirios Bersimis[1] · **Athanasios Sachlas**[1] · **Ross Sparks**[2]

Received: 26 November 2015 / Revised: 4 April 2017 /
Accepted: 17 April 2017 / Published online: 4 May 2017

Abstract Health and health service monitoring is among the most promising research area today and the world work towards efficient and cost effective health care. This paper deals with monitoring health service performance using more than one performance outcome variable (multi-attribute processes), which is common in most health services. Although monitoring whether a health service changes or improves over time is important this is well covered in the current literature. Therefore this paper focuses on comparing similar health services in terms of their performance. The proposed procedure is based on an appropriate control chart. The paper deals with firstly the case when no risk adjustment is required because the health services being compared treat the same patient case-mix which does not vary over time. Secondly it deals with comparing health services where risk adjustment is required because the patient case-mix they service do differ because they service either very different geographical locations or service very different demographics of the same population. The technology developed in this paper could be used for example to assess and compare health practitioners' competence over time, i.e. to decide if two doctors are equivalent in terms of their outcome performances. The waiting time random variable associated with the run length distribution of the control charts (as well as to competence testing) is

The authors are supported by the General Secretariat for Research and Technology (GSRT, Ministry of Education, Greece) research funding action "ARISTEIA II".

✉ Sotirios Bersimis
sbersim@unipi.gr

Athanasios Sachlas
asachlas@unipi.gr

Ross Sparks
Ross.Sparks@csiro.au

[1] Department of Statistics & Insurance Science, University of Piraeus, 80 Karaoli, Dimitriou str., 185 34, Piraeus, Greece

[2] CSIRO, Data61, Pembroke Road, Marsfield, New South Wales 2122, Australia

studied using a Markov Chain embedding technique. Numerical results are provided that exhibit the value of the proposed procedures.

Keywords Competence assessment · Control charting · Health services monitoring · Markov Chain embedding technique · Multi-attribute comparative processes · Process monitoring · Testing competence

Mathematics Subject Classification (2010) 60J20 · 62P10

1 Introduction

In many health applications, the event under study is binary, e.g. the person suffers from a disease or not, the patient survives after a difficult surgery or not, etc. In many cases, processes are characterized by more than one categorical variable and are called multi-attribute processes (Topalidou and Psarakis 2009; Woodall 1997).

A common practice in health-care applications when we work with multi-attribute processes is to use aggregated data (Burkom et al. 2004; Dubrawski 2011), i.e. contingency tables, etc. According to Ryan and Thompson (2002) we are unable to compare several health outcomes without aggregated data. However, data aggregation leads to loss of information and it significantly delays the detection of process shifts (Reynolds and Stoumbos 2000; 2004; Woodall and Montgomery 2004). Schuh et al. (2013) assumed an underlying Poisson process and studied the effect of aggregation. However, the comparison with other methods for varying aggregation periods, showed significant adverse effect of aggregation.

To improve the quality of medical practice and to ensure that a certain minimum standard is maintained, it is essential to monitor and assess doctors' performance and competence. Monitoring the performance of hospitals and health practitioners is a distinguished problem in health sector. A review of methods and systems involved in healthcare, public health, and syndromic surveillance was made by Tsui et al. (2008). Usually, risk adjustment is used to take into consideration the varying health conditions of the patients (Grigg and Farewell 2004; Sego et al. 2009; Steiner et al. 2000; Zhang et al. 2014; Zhang and Woodall 2015).

Recently, Maruthappu et al. (2014) presented a control chart for monitoring performance which takes into account both surgeon-specific and patient-specific characteristics. In this work, we aim in monitoring health practitioners' performance in an innovative way taking into account their relative performance (comparing directly their performance). For example, two endocrine surgeons can be online monitored for checking whether their performance on thyroidectomy under local anaesthesia with respect to two outcome measures - time taken to complete the operation, and pain experienced by the patient - has changed.

Another critical problem in health decision making is the assessment of competence among health practitioners. According to Rethans et al. (1990), competence is what a doctor is capable of doing while performance is what a doctor does in his day-to-day practice. To ensure that a hospital maintains a certain minimum standard of medical practice, it is crucial to hire equally experienced doctors and their performance maintained stable. Assume that two nephrologists perform renal biopsy alternately at the same hospital. The comparison of two doctors, can be seen from an acceptance sampling point of view. In this case, the aim is to assess if one doctor outperforms another doctor. Assume that we have a trainee and a consultant nephrologist and we wish to assess their competence at performing renal biopsy. Such situations were treated in Lim et al. (2002) through a control chart assessing each doctor independently. In this work, we assess the competence by introducing an appropriate hypothesis testing procedure.

To monitor doctors' performance and assess their competence we have to assume that the patients treated are "the same". However, this assumption is at times infeasible. A solution in this case is usually some risk adjustment to correct for patient differences. Thus, in the following section we describe a method for monitoring doctors' performance and assessing their competence, while in Section 6 we briefly discuss how the proposed procedure is applied under risk adjustment, using in this way both homogeneous and non-homogeneous Markov processes for dealing with the problems under study.

To solve several problems arising in health sector, associated to multi-attribute processes without aggregated data, in this paper we present two procedures that are defined on multivariate sequences of trials that aim to performance monitor and competence assess. The exact distribution of interest is studied using the Markov Chain embedding technique (Balakrishnan et al. 2009). We also introduce the necessary details for applying risk adjustment for monitoring two health practitioners' performance.

The paper is organized as follows: In Section 2 we present the proposed monitoring procedures while the methodology for calculating the run length associated distribution, which is based on Markov chain embedding, is given in Section 3. Numerical results are presented in Section 4 and two examples are given in Section 5. Three modifications of the method - the version to detect the deterioration of the performance of both doctors, a runs sum approach, and the risk-adjusted version - are discussed in Section 6 and, finally, in Section 7 we give some concluding remarks.

2 Monitoring Health Practitioners' Performance and Assessing Health Practitioners' Competence

In this section, we will present a unified method for monitoring doctors' performance and assessing doctors' competence.

2.1 Monitoring Health Practitioners' Performance

Assume that a hospital has two surgeons. To assess their performance, we can monitor their performance in a comparative manner, considering the one of them as a reference (e.g. the most experienced one). The basic idea of this paper is to monitor the performance of two doctors in real-time.

2.1.1 Online Monitoring of Health Practitioners' Performance with Respect to One Characteristic

Consider first the case that we monitor the two surgeons - surgeon A and surgeon B - with respect to one characteristic, e.g. the success of the operation. Thus, sequentially surgeries outcome are accrued in terms of a characteristic, say X_s^g, $g = A, B$, (s corresponds to the s-th surgery outcome) which takes two values representing whether the surgery was successful (0) or not (1), i.e.

$$X_s^g = \begin{cases} 0, & \text{if the } s\text{-th operation by surgeon } g \text{ is successful} \\ 1, & \text{if the } s\text{-th operation by surgeon } g \text{ is unsuccessful.} \end{cases}$$

The probability that the s-th operation conducted by surgeon g is successful with respect to X, is denoted by $\pi_{0,s}^g = P(X_s^g = 0)$ while the probability that the s-th operation conducted

by surgeon g is unsuccessful with respect to X is denoted by $\pi^g_{1,s} = P(X^g_s = 1) = 1-\pi^g_{0,s}$. For each pair of patients s, $s = 1, 2, \ldots$, we have two outcomes regarding the success or not of the operations. Table 1 gives an example of such a scheme for the first 15 pairs.

As we may observe in Table 1 at the first pair, the operation conducted by surgeon A is successful while the operation conducted by surgeon B is unsuccessful. At the second pair both surgeons conducted unsuccessful operations. Thus, in Table 1 we have a two dimensional sequence, with the first component referring to surgeon A and the second one referring to surgeon B. Since at each time point, data from a different pair of patients is accrued from the process, we assume that the results of each pair are independent from the previous one. Moreover, the performance of the first doctor is considered independent from the performance of the second doctor.

It is evident that surgeon A may performs well relative to surgeon B if a large number of 1s appear in the sequence of X^B_s. Analogously, surgeon B may performs well relative to surgeon A if a large number of 1s appear in the sequence of X^A_s. In order to monitor the process, we propose to compute the difference between the number of "failures" for surgeon A and the number of "failures" for surgeon B, considering as "failure" for the surgeon g the case that $X^g_s = 1$, $g = A, B$.

For the s-th pair, the probabilities associated with X^A_s, X^B_s, i.e.

$$p^{(s)}_{uv} = P\left(X^A_s = u, X^B_s = v\right), \quad u, v = 0, 1 \tag{1}$$

are

$$p^{(s)}_{00} = \pi^A_{0,s} \cdot \pi^B_{0,s}, \quad p^{(s)}_{10} = \pi^A_{1,s} \cdot \pi^B_{0,s}, \quad p^{(s)}_{01} = \pi^A_{0,s} \cdot \pi^B_{1,s}, \quad p^{(s)}_{11} = \pi^A_{1,s} \cdot \pi^B_{1,s} \tag{2}$$

since the two surgeons operate independently. Assuming the patients are "the same" (patients of equal severity) then the subscript s can be eliminated and the outcomes of the consecutive operations are considered homogeneous (i.e. having constant probabilities of success and failure). However, by introducing the subscript s we permit to the proposed procedure to cover the case that the patients are not "the same" (i.e. they have different characteristics that affect the final outcome of the operation) and they are subjected to risk adjustment (and as a result we permit the probabilities of success and failure varying for different patients).

Thus, having at hand the sequence of random variables X^A_s, X^B_s, we define a random variable D_n that is the cumulative difference between the number of "failures" for surgeon B and the number of "failures" for surgeon A, i.e.

$$D_n = \sum_{s=1}^{n} \left(X^B_s - X^A_s\right).$$

The difference $X^B_s - X^A_s$ can take the values -1, 0, or 1, thus it is obvious that D_n is a discrete variable taking values in $\mathbb{Z}$ with

$$E(D_n) = E\left(\sum\nolimits_{s=1}^{n} (X^B_s - X^A_s)\right) = \sum\nolimits_{s=1}^{n} E\left(X^B_s - X^A_s\right) = \sum\nolimits_{s=1}^{n} \left(E(X^B_s) - E(X^A_s)\right) = 0,$$

Table 1 A realization of a health process with the outcomes of the first 15 pairs of patients

Pair of patients (s)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
X^A_s	0	1	1	1	1	0	0	0	1	0	0	0	0	1	0	…
X^B_s	1	1	0	1	1	0	0	1	0	1	1	1	0	1	1	…

if the two doctors are equally effective and the probabilities of success of the two doctors are constant across the pairs of patients (assuming that the patients are "the same"). The number of "failures" for each surgeon at each time point (pair of operated patients), given in Table 1, is presented in Table 2. In this table we observe how the difference is evolving from one time point to the other.

If the difference exceeds a value k we declare that the surgeon B do not performs well, while if the difference is lower than $-k$ we declare that surgeon B performs well relative to surgeon A.

If instead of doctors we would like to monitor treatments, drugs, etc we can use the same methodology as above. A similar approach was used by Bersimis et al. (2015) in the context of clinical trials. However, this approach demands the difference of the number of successes of the 2 doctors to exceed a specific threshold while Bersimis et al. (2015) did not take the difference into account.

2.1.2 Online Monitoring of Health Practitioners' Performance with Respect to Two Characteristics

Let us now investigate how we can assess the surgeons' performance with respect to two dependent dichotomous characteristics. For example, we compare their performance with respect to the success of the surgery (X_1) and the occurrence of postoperative complications (X_2) (see e.g. Steiner et al. (2000)). X_1 and X_2 are binary variables defined, for pair of patients s, as

$$X_{i,s}^{g} = \begin{cases} 0, & \text{if the } s\text{-th operation conducted by surgeon } g \text{ is successful in terms of the } i\text{-th characteristic} \\ 1, & \text{if the } s\text{-th operation conducted by surgeon } g \text{ is unsuccessful in terms of the } i\text{-th characteristic,} \end{cases}$$

$i = 1, 2$.

To monitor the process, we propose again to use D_n. If the differences are small, then the two surgeons are equally efficient. If the difference exceeds a limit k we declare that surgeon A performs well relative to surgeon B, while if the difference is lower than $-k$ we declare that surgeon B is more efficient.

In order to determine the value of limit k, we define two new random variables that count the cases showing "failure" for surgeon A and surgeon B, respectively. Thus, on the sequence of events $\{X_{1,s}^{A}, X_{2,s}^{A}\}$, $\{X_{1,s}^{B}, X_{2,s}^{B}\}$, we define the two dimensional random variable (X_s^A, X_s^B) as follows:

$$X_s^A = X_{1,s}^A + X_{2,s}^A, \quad X_s^B = X_{1,s}^B + X_{2,s}^B.$$

(X_s^A, X_s^B) takes values on $\{0, 1, 2\} \times \{0, 1, 2\}$.

Table 2 Number of "failures" and cumulative number of "failures" for each surgeon at each pair of patients of the surgical process with one characteristic

Pair of patients (s)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
X_s^A	0	1	1	1	1	0	0	0	1	0	0	0	0	1	0	...
X_s^B	1	1	0	1	1	0	0	1	0	1	1	1	0	1	1	...
$X_s^B - X_s^A$	1	0	–1	0	0	0	0	1	–1	1	1	1	0	0	1	...
$D_n = \sum_{s=1}^{n} (X_s^B - X_s^A)$	1	1	0	0	0	0	0	1	0	1	2	3	3	3	4	...

The probabilities associated with X_s^A, X_s^B, i.e.

$$p_{uv}^{(s)} = P(X_s^A = u, X_s^B = v), \quad u, v = 0, 1, 2, \tag{3}$$

are

$$\begin{array}{lll} p_{00}^{(s)} = \pi_{00,s}^A \cdot \pi_{00,s}^B, & p_{10}^{(s)} = \pi_{10,s}^A \cdot \pi_{00,s}^B + \pi_{01,s}^A \cdot \pi_{00,s}^B, & p_{20}^{(s)} = \pi_{11,s}^A \cdot \pi_{00,s}^B, \\ p_{01}^{(s)} = \pi_{00,s}^A \cdot \pi_{10,s}^B + \pi_{00,s}^A \cdot \pi_{01,s}^B, & p_{11}^{(s)} = \pi_{10,s}^A \cdot \pi_{01,s}^B + \pi_{01,s}^A \cdot \pi_{10,s}^B + \pi_{01,s}^A \cdot \pi_{01,s}^B + \pi_{10,s}^A \cdot \pi_{10,s}^B, & p_{21}^{(s)} = \pi_{11,s}^A \cdot \pi_{01,s}^B + \pi_{11,s}^A \cdot \pi_{10,s}^B, \\ p_{02}^{(s)} = \pi_{00,s}^A \cdot \pi_{11,s}^B, & p_{12}^{(s)} = \pi_{10,s}^A \cdot \pi_{11,s}^B + \pi_{01,s}^A \cdot \pi_{11,s}^B, & p_{22}^{(s)} = \pi_{11,s}^A \cdot \pi_{11,s}^B, \end{array} \tag{4}$$

with

$$\pi_{ij,s}^g = P(X_{1,s}^g = i, X_{2,s}^g = j), \quad i, j = 0, 1, \quad g = A, B.$$

The $\pi_{ij,s}^g$'s are assumed to be known or at least estimated from a large enough historical data set.

Thus, having at hand the sequence of random variables X_s^A, X_s^B, we define the random variable

$$D_n = \sum_{s=1}^{n} (X_s^B - X_s^A).$$

Again, it is obvious that $X_s^B - X_s^A$ is a discrete variable taking values in $C \subseteq \mathbb{Z}$, $C = \{-2, -1, 0, 1, 2\}$ with expected value equal to zero under the assumption that the probabilities of success of the two doctors are equal and constant across the pairs of patients. The number of "failures" for each surgeon and the evolvement of the difference during time of a hypothetical health process is given in Table 3.

Analogously, we can extend the method to be used for a process with more than two characteristics.

2.1.3 Online Monitoring of Health Practitioners' Performance Taking into Account the Appearance of a Severe Complication by the Doctor'S Fault

Consider now the case where during the operation, a patient experiences death or another severe complication (*SC*) by assumed fault of the surgeon *A* or *B*. We assume that this case immediately terminates the comparison in favor of surgeon *B* or surgeon *A*.

In this case, X_s^A and X_s^B take the values 0, 1, 2, and ξ, where ξ means the occurrence of a *SC*. Thus, here the probabilities $p_{uv}^{(s)}$ associated with X_s^A and X_s^B are given by Eq. 3 and $p_{SC}^{(s)}$ corresponds to the probability of a *SC* in a patient treated by one of the two doctors. The $p_{SC}^{(s)} = 1 - \sum_{u,v} p_{uv}^{(s)}$.

Table 3 Number of "failures" and cumulative number of "failures" for each surgeon at each pair of patients of the surgical process with two characteristics

Pair of patients (s)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
$X_{1,s}^A$	0	1	1	1	1	1	1	0	1	1	1	0	0	1	0	...
$X_{2,s}^A$	0	0	1	0	0	0	1	1	0	0	0	1	0	0	1	...
$X_{1,s}^B$	1	1	0	1	1	0	0	1	0	1	1	1	1	0	0	...
$X_{2,s}^B$	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	...
$X_s^B - X_s^A$	2	1	–1	0	1	0	–1	0	0	1	1	1	2	0	-1	...
$D_n = \sum_{s=1}^{n} (X_s^B - X_s^A)$	2	3	2	2	3	3	2	2	2	3	4	5	7	7	6	...

In this case, we want three rules to assess comparatively the performance of the two surgeons:

- if $-k < D_n < k$ the two surgeons are equally efficient,
- if $D_n > k$ or we observe a *SC* by assumed fault of surgeon B, surgeon A is more efficient,
- if $D_n < -k$ or we observe a *SC* by assumed fault of surgeon A, surgeon B is more efficient.

2.1.4 Run Length Distribution of the Monitoring Scheme

In Section 3 we will study the run length distribution of the proposed procedure, by studying the waiting time L_D defined formally as

$$L_{D_n} = \min\{n : D_n > k \text{ or } D_n < -k\}, \quad n = 1, 2, \ldots$$

for the cases presented in Sections 2.1.1 and 2.1.2 and

$$L_{D_n} = \min\{n : D_n > k \text{ or } D_n < -k \text{ or a } SC \text{ happened}\}, \quad n = 1, 2, \ldots$$

for the case of Section 2.1.3.

Someone could argue why not to use the method exploiting the multinomial CUSUM chart proposed by Ryan et al. (2011). This method requires a prior knowledge of the direction of the shift. However, it is unlikely in practice to know in advance if the performance of a doctor has improved or worsened. In our method we do not need to know the direction of the shift. Moreover, although the method of Ryan et al. (2011) can be applied to the cases of Sections 2.1.1 and 2.1.2, it cannot be applied when a *SC* occurs.

2.2 Assessment of Two Doctors' Competence

When we want to assess two doctors' competence, we apply a strategy similar to acceptance sampling strategy. This resembles the designs of Bersimis et al. (2015) for the early termination of Phase II comparative clinical trials.

2.2.1 Comparison of Doctors' Competence with Respect to One and Two Characteristics

Assume that we want to evaluate whether a trainee surgeon (T) has sufficient experience and training to take on more severe cases. For this reason we can compare his/her competence with that of a consultant surgeon (C). The framework is the same as in the case of online monitoring with respect to one characteristic. Again, we consider that the two surgeons operate alternately on similar incidents as patients enter the operating room sequentially and we define X^g, $g = T, C$ for the s-th pair of patients, as

$$X_s^g = \begin{cases} 0, & \text{if the } s\text{-th operation by surgeon } g \text{ is successful} \\ 1, & \text{if the } s\text{-th operation by surgeon } g \text{ is unsuccessful.} \end{cases}$$

The probabilities associated with X_s^T, X_s^C, i.e. $p_{uv}^{(s)} = P(X_s^T = u, X_s^C = v)$, $u, v = 0, 1$, are the same as in Eq. 2.

It is again evident that surgeon C may be considered better than T if larger number of 1s than expected appeared in the sequence of X_s^T. Thus, here we are not interested in monitoring the performance of the surgeons over time, but whether there is evidence that the

two surgeons are not equivalent in performance. Here, the hypothesis is tested by observing whether D_n exceeds the threshold value k.

The generalization to the two characteristics is straightforward. The only difference is that the probabilities $p_{uv}^{(s)} = P(X_s^T = u, X_s^C = v)$, $u, v = 0, 1$, are not those given in Eq. 2 but those given in Eq. 4.

The waiting time L_{D_n} defined in the previous subsection can be used in order to define a sequential acceptance sampling scheme, in the sense that if one of the two doctors is significantly better than the other, we will quickly identify a value of D_n beyond the thresholds k or $-k$.

3 The Exact Distribution of L_{D_n}

The aim of this section is to establish the exact distribution of the discrete random variable L_{D_n}. To this end, we will exploit the widely used Markov Chain embedding technique (Balakrishnan et al. 2009).

In light of this, we first define an appropriate Markov Chain $\{V_s, s = 0, 1, 2, \ldots\}$, and then its state space and its transition probability matrix. The state space, which has $d = 2k + 1$ elements, has the form $\Omega = \{i : i = -k, \ldots, k\} \cup \{\alpha_d\}$, where α_d is the absorbing state and the indice i tracks the difference D_n. The Markov Chain is defined as

$$V_n = \begin{cases} i, & \text{if } D_n = i \\ \alpha_d, & \text{if} D_n > k \text{ or } D_n < -k \text{ or } SC \text{ happened,} \end{cases}$$

$n = 1, 2, \ldots$. By convention, it holds that $V_0 = 0$. The state α_d corresponds to all the possible absorbing states of the chain. Absorbing are all the states of the form i, for $i \geq k+1$ or $i \leq -(k+1)$. The transition probability matrix is

$$\mathbf{\Lambda}_{0,j}^{(s)} = [P(V_s = \alpha_\ell | V_{s-1} = \alpha_{\ell'})]_{d \times d},$$

where here $\alpha_{\ell'}$ and α_ℓ are states of Ω and s corresponds to the s-th patient. This setup provides us with the necessary tool to handle the case that the patients are not "the same" as in the case of risk adjustment which is presented at Section 6. The index j corresponds to the number of the characteristics we treat. Thus, $\mathbf{\Lambda}_{0,1}^{(s)}$ is the transition probability matrix in the case of one characteristic and $\mathbf{\Lambda}_{0,2}^{(s)}$ is the transition probability matrix in the case of two characteristics.

According to the Markov Chain embedding technique it holds that

$$P(L_{D_n} \leq n) = P(V_n = \alpha_d) = \boldsymbol{\pi}_0' \left(\prod_{s=0}^{n} \mathbf{\Lambda}_{0,j}^{(s)} \right) \mathbf{e}_d, \tag{5}$$

and

$$P(L_{D_n} \leq n) = P(V_n = \alpha_d) = \boldsymbol{\pi}_0' \mathbf{\Lambda}_{0,j}^{n} \mathbf{e}_d, \tag{6}$$

when $\mathbf{\Lambda}_{0,j}^{(s)} = \mathbf{\Lambda}_{0,j}$, i.e. all the patients assumed "the same" (patients of equal severity), $\boldsymbol{\pi}_0'$ is the $1 \times d$ vector of initial probabilities (vector of 0s, except the middle element which is equal to 1, since the starting point of the Markov chain is 0), and finally $\mathbf{e}_d'$ is an $1 \times d$ vector of the form $(0, 0, \ldots, 1)$. Then, the probability distribution of L_{D_n} may be calculated using the formula

$$P(L_{D_n} = n) = P(L_{D_n} \leq n) - P(L_{D_n} \leq n-1) = \boldsymbol{\pi}_0' \left(\prod_{s=0}^{n} \mathbf{\Lambda}_{0,j}^{(s)} - \prod_{s=0}^{n-1} \mathbf{\Lambda}_{0,j}^{(s)} \right) \mathbf{e}_d$$

and specifically for the case that there is not risk adjustment we may use the

$$P(L_{D_n} = n) = P(L_{D_n} \leq n) - P(L_{D_n} \leq n-1) = \boldsymbol{\pi}_1'(\boldsymbol{\Lambda}_{1,j})^{n-1}\mathbf{h},$$

where $\boldsymbol{\pi}_1$ is a $(d-1)\times 1$ vector with elements all the entries of $\boldsymbol{\pi}_0$ except the last one (this vanishes for non-degenerate cases) and $\boldsymbol{\Lambda}_{1,j}$ is matrix $\boldsymbol{\Lambda}_{0,j}$ without the last column and the last row. The vector $\mathbf{h}$ equals $\mathbf{1} - \boldsymbol{\Lambda}_{1,j}\mathbf{1}$, where $\mathbf{1}$ is a $(d-1)\times 1$ vector containing ones; exploiting the fact that the transition probability matrix $\boldsymbol{\Lambda}_{0,j}$ is stochastic. The vector $\mathbf{h}$ is filled up with the transition probabilities from the corresponding state to the absorbing state (for more details see Balakrishnan et al. (2009) and references therein).

The transition probability matrix as well as the vectors of initial probabilities for the cases of one and two characteristics is given and described in Appendices A.1 and A.2, respectively. Guidelines for the case that risk adjustment is applied is given in Section 6.

4 Numerical Results

In this section we present examples of monitoring the performance of two surgeons with respect to one and two characteristics. Moreover, we present examples of monitoring two surgeons taking into account the appearance of a *SC*, and of assessing two doctors' competence with respect to two characteristics.

4.1 Monitoring Doctors' Performance

Online monitoring of two doctors' performance with respect to one characteristic

Assume that the surgeons A and B acquiring patients of equal severity conduct successful operations with probability $\pi_0^A = \pi_0^B = 0.950$ and unsuccessful operations with probability $\pi_1^A = \pi_1^B = 0.050$. This is the in-control case (*ICC*) under the *Case A* given in Table 4. The same table also shows 3 different out-of-control cases (*OCC*): *OCC A*, *OCC B*, and *OCC C* which represent that although surgeon's A performance remains constant, the performance of surgeon B deteriorates. The 3 *OCC*s are ordered according to the value of a "non-centrality like statistic" defined as

$$\lambda_1 = \sqrt{\sum_i \frac{\left(\pi_i^{ICC} - \pi_i^{OOC}\right)^2}{\pi_i^{ICC}}}.$$

Table 4 The *ICC* and the *OCC*s for online monitoring of two doctors' performance with respect to one characteristic

	ICC				*OCC A*				*OCC B*				*OCC C*			
	π_0^A	π_0^B	π_1^A	π_1^B	π_0^A	π_0^B	π_1^A	π_1^B	π_0^A	π_0^B	π_1^A	π_1^B	π_0^A	π_0^B	π_1^A	π_1^B
Case A	0.95	0.95	0.05	0.05	0.95	0.90	0.05	0.10	0.95	0.80	0.05	0.20	0.95	0.75	0.05	0.25
Case B	0.85	0.85	0.15	0.15	0.90	0.85	0.10	0.15	0.93	0.85	0.07	0.15	0.95	0.85	0.05	0.15

Table 5 shows the average run length (*ARL*), the standard deviation of the run length (*SDRL*), and the percentage *ARL* improvement ($PI\%$) for the three *OCC*s for k from 4 to 10. According to Antzoulakos and Rakitzis (2008), the *ARL* is the mean of L_{D_n}, i.e.

$$ARL = E(L_{D_n}) = \boldsymbol{\pi}_1'(\mathbf{I} - \boldsymbol{\Lambda}_{1,j})^{-1}\mathbf{1},$$

while *SDRL*, is the standard deviation of L_{D_n}, i.e.

$$SDRL = \sqrt{2\boldsymbol{\pi}_1'\boldsymbol{\Lambda}_{1,j}(\mathbf{I} - \boldsymbol{\Lambda}_{1,j})^{-2}\mathbf{1} + \boldsymbol{\pi}_1'(\mathbf{I} - \boldsymbol{\Lambda}_{1,j})^{-1}\mathbf{1} - \left(\boldsymbol{\pi}_1'(\mathbf{I} - \boldsymbol{\Lambda}_{1,j})^{-1}\mathbf{1}\right)^2}.$$

Let us consider *OCC B* which claims that the probability of successful operation by surgeon *B* has decreased from 95% to 80%. For $k = 5$, the in-control *ARL* (ARL_0) equals to 378.947 while the out-of-control *ARL* (ARL_1) equals to 39.993, which corresponds to a 89.45% decrease. From Table 5 we observe that as k increases, the $PI\%$ increases too. Furthermore, as λ_1 increases, the $PI\%$ increases too.

Table 4 also presents *Case B* which corresponds to the case where the performance of surgeon *A* improves while surgeon *B* remains at the same level. The results for this case are similar to those of *Case A*, indicating that the proposed method has good performance regardless whether the performance of a doctor improves or worsens.

Here arises the question which k should the researcher choose. There are two options about the selection of k: the first one is according the desired ARL_0, while the latter is according the maximum $PI\%$ for given shift. Assume that by design we set the ARL_0 to be almost 370.4 which is the ARL_0 for a standard Shewhart control chart. Thus, for *Case A*

Table 5 The *ARL*, *SDRL*, and the $PI\%$ for online monitoring of two doctors' performance with respect to one characteristic

		ICC		*OCC A*			*OCC B*			*OCC C*		
	k	*ARL*	*SDRL*	*ARL*	*SDRL*	$PI\%$	*ARL*	*SDRL*	$PI\%$	*ARL*	*SDRL*	$PI\%$
Case A	4	263.158	216.398	95.342	65.832	63.77%	33.306	17.473	87.34%	24.995	12.108	90.50%
	5	378.947	310.942	117.319	76.265	69.04%	39.993	19.190	89.45%	29.999	13.274	92.08%
	6	515.789	422.674	138.510	84.870	73.15%	46.665	20.741	90.95%	34.999	14.339	93.21%
	7	673.684	551.595	159.191	92.181	76.37%	53.333	22.177	92.08%	40.000	15.330	94.06%
	8	852.632	697.705	179.568	98.595	78.94%	59.999	23.523	92.96%	45.000	16.260	94.72%
	9	1052.630	861.005	199.773	104.387	81.02%	66.667	24.795	93.67%	50.000	17.139	95.25%
	10	1273.680	1041.490	219.882	109.733	82.74%	73.333	26.006	94.24%	55.000	17.976	95.68%
	λ_1	0.000		0.229			0.688			0.918		
Case B	4	98.039	80.237	81.992	62.096	16.37%	60.760	40.175	38.02%	49.765	29.110	49.24%
	5	141.176	115.458	105.924	77.687	24.97%	74.102	45.784	47.51%	59.916	32.224	57.56%
	6	192.157	157.083	129.431	91.721	32.64%	87.052	50.413	54.70%	69.971	34.934	63.59%
	7	250.980	205.113	152.287	104.116	39.32%	99.781	54.397	60.24%	79.990	37.393	68.13%
	8	317.647	259.546	174.487	115.005	45.07%	112.395	57.955	64.62%	89.997	39.678	71.67%
	9	392.157	320.383	196.121	124.604	49.99%	124.950	61.219	68.14%	99.999	41.830	74.50%
	10	474.510	387.624	217.304	133.141	54.20%	137.477	64.271	71.03%	110.000	43.874	76.82%
	λ_1	0.000		0.140			0.224			0.280		

we choose $k = 5$ as this is the value that gives $ARL_0 \approx 370.4$. Alternatively, if we decide to achieve a decrease in *ARL* of at least 90%, we should select $k = 6$ as this is the lowest value that gives $PI\% \geq 90\%$.

Online monitoring of two doctors' performance with respect to two characteristics

Assume now the in-control case *ICC* under *Case A* given in Table 6, in which the probability of an unsuccessful operation is small. Table 6 also shows 3 *OCC*s, which represent increasing probability of unsuccessful operations. The *OCC*s are ordered according to the value of a "non-centrality like statistic" defined as

$$\lambda_2 = \sqrt{\sum_i \sum_j \frac{\left(\pi_{ij}^{IC} - \pi_{ij}^{OOC}\right)^2}{\pi_{ij}^{IC}}}.$$

As *Case B* (also given in Table 6) we consider the case where $\pi_{00} = 0.050, \pi_{01} = 0.100, \pi_{10} = 0.100, \pi_{11} = 0.750$. Here we assume that the probability of an unsuccessful operation is large. Three different *OCC*s representing decreasing probability of complications are shown in Table 6, as well. The 3 out-of-control cases are again ordered according to the value of λ_2.

Table 7 shows the *ARL*, the *SDRL*, and the $PI\%$ for various threshold values of k. Assume *OCC B* under *Case A*. The ARL_0 equals to 247.982 while the ARL_1 equals to 23.872, which corresponds to a 90.37% decrease. From Table 7 we observe that as k increases the percentage improvement $PI\%$ increases too. Furthermore, as λ_2 increases $PI\%$ increases too. Similar results hold for *Case B*, indicating again that the proposed method has good performance regardless whether the probability of unsuccessful operations is small or large.

The choice of k can also be chosen according to the desired in-control *ARL* or the desired percentage *ARL* reduction.

Online monitoring of doctors' performance taking into account the appearance of a severe complication by the doctor's fault.

Assume now the *ICC* under *Case C* in Table 6, in which the probability of the occurrence of a severe complication by a surgeon's assumed fault is 0.005. The 3 *OCC*s shown in

Table 6 The *ICC* and the *OCC*s for online monitoring of two doctors' performance with respect to two characteristic

			ICC		*OCC A*		*OCC B*		*OCC C*	
			X_2		X_2		X_2		X_2	
			0	1	0	1	0	1	0	1
Case A	X_1	0	0.950	0.020	0.850	0.050	0.750	0.090	0.650	0.125
		1	0.020	0.010	0.050	0.050	0.090	0.070	0.125	0.090
Case B	X_1	0	0.050	0.100	0.100	0.125	0.200	0.100	0.250	0.125
		1	0.100	0.750	0.125	0.650	0.100	0.600	0.125	0.500
Case C	X_1	0	0.950	0.020	0.900	0.025	0.850	0.030	0.800	0.050
		1	0.020	0.005	0.030	0.040	0.040	0.075	0.050	0.095
Case D	X_1	0	0.300	0.100	0.100	0.125	0.050	0.100	0.010	0.020
		1	0.100	0.500	0.125	0.650	0.100	0.750	0.020	0.950

Table 7 The *ARL*, *SDRL*, and the $PI\%$ for online monitoring of two doctors' performance with respect to two characteristic

		ICC		*OCC A*			*OCC B*			*OCC C*		
	k	*ARL*	*SDRL*	*ARL*	*SDRL*	$PI\%$	*ARL*	*SDRL*	$PI\%$	*ARL*	*SDRL*	$PI\%$
Case A	4	174.092	144.232	36.765	24.103	78.88%	20.001	11.204	88.51%	13.676	6.845	92.14%
	5	247.982	204.563	44.224	26.986	82.17%	23.872	12.301	90.37%	16.311	7.492	93.42%
	6	334.966	275.580	51.525	29.413	84.62%	27.727	13.281	91.72%	18.944	8.082	94.34%
	7	435.038	357.285	58.742	31.553	86.50%	31.576	14.186	92.74%	21.576	8.631	95.04%
	8	548.200	449.679	65.918	33.500	87.98%	35.423	15.032	93.54%	24.208	9.147	95.58%
	9	674.450	552.760	73.077	35.312	89.17%	39.269	16.595	94.18%	26.839	9.635	96.02%
	10	813.789	666.528	80.226	37.021	90.14%	43.115	16.595	94.70%	29.471	10.099	96.38%
	λ_2	0.000		0.261			0.892			1.945		
	p_{00}	0.950		0.850			0.750			0.650		
Case B	7	107.086	87.475	50.537	35.642	52.81%	27.319	16.504	74.49%	18.386	9.450	82.83%
	10	200.384	163.653	72.871	46.839	63.63%	37.531	19.771	81.27%	25.071	11.086	87.49%
	14	369.943	302.096	100.781	57.362	72.76%	50.910	23.140	86.24%	33.962	12.914	90.82%
	15	420.397	343.291	107.584	59.517	74.41%	54.246	23.893	87.10%	36.184	13.331	91.39%
	20	721.055	588.777	141.200	68.780	80.42%	70.915	27.334	90.17%	47.295	15.247	93.44%
	λ_2	0.000		0.261			0.892			1.945		
	p_{00}	0.750		0.650			0.600			0.500		
Case C	4	76.147	67.676	40.830	31.205	46.38%	25.433	17.308	66.60%	19.188	12.062	74.80%
	5	81.148	76.150	46.862	35.058	42.25%	29.464	19.268	63.69%	22.356	13.398	72.45%
	6	89.543	82.467	52.168	38.489	41.74%	33.304	21.105	62.81%	25.429	14.678	71.60%
	7	93.145	87.148	56.873	41.699	38.94%	36.898	22.942	60.39%	28.361	15.976	69.55%
	8	95.540	90.602	61.089	44.767	36.06%	40.307	24.774	57.81%	31.184	17.283	67.36%
	9	97.129	93.140	64.881	47.734	33.20%	43.527	26.620	55.19%	33.894	18.612	65.10%
	10	98.182	94.996	68.300	50.610	30.44%	46.575	28.475	52.56%	36.498	19.960	62.83%
	λ	0.000		0.504			1.008			1.317		
	p_{00}	0.950		0.900			0.850			0.800		

Table 6 represent increasing probability of unsuccessful operations and constant probability of appearance of a severe complication. The *OCC*s are ordered according to the value λ_2.

Table 7 shows the *ARL*, the *SDRL*, and the $PI\%$ for the three *OCC*s for k from 4 to 10. Consider *OCC B* which claims that the probability of successful operation by the surgeons has decreased from 95% to 85%. For the case of $k = 5$, the ARL_0 equals to 81.148 while the ARL_1 equals to 29.464, which corresponds to a 63.69% decrease. Contrary to *Cases A* and *B*, we observe that as k increases, the $PI\%$ decreases. However, as λ_2 increases, the $PI\%$ increases too.

4.2 Assessment of Two Doctors' Competence

The *Case D* given in Table 6 concerns the case we want to assess two surgeons' competence with two characteristics. The two surgeons conduct successful operations with probability $\pi_{00} = 0.300$ while both surgeons are unsuccessful with $\pi_{11} = 0.500$. Surgeon A is

successful and surgeon B is unsuccessful with probability $\pi_{01} = 0.100$ while surgeon A is unsuccessful and surgeon B is successful with probability $\pi_{10} = 0.100$. Table 6 also shows 3 *OCC*s which represent increasing probability of unsuccessful operations. The *OCC*s are ordered according to the value λ_2.

As we have already discuss, this case resembles a comparative acceptance sampling scheme. The results for several ks from 10 to 40 are shown in Table 8. Assume the *OCC B* which corresponds to an increase in the probability of unsuccessful operation from 50% to 75%. For $k = 21$, we will identify that the two surgeons have different competence before the 64th pair of operations with power 91.3%. From Table 8, we observe that as k increases, the power $\gamma = P(D_n \leq c|k, \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ increases, while as λ_2 increases, γ increases.

5 Illustrative Examples

Here we will present two illustrative examples concerning the online monitoring of two surgeons' performance with respect to one and two characteristics, respectively. Assume the *Case A* of Table 4. If we set by design the ARL_0 to be almost 370.4, we choose $k = 5$ which gives ARL_0 equal to 378.947 (see Table 5). Thus, we have to set the upper and lower control limit of the control chart to 5 and -5, respectively.

From Fig. 1 we observe that until the 19th day the two surgeons had similar performance (the line regresses around the central line). After the 20th day we observe that the cumulative

Table 8 The power of the comparison of the competence of two doctors with respect to two characteristics

		OCC A	*OCC B*	*OCC C*
k	$c_{0.05}$		power γ	
10	16	0.153	0.296	0.647
11	19	0.180	0.356	0.744
12	22	0.203	0.408	0.816
13	26	0.249	0.494	0.896
14	30	0.290	0.568	0.943
15	34	0.327	0.631	0.969
16	38	0.361	0.684	0.983
17	43	0.412	0.751	0.993
18	48	0.459	0.804	0.997
19	53	0.501	0.846	0.999
20	59	0.556	0.890	1.000
21	64	0.589	0.913	1.000
22	70	0.633	0.938	1.000
23	77	0.685	0.960	1.000
24	83	0.719	0.971	1.000
25	90	0.758	0.981	1.000
30	127	0.895	0.998	1.000
35	172	0.991	1.000	1.000
40	222	0.992	1.000	1.000

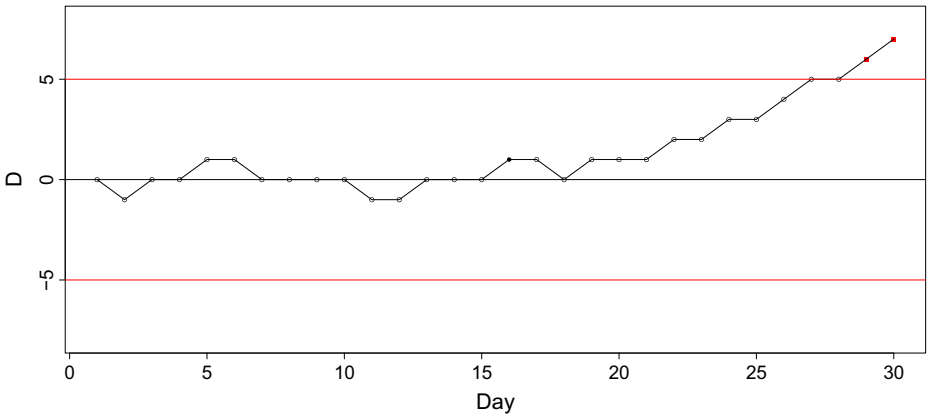


Fig. 1 The control chart for online monitoring of two surgeons' performance with respect to one characteristic

difference D_n increased, which means that the performance of surgeon B worsened. Finally, at day 29, D_n went above the upper control limit. This means that we must take action in order to identify what went wrong with the performance of surgeon B.

Assume now that a death occurred by assumed fault of surgeon B at day 16, depicted as a solid cycle in Fig. 1. In this case we ought to stop the process declaring that surgeon A is more efficient.

The second example concerns the online monitoring of two surgeons' performance with respect to two characteristics. Consider the *Case C* of Table 6. Setting by design the ARL_0 to almost 90.0, the appropriate choice of k is 6 which gives ARL_0 equal to 89.543 (see Table 7). This, drives us to set the upper control limit of the control chart to 6 and the lower control limit to -6. The Fig. 2 shows that from day 9 the D_n increased until day 15 where D_n surpassed the upper control limit. Appropriate corrective actions were taken and two days later the D_n started to decrease and eventually was stabilized after the 23rd day.

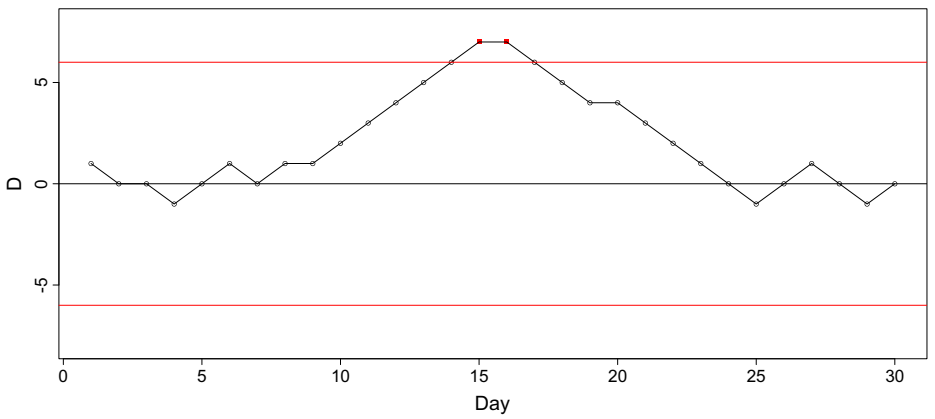


Fig. 2 The control chart for online monitoring of two surgeons' performance with respect to two characteristic

6 Extended Schemes

6.1 Detection of Deterioration of the Performance of Both Doctors

The proposed method compares two doctors assuming that one of them has a stable high performance. In this subsection we outline a modification of the method in order to detect deterioration of the performance of both doctors.

To this end, along with the monitoring of the D_n, we can monitor the performance of one of the doctors, by depicting in the same chart the sequence X_n of his surgery outcomes, i.e. success (0) or failure (1). Thus, on the chart we will depict a sequence of the form of 0001101110111. If we identify a run of length, say r, of failures then we stop the comparison processes declaring that the doctor's performance has worsened, or both doctors' performance has worsened if simultaneously we have small value of D_n. The use of runs rules in the statistical process control has been discussed by Koutras et al. (2007).

Denoting by $ARL^{(1)}$ the in-control ARL of our method and by $ARL^{(2)}$ the in-control ARL of the individual doctor's chart, the total in-control ARL ($ARL^{(t)}$) is approximated (see Koutras et al. (2006)) by

$$\frac{1}{ARL^{(t)}} \cong \frac{1}{ARL^{(1)}} + \frac{1}{ARL^{(2)}}.$$

For the computation of $ARL^{(2)}$ we exploit a generalized geometric distribution (Koutras et al. 2007).

Assuming that we monitor doctor A, Table 9 shows the $ARL^{(1)}$, $ARL^{(2)}$, and the approximated and simulated $ARL^{(t)}$ for online monitoring the two doctors' performance with respect to one characteristic for $k = 4, 5, 6$ and runs of length $r = 2, 3$. If we consider the *ICC* of *Case A* with $k = 5$ and $r = 2$, we have $ARL^{(1)} = 378.947$, $ARL^{(2)} = 420.000$,

Table 9 The $ARL^{(1)}$, $ARL^{(2)}$, and the approximate and simulated $ARL^{(t)}$ for online monitoring of two doctors' performance with respect to one characteristic for $k = 4 - 10$ and $r = 2, 3$ under *Case A*

		ICC				*OCC A*			
k	r	$ARL^{(1)}$	$ARL^{(2)}$	*Approx* $ARL^{(t)}$	*Simul* $ARL^{(t)}$	$ARL^{(1)}$	$ARL^{(2)}$	*Approx* $ARL^{(t)}$	*Simul* $ARL^{(t)}$
4	2	263.158	420.000	161.787	179.896	95.342	420.000	77.703	67.566
	3	263.158	1706.140	227.992	259.679	95.342	1706.140	90.296	93.273
5	2	378.947	420.000	199.209	222.058	117.319	420.000	91.703	76.039
	3	378.947	1706.140	310.077	370.536	117.319	1706.140	109.771	113.740
6	2	515.789	420.000	231.496	258.457	138.510	420.000	104.160	82.511
	3	515.789	1706.140	396.056	497.893	138.510	1706.140	128.110	132.924
		OCC B				*OCC C*			
		$ARL^{(1)}$	$ARL^{(2)}$	*Approx* $ARL^{(t)}$	*Simul* $ARL^{(t)}$	$ARL^{(1)}$	$ARL^{(2)}$	*Approx* $ARL^{(t)}$	*Simul* $ARL^{(t)}$
4	2	33.306	420.000	30.859	22.911	24.995	420.000	23.591	16.471
	3	33.306	1706.140	32.668	31.931	24.995	1706.140	24.634	23.675
5	2	39.993	420.000	36.516	24.748	29.999	420.000	27.999	17.560
	3	39.993	1706.140	39.077	37.594	29.999	1706.140	29.481	27.549
6	2	46.665	420.000	41.999	26.106	34.999	420.000	32.307	18.314
	3	46.665	1706.140	45.423	42.970	34.999	1706.140	34.295	31.253

approximate $ARL^{(t)} = 199.209$, and simulated $ARL^{(t)} = 222.058$. For the *OCC A* we have $ARL^{(1)} = 117.319$, $ARL^{(2)} = 420.000$ (doctor A remains stable), approximate $ARL^{(t)} = 91.703$, and simulated $ARL^{(t)} = 76.039$.

Figure 3 shows three examples of the modified version. In all cases, D_n is depicted. With a red square we have visualized the time point that the cumulative difference exceeds the upper or lower control limit, while with a red star we have visualized the time points that doctor A has a failure. Specifically, in Fig. 3a although doctor A has a stable performance (we see a few non-continuous red stars), the D_n went under the lower control limit at day 30. This means that the D_n decreased and eventually went below the lower control limit due to the many failures of doctor A. Thus, doctor B is more efficient than doctor A. In Fig. 3b the D_n remains between the control limits, which means that both doctors are equivalent. However, at day 25 a run of failures of length 3 appeared for doctor A. This means that we have to stop the monitoring procedure declaring that doctor B is more efficient. Finally, in Fig. 3c both "rules" give an out-of-control signal. For the first time, the D_n went below the lower control limit at day 24, which means that doctor A is responsible for the out-of-control signal. However, 3 days later a run of failures of length 3 appeared for doctor A. This means that we have to stop the monitoring procedure.

6.2 A Run Sum Approach

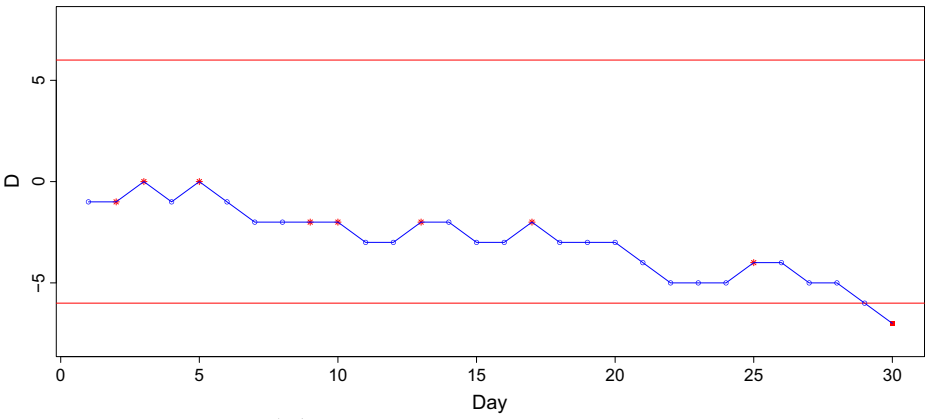
In this subsection we modify our method following the rational of the run sum control chart (Montgomery 2013). As we will show, this extension appears to be better if we aim at large values of in-control *ARL*.

In this case we assume that the D_n run sum type statistic equals to $D_{n-1} + 1$ if the difference $X^B - X^A$ is positive (i.e. we have a failure for doctor B) for $D_{n-1} > 0$ while it equals to 0 if the $D_{n-1} < 0$. Analogously, D_n equals to $D_{n-1} - 1$ if the difference $X^B - X^A$ is negative (i.e. we have a failure for doctor A) for $D_{n-1} < 0$ while it equals to 0 if the $D_{n-1} > 0$. Thus, the D_n run sum type statistic is

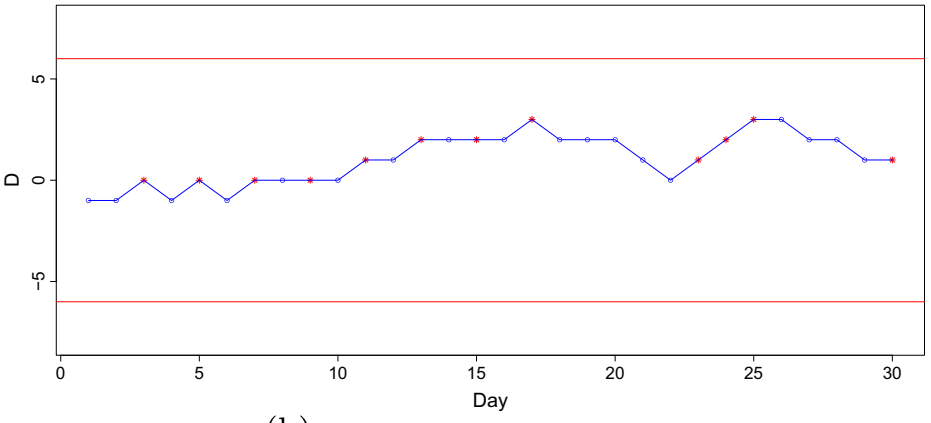
$$D_n = \begin{cases} D_{n-1}, & \text{if } X^B - X^A = 0 \\ D_{n-1} + 1, & \text{if } X^B - X^A > 0 \text{ and } D_{n-1} > 0 \\ 0, & \text{if } X^B - X^A > 0 \text{ and } D_{n-1} < 0 \\ D_{n-1} - 1, & \text{if } X^B - X^A < 0 \text{ and } D_{n-1} < 0 \\ 0, & \text{if } X^B - X^A < 0 \text{ and } D_{n-1} > 0, \end{cases}$$

The transition probability matrix for the case of one characteristic is given in Appendix A.3.

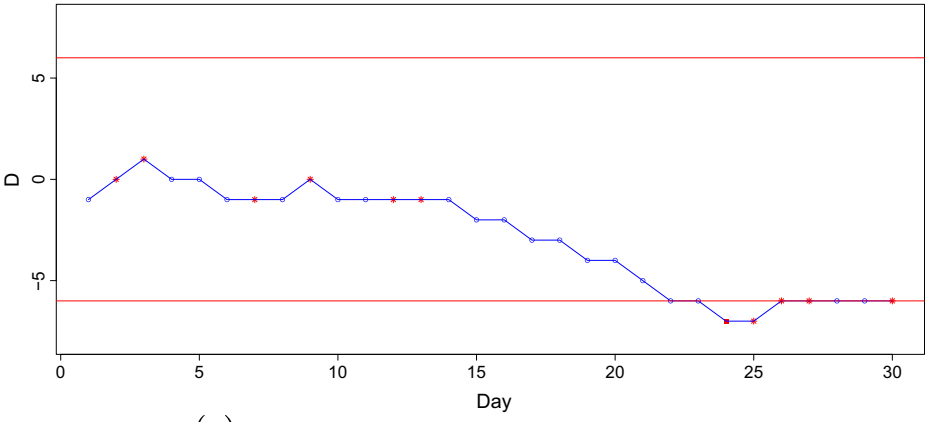
Table 10 presents the *ARL*, the *SDRL*, and the $PI\%$ for the *Case A* of Table 4 for k from 3 to 10. We observe that in both cases the run sum modification achieves very high ARL improvement. Let us now compare these results to those of Table 5. In order the comparison to be valid, we should compare the $PI\%$ for similar in-control *ARL*s. For example, see the results for $k = 4$ of Table 5 with in-control *ARL* equal to 263.158 and the results for $k = 3$ of Table 10 with in-control *ARL* equal to 231.579. For the *OCC A*, the proposed method gives $PI\% = 63.71\%$ while for the run sum extension we have less $PI\%$ (equal to 63.71%). For the *OCC B*, we have $PI\% = 87.34\%$ for the proposed method and $PI\% = 86.80\%$ for the extended one. For the *OCC C*, we have $PI\% = 90.50\%$ for the proposed method and $PI\% = 90.38\%$ for the extended one. Similar results we have for $k = 5$ of Table 5 with in-control $ARL = 515.789$ and for $k = 3$ of Table 10 with in-control $ARL = 484.211$.



(a) D_n gives an out-of-control signal



(b) X_n gives an out-of-control signal



(c) Both D_n and X_n give an out-of-control signal

Fig. 3 The modified control chart

Table 10 The *ARL*, *SDRL*, and the $PI\%$ under the run sum modification for online monitoring of two doctors' performance with respect to one characteristic under *Case A*

	ICC		*OCC A*			*OCC B*			*OCC C*		
k	*ARL*	*SDRL*	*ARL*	*SDRL*	$PI\%$	*ARL*	*SDRL*	$PI\%$	*ARL*	*SDRL*	$PI\%$
3	231.579	208.918	93.318	77.887	59.70%	30.578	21.014	86.80%	22.283	14.156	90.38%
4	484.211	452.464	151.880	130.560	68.63%	42.344	29.447	91.25%	30.028	19.147	93.80%
5	989.474	948.249	237.357	209.903	76.01%	56.541	40.305	94.29%	38.983	25.357	96.06%
6	2000.000	1948.990	362.299	328.775	81.89%	73.712	54.125	96.31%	49.349	32.981	97.53%
7	4021.050	3960.020	545.970	506.149	86.42%	94.495	71.533	97.65%	61.352	42.236	98.47%
8	8063.160	7991.930	816.199	769.959	89.88%	119.652	93.283	98.52%	75.249	53.374	99.07%
9	16147.400	16065.800	1214.110	1161.340	92.48%	150.105	120.291	99.07%	91.341	66.689	99.43%
10	32315.700	32223.100	1800.280	1740.880	94.43%	186.969	153.666	99.42%	109.974	82.522	99.66%

The only difference is that now the extended method gives a slight larger $PI\%$ for the *OCC B*. Finally, let us consider two cases with larger in-control *ARL* ($k = 10$ of Table 5 with in-control $ARL = 1052.630$ and $k = 5$ of Table 10 with in-control $ARL = 989.474$). In this case, the extended method achieves higher percentage *ARL* improvement in all cases apart from *OCC A*. Concluding, we can say that the proposed method is always better for small shifts in the doctor's performance while the extended method with the run sum approach is better for larger in-control *ARL*s and moderate and large shifts in the doctor's performance.

6.3 A Risk-Adjusted Procedure

Up until this stage we have assumed that the patients are matched with very similar characteristics over time when comparing surgeons, but this may only be possible for very busy surgeons with nearly identical patients over time. Thus, we now outline a process for comparing two surgeons' performance when their patient case mix differs locally in time but overlaps in the long-term, e.g. several years.

Phase I monitoring for the risk adjustment approach uses training data involving recent past surgery performance to fit a logistic regression model on whether the surgery of two surgeons was a success or not as a response. The explanatory variables in the logistic regression are the patient risk factors such as age, gender, patient co-morbidities, patient social disadvantage index, etc. An indicator variable for the two surgeons is included as a factor in the model using the combined data from both surgeons to fit this model. If the surgeon indicator variable is not significant as an explanatory variable and it failures to interact with any of the risk factors or with time then the two surgeons performances do not differ significantly. In addition the time variable is not significant and time fails to interact with the surgeon indicator variable then these two surgeons are not significantly different from each other and are in-control (retrospectively speaking). This model will have adjusted for the differing case mix of the patients used across time and therefore is a fair comparison of surgery performance. If the surgeons use different hospitals to perform their surgery and these do not overlap in some way then the comparison may still be unfair because hospital differences is difficult to be adjusted for. However, as this is beyond the scope of this paper, we will assume that both surgeons use the same hospitals. If we assume that there are

no differences between the two surgeons, the logistic regression model fitted in the Phase I analysis provides practitioners a risk adjustment tool for monitoring differences between two surgeons prospectively. It is worth noting that by updating the fitted logistic regression model, monitoring changes in its parameters is useful for monitoring only linear changes in the logit response, whereas the control chart approach is appropriate for both linear and non-linear changes and therefore is more appropriate.

The adjusted risk of the surgeon g's probability of failure for patient with risk factors $\mathbf{x}$ is $\pi^g_{1,s} = \theta^g_{\mathbf{x}}\rho^g_s$ where $\theta^g_{\mathbf{x}}$ is the corresponding risk adjustment to the failure for an average patient with risk factors $\mathbf{x}$ and ρ^g_s is the baseline risk of the doctor. We want to use the same risk adjustment for both surgeons $g = A$ and B. If do this using the same logistic regression model for both surgeons using the following model:

$$logit(\pi^g_{1,s}) = \mathbf{x}'_s\beta + \alpha\delta^g + \mathbf{x}'_s\gamma^g$$

where β is the risk adjustment parameters, e.g., how much do we adjust for increases in age, $\delta^g = 1$ if operated on by surgeon g otherwise it is zero, γ^g is the influence the surgeon g have on the adjustment risk factors, e.g., the surgeon performance is not good for older patients, α is the additive influence of the first surgeon, and γ^g is the influence of the g-th surgeon on the covariates $\mathbf{x}$ (e.g. where surgeon A performs worse for older patients than surgeon B). If the risk factors and surgeons interact then the surgeons do differ in performance. This approach is very similar to Phase I analysis outlined earlier but is only useful if you wish to monitoring performance monthly or quarterly and each surgeon performs enough surgeries to warrant this approach. If vector γ^g is equal to the zero vector and $\delta^g = 0$ then we are interested in monitoring whether these two surgeons performance remains equal over time using a prospective control chart. We still need to adjust for case mix to remove the variation in well-being of the patient during surgery to give our approach a better chance of detecting drifts in surgery performance. This model could be refitted using monthly or quarterly data to assess changes in performance over time.

However it is important to realize that this model is only good at monitoring changes that are linear on the logit scale, and it is not good at monitoring changes that are non-linear on the logit scale. For example, if the surgeon B performs temporal changes for a week in a quarterly evaluation of performances while he is on medication and this is not recorded, then the logistic regression is unlikely to find this because this will occur in the residuals of the model but the control chart will have a good chance of finding it if it is a big effect.

Then, take the probabilities from the above logistic regression model, fitted using in-control data (often in Phase I), and plug them in Eq. 3 in order to compute the exact distribution of L_{D_n}, after adjusting for the characteristics of each patient.

7 Conclusions

The aim of the paper was to present a procedure for monitoring multi-attribute comparative health processes. More specifically, we presented two cases: the comparison of doctors' performance and the assessment of doctors' competence. The latter can be treated as an acceptance sampling. The main idea in both cases is to use the difference between two quantities: the difference between the number of "failures" for doctor A and the number of "failures" for doctor B.

To compare the doctors, we consider the one of them as reference (a doctor with stable high performance). The reference doctor may be the most experienced one, the one with the most years of service, etc. In such a way, we assume that the good performance of the reference doctor is assured.

The proposed method can also be modified to handle the comparison of more than two doctors (say ℓ). For example, we can use as test statistic the cumulative sum of the maximum difference between the number of "failures" for doctor i and the number of "failures" for doctor j, for $i \neq j, i, j = 1, 2, \ldots, \ell$.

A third modification exploiting the run sum approach was also presented. In this case the test statistic goes to 0 whenever the difference $X^B - X^A$ is positive and simultaneously D_{n-1} is negative or whenever $X^B - X^A$ is negative and simultaneously D_{n-1} is positive. Otherwise, it increases or decreases by one unit.

The exact run length distribution is derived, using the Markov Chain embedding technique. In the proposed procedure we do not need to know the direction of the shift.

For the above mentioned procedures we have assumed that the patients are matched. This is more plausible when patients have similar characteristics. For the case where we cannot make the above assumption we should adjust the probability of successful operation for several characteristics of the patients. For this reason we briefly discussed a risk-adjusted method.

A Appendix

A.1 The Transition Probability Matrix for the Process with One Characteristic

In the case where we assess only one characteristic, the transition probability matrix is

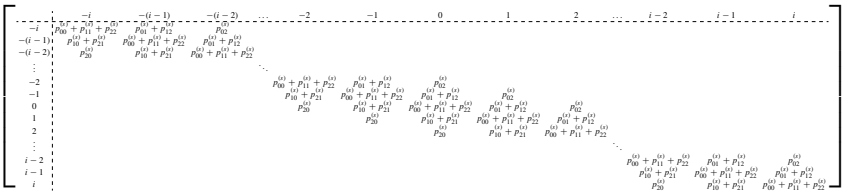
$$\boldsymbol{\Lambda}_{1,1}^{(s)} = \begin{array}{c|cccccccccccccccccc} & -i & -(i-1) & -(i-2) & \cdots & -2 & -1 & 0 & 1 & 2 & \cdots & i-2 & i-1 & i \\ \hline -i & p_{00}^{(s)}+p_{11}^{(s)} & p_{01}^{(s)} & & & & & & & & & & & \\ -(i-1) & p_{10}^{(s)} & p_{00}^{(s)}+p_{11}^{(s)} & p_{01}^{(s)} & & & & & & & & & & \\ -(i-2) & & p_{10}^{(s)} & p_{00}^{(s)}+p_{11}^{(s)} & \ddots & & & & & & & & & \\ \vdots & & & & \ddots & p_{00}^{(s)}+p_{11}^{(s)} & p_{01}^{(s)} & & & & & & & \\ -2 & & & & & p_{10}^{(s)} & p_{00}^{(s)}+p_{11}^{(s)} & p_{01}^{(s)} & & & & & & \\ -1 & & & & & & p_{10}^{(s)} & p_{00}^{(s)}+p_{11}^{(s)} & p_{01}^{(s)} & & & & & \\ 0 & & & & & & & p_{10}^{(s)} & p_{00}^{(s)}+p_{11}^{(s)} & p_{01}^{(s)} & & & & \\ 1 & & & & & & & & p_{10}^{(s)} & p_{00}^{(s)}+p_{11}^{(s)} & p_{01}^{(s)} & & & \\ 2 & & & & & & & & & p_{10}^{(s)} & p_{00}^{(s)}+p_{11}^{(s)} & & & \\ \vdots & & & & & & & & & & \ddots & & & \\ i-2 & & & & & & & & & & & p_{00}^{(s)}+p_{11}^{(s)} & p_{01}^{(s)} & \\ i-1 & & & & & & & & & & & p_{10}^{(s)} & p_{00}^{(s)}+p_{11}^{(s)} & p_{01}^{(s)} \\ i & & & & & & & & & & & & p_{10}^{(s)} & p_{00}^{(s)}+p_{11}^{(s)} \end{array}$$

Matrix $\boldsymbol{\Lambda}_{1,1}^{(s)}$ includes the probabilities of the transitions of the Markov chain from state to state. More specifically, the chain moves from state i to state i, $i = -k, \ldots, k$ with probability $p_{00}^{(s)} + p_{11}^{(s)}$; from state i moves to state $i+1, i = -k, \ldots, (k-1)$, with probability $p_{01}^{(s)}$; and from state i moves to state $i-1$, $i = -(k-1), \ldots, k$, with probability $p_{10}^{(s)}$. For any other case, the chain moves to the absorbing state. The empty cells of $\boldsymbol{\Lambda}_{1,1}^{(s)}$ are filled by zero. Thus, the (row) vector of initial probabilities of the Markov chain is $\boldsymbol{\pi}_0' = (0, 0, \ldots, 1, \ldots, 0)_{1\times d}$.

A.2 The Transition Probability Matrix for the Process with Two Characteristics

In the case where we assess two characteristics, the transition probability matrix is

$$\boldsymbol{\Lambda}_{1,2}^{(s)}=$$



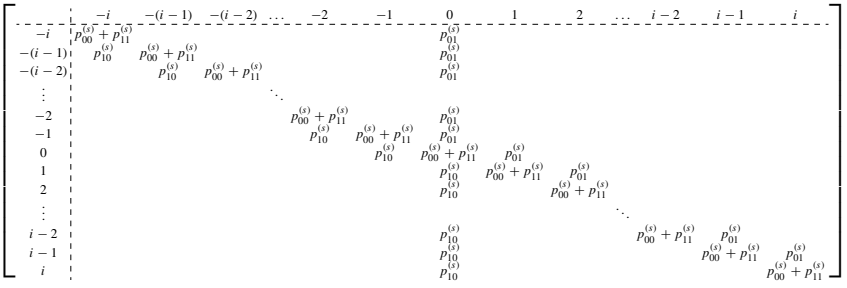
Matrix $\boldsymbol{\Lambda}_{1,2}^{(s)}$ includes the probabilities of the transitions of the Markov chain from state to state. More specifically, the chain moves from state i to state i, $i=-k,\ldots,k$ with probability $p_{00}^{(s)}+p_{11}^{(s)}+p_{22}^{(s)}$; from state i moves to state $i+1$, $i=-k,\ldots,(k-1)$, with probability $p_{01}^{(s)}+p_{12}^{(s)}$; from state i moves to state $i-1$, $i=-(k-1),\ldots,k$, with probability $p_{10}^{(s)}+p_{21}^{(s)}$; from state i moves to state $i+2$, $i=-k,\ldots,(k-2)$, with probability $p_{02}^{(s)}$; from state i moves to state $i-2$, $i=-(k-2),\ldots,k$, with probability $p_{20}^{(s)}$. For any other case, the chain moves to the absorbing state. The empty cells of $\boldsymbol{\Lambda}_{1,2}^{(s)}$ are filled by zero. Thus, the (row) vector of initial probabilities of the Markov chain is $\boldsymbol{\pi}_0'=(0,0,\ldots,1,\ldots,0)_{1\times d}$.

In a similar way we can construct the transition probability matrix for the case of more than two characteristics.

A.3 The Transition Probability Matrix for the Run Sum Approach

In the case of the run sum modification, the transition probability matrix is

$$\boldsymbol{\Lambda}_{1,2}^{(s)}=$$



Matrix $\boldsymbol{\Lambda}_{1,1}^{(s)}$ includes the probabilities of the transitions of the Markov chain from state to state. More specifically, the chain moves from state i to state i, $i=-k,\ldots,k$ with probability $p_{00}^{(s)}+p_{11}^{(s)}$; from state i moves to state $i+1$, $i=1,\ldots,(k-1)$, (i.e., if the difference is positive) or to 0 if the difference is negative, with probability $p_{01}^{(s)}$; and from state i moves to state $i-1$, $i=-(k-1),\ldots,-1$, (i.e., if the difference is negative) or to 0 if the difference is positive, with probability $p_{10}^{(s)}$. For any other case, the chain moves to the absorbing state. The empty cells of $\boldsymbol{\Lambda}_{1,1}^{(s)}$ are filled by zero. Thus, the (row) vector of initial probabilities of the Markov chain is $\boldsymbol{\pi}_0'=(0,0,\ldots,1,\ldots,0)_{1\times d}$.

References

Antzoulakos D, Rakitzis A (2008) The modified r out of m control chart. Commun Stat Simul Comput 37(2):396–408

Balakrishnan N, Bersimis S, Koutras MV (2009) Run and frequency quota rules in process monitoring and acceptance sampling. J Qual Technol 41(1):66–81

Bersimis S, Sachlas A, Papaioannou T (2015) Flexible designs for phase ii comparative clinical trials involving two response variables. Stat Med 34(2):197–214

Burkom HS, Elbert Y, Feldman A, Lin J (2004) Role of data aggregation in biosurveillance detection strategies with applications from essence. Morb Mortal Wkly Rep 53(Suppl):67–73

Dubrawski A (2011) The role of data aggregation in public health and food safety surveillance. In: Kass-Hout T, Zhang X (eds) Biosurveillance - Methods and Case Studies. FL: CRC/Taylor & Francis Group, Boca Raton

Grigg O, Farewell V (2004) An overview of risk-adjusted charts. J R Stat Soc Ser A 167:523–539

Koutras MV, Bersimis S, Antzoulakos DL (2006) Improving the performance of the chi-square control chart via runs rules. Methodol Comput Appl Probab 8(3):409–426

Koutras MV, Bersimis S, Maravelakis PE (2007) Statistical process control using shewhart control charts with supplementary runs rules. Methodol Comput Appl Probab 9(2):207–224

Lim TO, Soraya A, Ding LM, Morad Z (2002) Assessing doctors' competence: application of cusum technique in monitoring doctors' performance. Int J Qual Health Care 14(3):251–258

Maruthappu M, Carty MJ, Lipsitz SR, Wright J, Orgill D, Duclos A (2014) Patient- and surgeon-adjusted control charts for monitoring performance. BMJ Open 4:e004046. doi:10.1136/bmjopen-2013-004046

Montgomery DC (2013) Introduction to Statistical Quality Control, 7th edn. Wiley, New Jersey

Rethans JJ, Van Leeuwen Y, Drop R, Vand Der Vleuten C, Sturmans F (1990) Competence and performance: two different concepts in the assessment of quality of medical care. Fam Pract 7(3):168–174

Reynolds MR, Stoumbos ZG (2000) A general approach to modeling cusum charts for a proportion. IIE Trans 32:515–535

Reynolds MR, Stoumbos ZG (2004) Should observations be grouped for elective process monitoring? J Qual Technol 36(4):343–366

Ryan AG, Wells LJ, Woodall WH (2011) Methods for monitoring multiple proportions when inspecting continuously. J Qual Technol 43:237–248

Ryan SA, Thompson CB (2002) The use of aggregate data for measuring practice improvement. Semin Nurse Manag 10(2):90–94

Schuh A, Woodall WH, Camelio JA (2013) The effect of aggregating data when monitoring a poisson process. J Qual Technol 45(3):260–272

Sego LH, Reynolds MR Jr, Woodall WH (2009) Risk-adjusted monitoring of survival times. Stat Med 28:1386–1401

Steiner SH, Cook RJ, Farewell VT, Treasure T (2000) Monitoring surgical performance using risk-adjusted cumulative sum charts. Biostatistics 1:441–452

Topalidou E, Psarakis S (2009) Review of multinomial and multi-attribute quality control charts. Qual Reliab Eng Int 25(7):773–804

Tsui KL, Chiu W, Gierlich P, Goldsman D, Liu X, Maschek T (2008) A review of healthcare, public health, and syndromic surveillance. Qual Eng 20:435–450

Woodall WH (1997) Control charting based on attribute data: Bibliography and review. J Qual Technol 29(2):172–183

Woodall WH, Montgomery DC (2004) Some current directions in the theory and application of statistical process monitoring. J Qual Technol 46(1):78–93

Zhang M, Megahed FM, Woodall WH (2014) Exponential cusum charts with estimated control limits. Qual Reliab Eng 30:275–286

Zhang X, Woodall WH (2015) Dynamic probability control limits for risk-adjusted bernoulli cusum charts, Stat Med. doi:10.1002/sim.6547