Inspecting the quality of care: a comparison of CUSUM methods for inter hospital performance

Gomon, D.; Sijmons, J.; Putter, H.; Dekker, J.W.; Tollenaar, R.; Wouters, M.; ... ; Signorelli, M.

Health Services and Outcomes Research Methodology (2024) 24:281–303
https://doi.org/10.1007/s10742-023-00315-0

Inspecting the quality of care: a comparison of CUSUM methods for inter hospital performance

Daniel Gomon[1] · Julie Sijmons[2,3] · Hein Putter[1,4] · Jan Willem Dekker[5] · Rob Tollenaar[2,7] · Michel Wouters[6,7] · Pieter Tanis[3,8] · Marta Fiocco[1,4] · Mirko Signorelli[1]

Received: 21 June 2023 / Accepted: 30 September 2023 / Published online: 24 October 2023

Abstract

During the past 14 years, a clinical audit has been used in the Netherlands to provide hospitals with data on their performance in colorectal cancer care. Continuous feedback on the quality of care provided at each hospital is essential to improve patient outcomes. It is unclear which methods should be used to generate most informative output for the identification of potential quality issues. Our aim is to compare the commonly employed funnel plot with existing cumulative sum (CUSUM) methodology for the evaluation of postoperative survival and hospital stay outcomes of patients who underwent colorectal surgery in the Netherlands. Data from the Dutch ColoRectal Audit on 25367 patients in the Netherlands who underwent surgical resection for colorectal cancer in 71 hospitals between 2019 and 2021 is used to compare four methods for the detection of deviations in the quality of care. Two methods based on binary outcomes (funnel plot, binary CUSUM) and two CUSUM charts based on survival outcomes (BK-CUSUM and CGR-CUSUM) are considered. A novel approach for determining hospital specific control limits for CUSUM charts is proposed. The ability to detect deviations as well as the time until detection are compared for the four methods. Charts were constructed for the inspection of both postoperative survival and hospital stay. Methods using survival outcomes always yielded faster detection times compared to approaches employing binary outcomes. Detections between methods mostly coincided for postoperative survival. For hospital stay detections varied strongly, with methods based on survival outcomes signalling over half the hospitals. Further pros and cons as well as pitfalls of all methods under consideration are discussed. Methodology for the continuous inspection of the quality of care should be tailored to the specific outcome. Properly understanding how the mechanism of a control chart functions is crucial for the correct interpretation of results. This is particularly true for CUSUM charts, which require the choice of a parameter that greatly influences the results. When applying CUSUM charts, consideration of these issues is strongly recommended.

Keywords Quality control · Survival analysis · Cumulative sum (CUSUM)

Abbreviations

CUSUM	CUmulative SUM

Extended author information available on the last page of the article

CGR-CUSUM	Continuous time Generalised Rapid response CUSUM (Gomon et al. 2022)
BK-CUSUM	Biswas & Kalbfleisch CUSUM (Biswas and Kalbfleisch 2008)
DCRA	Dutch ColoRectal Audit
OR	Odds Ratio

1 Background

Continuous evaluation of the quality of care in healthcare institutions is vital, and allows for the detection and swift resolution of deviations in performance. Almost a century ago, industrial processes were faced with a similar problem resulting in the creation of the field of statistical process control. Consequently, many of the methods (called *control charts*) developed for industrial production lines found their way into health care inspection. It quickly became apparent that the inspection of patient outcomes is more complicated due to the presence of underlying risk factors, called *covariates/prognostic factors*. This resulted in the development of *risk-adjusted* control charts, allowing for covariates to be incorporated into the model. One drawback of using control charts for clinical outcomes is that they are primarily constructed for discrete, mostly binary outcomes such as failure/success or dead/alive. In the rest of this article such methods will be called *binary control charts*. However, patient outcomes are often considered in continuous time, where one is interested in whether a patient is still alive or has experienced an event during follow-up. Recently *continuous time* extensions of control charts have been developed, allowing to incorporate survival outcomes into *survival control charts*.

The current study focuses on the scenario within the context of the continuous evaluation of the quality of care at a single hospital, where this hospital is initially performing well (*in-control*) and might perform sub-optimally (*out-of-control*) later on. When constructing control charts for a hospital over a study period, extreme values of the chart support the hypothesis that the quality at this hospital has deteriorated. To determine when the quality is no longer acceptable, a *control limit* is used. When the value of the chart exceeds the control limit, a signal is produced indicating a possible deterioration in performance at the specific hospital being evaluated. Control limits are usually determined by guaranteeing that either hospitals performing on target will not be detected for a long time, or by limiting the number of false alarms over the study period (similar to the confidence levels in the funnel plot). The value of the control limit should be hospital specific, as hospitals vary in the number of patients treated per time unit (Gomon et al. 2022). Due to a lack of algebraic results for risk-adjusted control charts, control limits are usually determined using a simulation study. This significantly complicates the use of control charts in medical applications, as the variability introduced by the simulation procedure may favour one hospital over another. We propose an approach to determine control limits which attempts to eliminate the unfairness resulting from the variability in the simulation study.

This article focuses on *cumulative sum* (CUSUM) control charts, originally introduced by Page (Page 1954) and later applied for assessing healthcare quality (Steiner et al. 2000). Binary CUSUM charts and their applications in health care have been extensively studied and compared with other binary control charts (Woodall 2006; Tsui et al. 2011, 2008; Mahmoud et al. 2008; Jiang et al. 2012; Grigg et al. 2003; Fatt Gan et al. 2020). CUSUM charts have emerged as one of the most suitable charts for the detection of small fluctuations in the quality of health care and have recently been shown to yield faster detections than other control charts for sustained shifts in quality (Diko et al. 2019). This property is especially useful in the context of audit monitoring, where the primary goal is that of identifying persistent changes in the quality of care. Several extensions of CUSUM charts to survival outcomes have been developed (Biswas and Kalbfleisch 2008; Gomon et al. 2022; Begun et al. 2019; Sego et al. 2009) as well as some continuous time extensions of other binary charts (Grigg 2018; Steiner and Jones 2009). All discussed survival charts have been shown to potentially detect deviations faster compared to their binary counterparts, both on real data sets as well as in simulation studies. Little research has been done however on the possible drawbacks of using survival CUSUMs, as well as the use of control charts for the detection of an increase in the quality of care.

We aim to fill this gap in research, detailing when the use of survival charts may be preferable and when it is not. We also highlight some dangers of using CUSUMs for the inspection of clinical outcomes. A data set from the Dutch ColoRectal Audit (DCRA) about colorectal surgery procedures in the Netherlands is used to compare the performance of the binary CUSUM (Steiner et al. 2000), the Biswas & Kalbfleisch CUSUM (Biswas and Kalbfleisch 2008) and the CGR-CUSUM (Gomon et al. 2022) with the commonly used funnel plot (Spiegelhalter 2005). Although the funnel plot was not developed for the repeated inspection of a process, it is commonly used for this purpose in the medical field [e.g. Griffen et al. (2012); Verburg et al. (2017)], as well as currently by the DCRA. Some of the pitfalls of funnel plots are discussed in Willik et al. (2020) We show that CUSUM control charts are more appropriate for continuously monitoring the quality of care, possibly reducing the number of false detections as compared to the funnel plot while simultaneously providing new insights into the causes of deviations in the quality of care.

We propose the use of CUSUM charts for the inspection of an “inverse” survival problem where longer survival times are problematic and demonstrate its use for the inspection of hospital stay duration. Additionally, we investigate when and whether the use of survival/generalized control charts such as the CGR-CUSUM is appropriate in a medical setting.

This article is organised as follows. In Section “Methods” we give an overview of the four methods and propose as a novel procedure for determining control limits for multiple hospitals. Section “Results” describes the resulting performance of the considered methods on the DCRA data set followed by a Discussion. The article ends with a section describing our main “Conclusions”.

2 Methods

Consider the setting where multiple hospitals are performing the same medical procedure over the duration of a study period. Our goal is to compare four different methods that can be used to inspect the quality of care at different hospitals. These methods are used to distinguish between hospitals providing unsatisfactory care from hospitals that are performing well, which we define as *out-of-control* and *in-control* hospitals respectively. This helps the hospitals to identify where it might be necessary to evaluate the care process with plans for subsequent improvement. As early interventions are crucial, we are primarily interested in the time that each method takes to detect deviating hospitals. We are also interested in evaluating which hospitals are detected by each method. The comparison is performed on a real-life study about surgical resection of colon cancer.

We start with a description of a real-life data set on colon cancer surgery. Afterwards we introduce a binary outcome and give two examples of commonly used (binary) charts in this context: the funnel plot (Spiegelhalter 2005) and Bernoulli Cumulative SUM (CUSUM) chart (Steiner et al. 2000). We then introduce survival outcomes and present two survival charts: the Biswas & Kalbfleisch CUSUM (BK-CUSUM) (Biswas and Kalbfleisch 2008) and the Continuous time Generalised Rapid response CUSUM (CGR-CUSUM) (Gomon et al. 2022). We consider a method to determine control limits for the CUSUM charts and conclude this section with a discussion on an alternative use of quality control charts: the detection of an increase in the quality of care.

2.1 Data

The Dutch ColoRectal Audit (DCRA) is a nationwide clinical audit in the Netherlands that includes all patients that underwent surgery for primary colorectal cancer. The audit monitors, evaluates and improves the colorectal cancer care and has a completeness up to 95% which is externally validated (Van Leersum et al. 2013).

A total of 25367 patients who underwent a surgical resection in the period 01/01/2019 up until 31/12/2021 were included in this study, with 71 hospitals performing surgeries. The following patient, disease and procedural characteristics were extracted from the database for risk-adjustment: sex, body mass index (BMI), age, Charlson Comorbidity Index (CCI), American Society of Anesthesiologist (ASA) score, solitary of synchronous tumor, preoperative tumor complications (e.g. obstruction/ileus, perforation, anemia or peritumoral abscess), T-stage, M-stage, emergency or elective resection, additional resection due to tumor ingrowth or metastasis (Kolfschoten et al. 2011). A summary of these prognostic factors can be found in Table 1. BMI showed the highest percentage of missing data which is only 1.0%, therefore complete case analysis was performed. Outcomes considered were postoperative mortality within 90 days after resection (average 2.4%) and length of hospital stay in days (mean 7.05, SD 8.93, median 4).

2.2 Binary charts

After a surgical procedure, the health of a patient is closely monitored. Due to practical constraints and medical necessity, this usually happens for a fixed amount of time if no complications arise. For this reason, the treating hospital often only knows the vital status of a patient in that specific time period. This practical limitation has made the use of binary outcomes popular where the vital status of a patient is only considered at the end of the follow-up period.

This can be formalised as follows: suppose we have $j = 1, ..., k$ hospitals and consider for each patient $i = 1, ..., n_j$ being treated at hospital j the binary outcome $X_{i,j}$, which is equal to zero if a patient is alive 90 days after surgery and one if the patient is deceased. For duration of stay, consider $X_{i,j}$ to be zero if a patient is still at the hospital 21 days after surgery and one if the patient was discharged. Outcomes for some patients may be censored, indicating that we did not observe the outcome. This can happen for a number of reasons, for example when a patient can no longer be reached for follow-up. We assume that these patients had desirable outcomes (not deceased), but it is also possible to not include these patients into the study. Furthermore, for each patient we have p prognostic factors denoted by $\boldsymbol{Z}_i$.

Table 1 Description of the characteristics of patients in the DCRA data set between 01/01/2019 and 31/12/2021

71 hospitals, 25367 patients	Number	Percent (%)
Gender		
Female	11857	46.7
BMI		
< 18.5	448	1.8
18.5–25	10013	39.9
25–30	9869	39.3
> 30	4763	19
Age (years)		
≤ 60	5800	22.9
61–70	6612	26.1
71–80	8506	33.5
≥81	4448	17.5
Charlson comorbidity index		
0	12920	50.9
1	5533	21.8
2	3630	14.3
3	1691	6.7
4	777	3.1
5	277	1.1
6	286	1.1
7	137	0.5
8	56	0.2
9	25	0.1
≥ 10	35	0.1
ASA score		
1-2	16773	66.1
3-5	8594	33.9
Double tumor		
Yes	698	2.8
Preoperative tumor complications		
Yes	8297	32.8
T stage		
0	149	0.6
1	3457	13.6
2	5490	21.6
3	12620	49.7
4	3651	14.4
M stage		
Yes	2385	9.4
Emergency resection		
Yes	2125	8.4
Additional resection metastasis		
Yes	741	2.9
Additional resection ingrowth		
No	23213	91.5
Extensive	975	3.8
Restricted	1179	4.6

There are many disadvantages associated with the use of a binary outcome. The DCRA uses a follow-up time of 90 days for postoperative mortality, meaning that deaths later than 90 days after surgery will be ignored. Moreover, the outcome does not reflect how long after surgery a patient passed away. As a consequence, there is no difference in outcome between a patient who passed away one day after surgery and a patient who died 90 days after surgery. Additionally, the choice of follow-up duration is often relatively arbitrary. Choosing a slightly different follow-up time (e.g. 85 days instead of 90) can already significantly change the results of quality control methods. Finally, the use of a binary outcome introduces a time delay in the information stream. As vital status is only registered 90 days after surgery, no information on a patient is known before that point in time. This means that binary outcomes provide "outdated" information, potentially leading to delays in the detection of deviations.

In the following sections, we describe two methods based on binary outcomes that can be used to inspect the quality of care. We introduce the term *failure*, which is either death (< 90 days) or prolonged hospital stay (> 21 days). We consider the problem of inspecting the quality of care in a specific *study period*, which is between 01/01/2019 and 31/12/2021 for the DCRA.

2.2.1 Funnel plot

The funnel plot (Spiegelhalter 2005) can be used to compare the proportion of *failures* between different hospitals. Suppose that p_0 is an acceptable (baseline) probability of failure for a patient at the end of follow-up. Usually, such a probability is not known and the average failure probability over all patients at all hospitals is considered instead. We would now like to test the null hypothesis that patients at hospital j have an acceptable failure probability against the alternative that they do not. For this, we consider the proportion of failures at the hospital during the study period: $\gamma_j = \frac{\sum_{i=1}^{n_j} X_{i,j}}{n_j}$. By the central limit theorem, this proportion is asymptotically normally distributed. This means that we can conclude that the probability of failure at this hospital is not in an acceptable range with confidence level $1 - \alpha$ when the proportion is outside the prediction limits:

$$\gamma_j \notin \left[p_0 + \xi_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n_j}}, p_0 - \xi_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n_j}} \right], \tag{1}$$

where ξ_α is the α−th quantile of the standard normal distribution. This means that if γ_j is larger/smaller than the upper/lower boundary, the probability of failure is larger/smaller than baseline, indicating that the quality of care at this hospital is worse/better than expected. Note that the prediction limits only depends on the number of patients treated at a hospital and baseline failure probability.

Simply comparing the proportions of failure does not adequately capture the complexity of treatment between patients. Some patients might have worse prognostic factors than others, making an undesirable outcome more likely. To account for this, a risk-adjusted procedure can be considered where for each patient an individual probability of failure p_i is modelled using logistic regression. We determine the expected number of failures at a hospital as $E_j = \sum_{i=1}^{n_j} p_i$. A risk-adjusted proportion of failures is then given by $\gamma_j^{\mathrm{RA}} = \frac{O_j}{E_j} \cdot p_0$

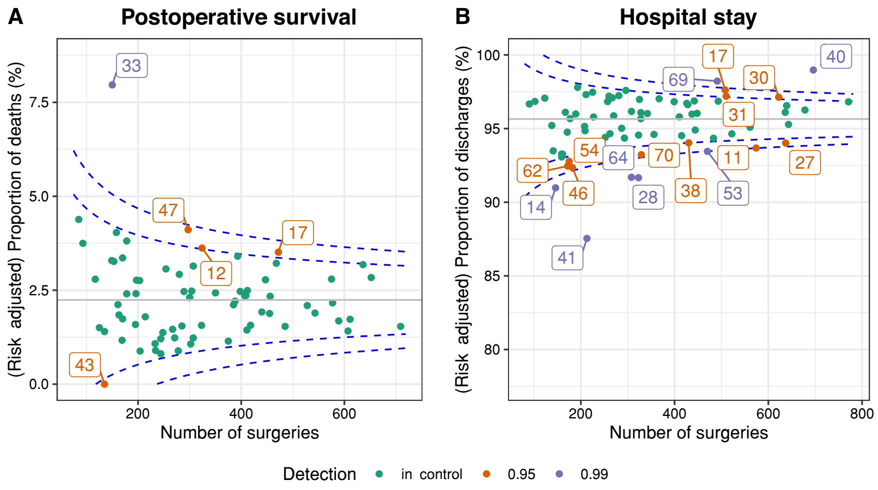

Fig. 1 Funnel plots of DCRA data between 01/01/2019 and 31/12/2021 for two different outcomes: **a** Postoperative survival 90 days after surgery; **b** Hospital stay 21 days after surgery. The dashed blue lines indicate the 0.95 and 0.99 prediction limits. Each dot represents a hospital, with the colour representing whether it falls outside of the specified prediction limits

where O_j is the observed number of failures at hospital j. For the risk-adjusted procedure, γ_j^{RA} is used instead of γ_j in Eq. (1) to draw conclusions.

When constructing a (risk-adjusted) funnel plot, we determine the proportions γ_j^{RA} for all k hospitals ($j = 1, ..., k$) and plot them against the number of treated patients at that hospital in a scatter plot, along with the prediction limits in Eq. (1). An example of risk-adjusted funnel plot is given in Fig. 1.

The goal of funnel plots from a mathematical point of view is to check for an increase/decrease in the failure probability at an institution over a fixed time period. In practice funnel plots are often used to continuously compare the performances of hospitals (Warps et al. 2021), with many consecutive funnel plots being constructed over overlapping time periods. This approach introduces multiple problems, such as an increased probability of a type I error incurred by repeatedly performing a dependent testing procedure. In addition, hospitals that have had a good historical performance and end up in the right lower quadrant of Fig. 1a may not be detected by future funnel plots due to the buffer they have built up in previous years. Finally, with this approach it is not clear how to handle past information of hospitals that have been signalled by a funnel plot. Due to these disadvantages, control charts that allow for the continuous inspection of the quality of care should be used for the intended goal. We focus on such control charts in the following sections.

2.2.2 Binary CUSUM

The binary cumulative sum (CUSUM) chart (Steiner et al. 2000) is a control chart which can be used to test for an increase or decrease in the failure probability of a process. CUSUM charts can be used to test hypotheses sequentially, meaning that the test can be performed after observing each individual outcome. This is not the case for the funnel plot, where the test is performed at the end of the study duration. Consider a single hospital j and suppose we want to test whether the post surgery failure probability has increased from

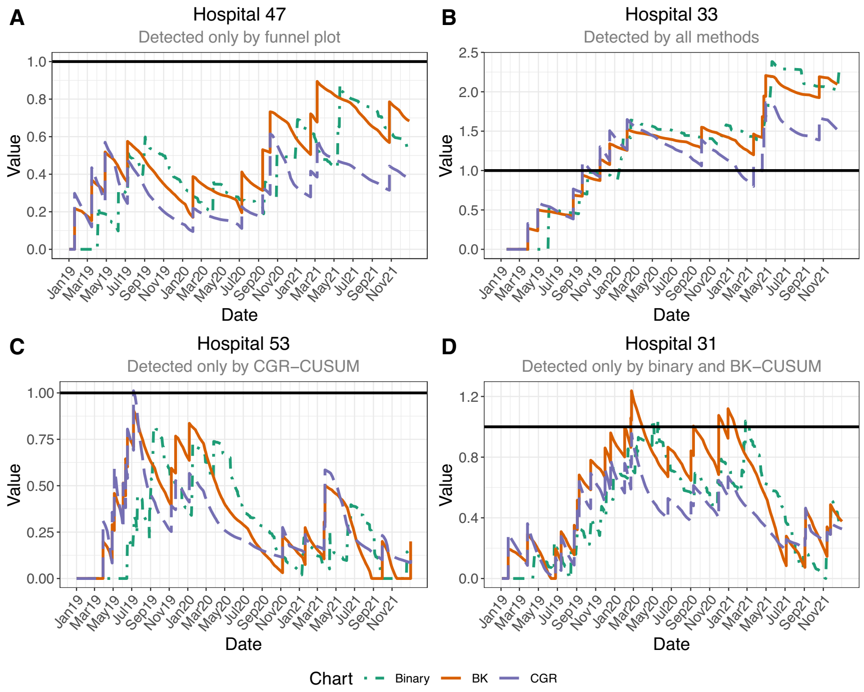

Fig. 2 Binary (dot-dash), BK- (solid) and CGR-CUSUM (dashed) control charts for postoperative survival for four hospitals: **a** 47, **b** 33, **c** 53, **d** 31. Charts were scaled with respect to their control limits, resulting in a shared control limit with value one. A signal is produced when a chart surpasses the control limit

p_0 to p_1 (with $p_1 > p_0$) starting from some chronological patient $\nu \geq 1$. In other words, we are looking for a change point in the failure probability at a single hospital. Again, p_0 is usually not known in practice and therefore determined as an average over all patients. Choosing p_1 can pose considerable challenges; therefore we consider the Odds Ratio $OR = \frac{p_1(1-p_0)}{p_0(1-p_1)} =: e^\theta$. Choosing $e^\theta > 1$ results in a test for an increase in failure rate, while $e^\theta < 1$ produces a test for detecting a decrease in failure rate. The odds ratio is often chosen to be equal to two (Steiner et al. 2000), but a practical reason for this choice is usually lacking.

The binary CUSUM for hospital j after observing the outcome of the n−th chronological patient is given by:

$$S_{n,j} = \max\left(0, S_{n-1,j} + W_{n,j}\right), \tag{2}$$

where

$$W_{n,j} = X_{n,j} \ln\left(e^\theta\right) + \ln\left(\frac{1}{1 - p_0 - e^\theta p_0}\right). \tag{3}$$

For a risk-adjusted procedure the patient specific baseline probability of failure $p_{0,i}$ is first modelled using logistic regression and afterwards substituted for p_0 in Eq. (3). Note that

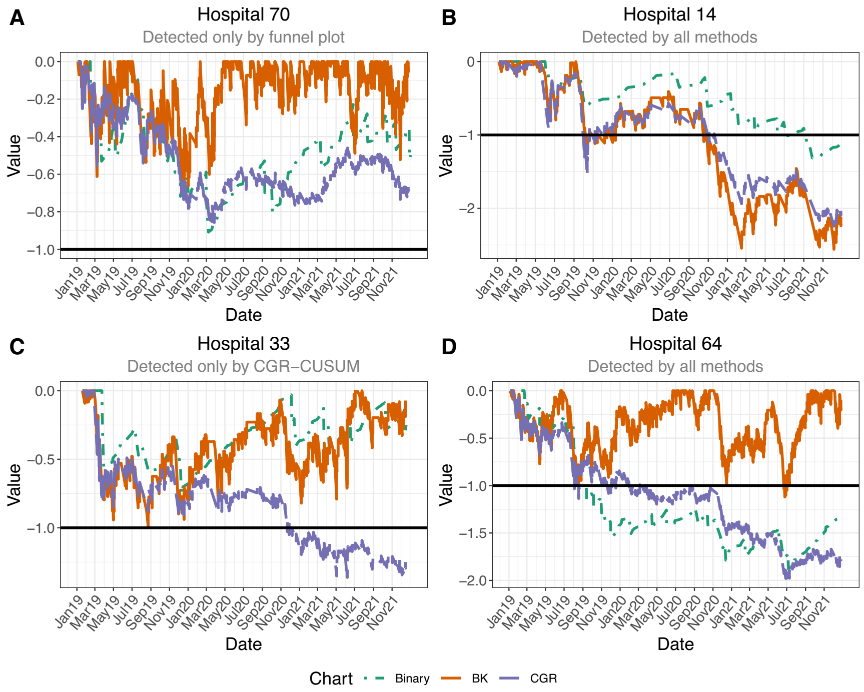

Fig. 3 Binary (dot-dash), BK- (solid) and CGR-CUSUM (dashed) control charts for hospital stay for four hospitals: **a** 70, **b** 14, **c** 33, **d** 64. Charts were scaled with respect to their control limits, resulting in a shared control limit with value one. A signal is produced when a chart dips below the control limit

$W_{n,j}$ is positive if patient n has failed, and negative otherwise. This means that the value of the binary CUSUM chart increases when patients fail and decreases when favourable outcomes are observed.

When constructing a binary CUSUM chart, the value of the chart is plotted against the chronological patient number or against the time at which the outcome was observed for said patient. We signal a change in the failure rate when the value of the chart exceeds the value of a pre-specified *control limit h*. We discuss how to determine such a control limit in Section "Control limits". Note that the value of the chart is cut-off at zero, meaning that the CUSUM cannot build up a buffer when the proportion of observed desirable outcomes is large. Examples of binary CUSUM charts can be seen in Figs. 2 and 3.

In addition to employing a binary outcome, the major limitation of a binary CUSUM chart is the need to specify the odds ratio in advance. In most practical scenarios, there will be no information about the expected increase in the failure probability at sub-optimally performing hospitals. Choosing an unsuitable value for the odds ratio may cause the procedure to lose statistical power or cause a delay in detections. Grigg et al. (2003); Grigg and Farewell (2004) describe this problem and other considerations when using risk-adjusted CUSUM charts.

2.3 Survival charts

Instead of considering a binary outcome, it is often advantageous to consider the time until event for each patient, also known as a survival outcome. We then consider $T_{i,j}$ to be the time (f.e. in days) until death or discharge of patient i at hospital j. In contrast to binary outcomes, *failure* here means the observed death or discharge of a patient. Considering the discharge of a patient as a failure might seem unnatural, as discharge is usually considered a positive outcome. The aforementioned dilemma is examined in Section "Lower CUSUM". For the rest of this section we drop the subscript j and consider the problem of detecting an increase in the failure rate at a *single* hospital.

We model the survival times using a Cox proportional hazards model (Cox 1972) where the patient specific hazard rate is given by $h_i(t) = h_0(t)\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_i)$; the term $h_0(t)$ represents the baseline hazard and $\boldsymbol{\beta}$ a coefficient vector indicating the risk associated with the corresponding prognostic factors. The baseline hazard indicates an acceptable failure probability and is usually not known in practice, similarly to p_0 for binary charts. The baseline hazard (and coefficient vector) is usually estimated by fitting a Cox model on all hospitals together, thereby recovering an "average" performance measure. As we are only interested in following a patient during their participation in the study, we consider the risk indicator $Y_i(t)$ which is one if a patient is *at risk* of failure at time t and zero otherwise. A patient is not at risk of failure when they have not entered the study yet, or after they have failed or their observation has been censored. To keep track of the number of failures at a hospital in real time, we introduce a counting process $N(t)$, which is equal to the number of failures at a hospital at time t after the start of the study. As for the binary outcomes, survival outcomes can also be right-censored. In our case we again consider censored observations as having a desirable outcome at the censoring time, but other assumptions are also possible.

There are many advantages to the use of survival outcomes instead of binary outcomes. First of all, the binary charts required a choice for an outcome threshold (i.e. 90 days) whereas for survival outcomes this is no longer necessary. The absence of such an (arbitrary) choice therefore no longer impacts the resulting conclusions, making the use of survival charts consistent. This is a major advantage, especially when there is no clear motivation for the choice of a threshold. Additionally, the survival outcome holds the most recent information about the status of a patient at any point in time, possibly leading to quicker detections. The disadvantages of using survival outcomes primarily relate to practical considerations. The registration of survival outcomes requires a continuous follow up of patients, which is often not feasible for long periods after the initial procedure. For this reason, times to event may only be known exactly in a fixed period after surgery, as is the case for postoperative survival in the DCRA data. The duration of stay is known exactly for each patient.

In the following two sections we describe two control charts that use survival outcomes, with the goal to detect an increase in the failure rate at a single hospital during the study period.

2.3.1 BK-CUSUM

The Biswas & Kalbfleisch CUSUM (Biswas and Kalbfleisch 2008) (BK-CUSUM) is a control chart for survival outcomes, used to test for a change in the failure rate of a process. This chart can be seen as the survival analogue of the binary CUSUM in Section "Binary CUSUM". The BK-CUSUM however is used to test slightly different hypotheses. Where the binary CUSUM was used to test for an increase in failure rate starting from a chronological patient $\nu \geq 1$, the BK-CUSUM is used to test for a sudden change in the failure rate of all patients in the study at some time $s \geq 0$ after the start of the study. This sudden change is then described by an increase in the baseline hazard rate from $h_0(t)$ to $h_0(t)\exp(\theta)$, with $\exp(\theta)$ called the *hazard ratio*. The hazard ratio can be seen as an analogue of the Odds Ratio in binary CUSUM charts. It has to be chosen in advance, with a wrong choice potentially leading to delayed and/or false detections (Gomon et al. 2022). To facilitate testing the described hypotheses, we consider a time-constrained counting process $N(s,t) = N(t) - N(s)$ for $t \geq s$, which keeps track of the failures at a hospital between times s and t.

The BK-CUSUM is given by:

$$BK(t) = \max_{s:0 \leq s \leq t} \left\{ \theta N(s,t) - \left(e^{\theta} - 1\right)\Lambda(s,t) \right\}, \tag{4}$$

where $\Lambda(s,t) = \sum_{i=1}^{n_j} \int_s^t Y_i(u)h_i(u)du$ is the accumulated cumulative hazard at a hospital between times s and t. Heuristically, the accumulated cumulative hazard indicates how much risk was built up by patients present at a specific hospital between times s and t. Note that the value of the BK-CUSUM will increase by θ any time a failure is observed and drift downwards at all other time points, with the downward slope depending on the current amount of patients at the hospital and their risk of failure. Even though the BK-CUSUM does not have an explicit cut-off at zero, the maximisation term implicitly stops the value of the chart dropping below zero. As with the binary CUSUM, the value of the BK-CUSUM is plotted against study time and a signal is produced when the value exceeds a pre-defined control limit h.

In parallel to the binary CUSUM, the choice of a hazard rate e^{θ} complicates the use of the BK-CUSUM, as it is usually not known in advance what the increase in hazard rate will be at sub-optimally performing hospitals. (Biswas and Kalbfleisch 2008) chose to use a value of $e^{\theta} = 2$, corresponding to the use of $OR = 2$ in binary CUSUM charts. Even though this choice seems to be commonplace (Gomon et al. 2022; Steiner et al. 2000; Biswas and Kalbfleisch 2008), there is no guarantee that an OR equal to 2 will perform well in real life applications.

2.3.2 CGR-CUSUM

The Continuous time Generalised Rapid response CUSUM (Gomon et al. 2022) (CGR-CUSUM) is a control chart for survival outcomes, similar to the BK-CUSUM. There are two key differences between the CGR- and BK-CUSUM. First of all, the CGR-CUSUM can be used to test the continuous time alternative to the discrete time hypotheses considered for the binary CUSUM: the detection of a change in failure rate starting from some chronological patient $\nu \geq 1$, ignoring the information of all patients before this patient. The BK-CUSUM can be used to test for a change point in time, considering the information of all patients still in the study. Secondly, the CGR-CUSUM is also used to test for a sudden increase of $\exp(\theta)$ in the baseline hazard rate, but θ is determined "automatically" using a maximum likelihood estimator. Suppose at time $t > 0$ after the start of the study, n patients have been treated at a specific hospital. The CGR-CUSUM chart is then given by:

$$CGR(t) = \max_{1 \le \nu \le n} \left\{ \hat{\theta}_{\ge \nu} N_{\ge \nu}(t) - \left(e^{\hat{\theta}(t)} - 1 \right) \Lambda_{\ge \nu}(t) \right\}, \tag{5}$$

where $N_{\ge \nu}(t)$ counts the number of failures at time t after the entry of patient ν and $\Lambda_{\ge \nu}(t) = \sum_{i \ge \nu} \int_0^t Y_i(u) h_i(u) du$ is the accumulated cumulative hazard rate of those patients. The maximum likelihood estimator is then given by:

$$\hat{\theta}(t) = \max \left(0, \ln \left(\frac{N_{\ge \nu}(t)}{\Lambda_{\ge \nu}(t)} \right) \right).$$

Details on the calculation of the CGR-CUSUM are described in Appendix A. Similarly to the BK-CUSUM, the CGR-CUSUM will jump up at any observed failure and drift downwards when no failures are observed. There is also an implicit cut-off at zero due to the maximisation term. In contrast, the CGR-CUSUM no longer makes upwards jumps of fixed size due to the updating maximum likelihood estimate $\hat{\theta}(t)$. This estimate becomes large when recent patients fail rapidly compared to their accumulated cumulative hazard, and small when failures happen infrequently. The maximum likelihood estimate provides an indication on how the recent failure rate at a hospital compares to the baseline failure rate (usually the average over all hospitals). The value of the CGR-CUSUM is plotted against study time and a signal is produced when the value of a control limit h is exceeded.

The value of the maximum likelihood estimator $\hat{\theta}(t)$ can become very unstable, especially when some patients fail very quickly after their study entry. For this reason, (Gomon et al. 2022) chose to limit the value of the estimator between $0 \le e^{\hat{\theta}(t)} \le 6$, thereby aiming to detect an increase in the hazard rate of by at most a factor 6. While the difference between the hypotheses concerning the increase in a failure rate from a specific patient in the CGR-CUSUM and a change at some point in time for the BK-CUSUM may seem trivial, it changes the considered problem significantly. The CGR-CUSUM assumes that at

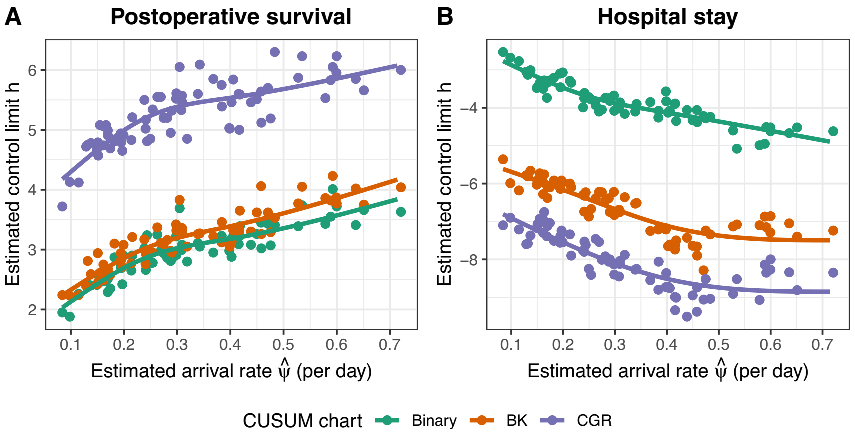

Fig. 4 Estimated control limits vs estimated daily arrival rate of patients for all hospitals in the DCRA data set, with each dot representing the simulated control limit for a single hospital. The left panel **a** indicates the control limits determined for postoperative survival and the right panel **b** for hospital stay. The lines represent a monotonous cubic regression spline fit through the simulated control limits. Control limits used in the study were determined from the fitted lines

some point in time all *future* patients will have an increased probability of failure, whereas the BK-CUSUM assumes that at a certain point in time all *current and future* patients have an increased failure probability.

2.4 Control limits

Even though CUSUM charts give a visual representation of a hospital's performance, we require a practical method to determine when a change in quality should be signalled. This is achieved using a *control limit*, a numeric value indicating when the value of a chart has become too large. A CUSUM chart is constructed until the time when its value exceeds the control limit, signalling a change in the failure rate at that specific hospital.

A natural question that arises is how this control limit should be chosen. Two approaches are commonly used: either the expected time until detection or the probability of wrongfully detecting an in-control hospital during the study period is restricted. For a single CUSUM chart, define the *run length* as the time since the start of the study until the first detection. The first approach chooses a control limit such that the expected run length of an in-control hospital is restricted to some suitable quantity. Using the second approach, the control limit is chosen such that an arbitrary in-control hospital has at most probability α to be detected during the study period. In other words, the type I error probability is restricted over a certain time period. To the best of our knowledge, for both approaches no algebraic results for determining control limits for risk-adjusted survival CUSUM charts are available. For this reason, Monte Carlo simulation methods are usually employed to determine a control limit.

A commonly overlooked problem for CUSUM charts is that hospitals can differ in the number of patients treated during the study period. (Gomon et al. 2022) have shown that this difference warrants the use of control limits varying depending on the volume of patients treated at a hospital. Mathematically we model the number of patients treated at a specific hospital by using a Poisson process with rate ψ. For each hospital we therefore estimate a *Poisson arrival rate* $\hat{\psi}$ using the Poisson maximum likelihood estimator and determine a hospital specific control limit for this rate. The estimated rate $\hat{\psi}$ can be seen as the expected number of people to be treated at a hospital per time unit (e.g. per day).

We choose a simulation approach to determine hospital specific control limits. Suppose we want to estimate the control limit associated with an estimated arrival rate of $\hat{\psi}_j$ for hospital j over a study period of length T. We fit a Cox model on all available data (all hospitals):

$$h_i(t) = h_0(t)e^{\boldsymbol{\beta}^\top \mathbf{Z}_i}$$

to obtain an estimate for the baseline hazard $\widehat{h_0}(t)$ and regression coefficients $\widehat{\boldsymbol{\beta}}$. N dummy hospitals are then generated by bootstrapping patient characteristics from the full data set, with the amount of patients at each simulated hospital determined by independently sampling from a Poisson process with rate $\hat{\psi}_j$. Survival times are then generated using the inverse transform method described in Bender et al. (2005). CUSUM charts are then constructed over the study duration T for each hospital in the dummy data set and a value for the control limit is chosen such that at most a proportion α of the simulated hospitals is signalled using this value. This is the empirical analogue of choosing a type I error rate of α over a study duration of T.

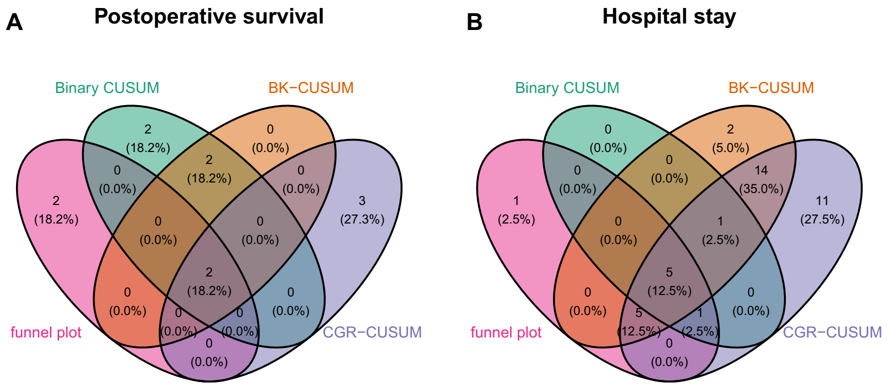

Fig. 5 Venn diagrams detailing the numbers of (shared) detections between the considered methods for: **a** postoperative survival; **b** Hospital stay. Each ellipse represents a method, from left to right: funnel plot, Binary CUSUM, BK-CUSUM and CGR-CUSUM

To compensate for variance in the simulation procedure we propose to determine control limits for hospitals with different arrival rates ψ and fit monotone cubic regression splines (see Section 5.3.1 of Wood (2017)) on the determined control limits. We can then use the smoothed values as control limits instead, see Fig. 4 for an example. The reason for fitting monotone functions on the estimated control limits stems from Gomon et al. (2022), who showed that a variant of the CGR-CUSUM is asymptotically normally distributed with variance increasing with the arrival rate ψ. Heuristically this means that in-control hospitals treating a larger number of patients are more likely to be detected than hospitals treating fewer patients over a fixed study period.

2.5 Lower CUSUM

For certain studies, detecting a decrease in the failure rate may be of greater interest than an increase. The goal would then be to detect hospitals performing better than what is deemed to be on target. This is easily achieved by choosing a value of $\exp(\theta) < 1$ for the binary and BK-CUSUM in Eqs. (2) and (4). For the CGR-CUSUM this is achieved by restricting the maximum likelihood $\frac{1}{6} \leq \exp(\widehat{\theta}(t)) \leq 0$ in Eq. (5). Page (1954) suggested a procedure where we test for both an increase and decrease in the failure rate at the same time by plotting the two procedures in a single plot, with (usually) the CUSUM for a decrease in failure rate facing downwards. This is often called a *two-sided* CUSUM procedure, with one *upper* and one *lower* CUSUM chart. We will adhere to the usual naming convention by referring to CUSUMs for an increase/decrease in failure rate as upper and lower CUSUM charts respectively.

The lower CUSUM procedure solves the problem discussed in section "Survival charts", which is that a discharge from a hospital is not usually regarded as "failure". We are indeed interested in detecting a **decrease** in the discharge rate at hospitals and will therefore construct lower CUSUM charts for this outcome.

Table 1 (continued) BMI stands for Body Mass Index and ASA for American Society of Anesthesiology

3 Results

In this section, we compare the performance of the four methods mentioned earlier for detecting changes in the quality of care at the hospitals in the DCRA data set. First we describe the approach taken to ensure comparability of the results obtained from the methods, then we look at the resulting detections for postoperative survival and hospital stay separately.

We construct the funnel plot with 95 percent prediction limits (Eq. (1)), including all 71 hospitals in the data set. For the CUSUM charts we first determine hospital specific control limits as described in Section "Control limits" with type I error guarantee $\alpha \leq 0.05$ over the duration of the study (three years) and $\exp(\theta) = 2$ (upper) or $\exp(\theta) = 0.5$ (lower) for both the binary and BK-CUSUM. In this way, the type I error of both funnel plot and CUSUM procedures over the study duration is restricted to five percent. The funnel plot and CUSUM charts are constructed for both outcomes. The resulting control limits for CUSUM charts can be seen in Fig. 4. We determine binary, BK- and CGR-CUSUM charts for all hospitals in the DCRA data set using their associated smoothed control limits. Chart and control limit calculations were performed using the `R` packages `success`(Gomon et al. 2023) and `mgcv` (Wood 2011). The resulting charts for all the hospitals are available as interactive `plotly` (Sievert 2020) charts in Additional files 1-2. The code used to perform the analyses is available at https://github.com/d-gomon/DCRA_CUSUM.

3.1 Postoperative survival

A Venn diagram detailing the number of detected hospitals for postoperative survival is shown in Fig. 5a. Notably, only 2 hospitals were detected by all methods. The funnel plot detected 4 hospitals in total, with 2 of those hospitals not detected by any of the CUSUM charts. The BK-CUSUM signalled 4 hospitals in total, also detected by the binary CUSUM. Furthermore, the binary CUSUM identified 2 additional hospitals that were not detected by any of the other charts. The CGR-CUSUM was the sole chart to detect 3 hospitals, but shared only 2 detections with the rest of the methods in total. Detection times in days since the start of the study for all hospitals can be found in Additional file 3. The BK-CUSUM had faster detection times than the binary CUSUM for all hospitals except for one. The CGR-CUSUM was fastest in detecting the 2 hospitals signalled by all other methods. Due to the stark contrast in detected hospitals, no meaningful comparison summaries of signalling times can be made. We attempt to explain the differences in these detections by looking at some of them individually. The CUSUM charts for hospital 47, which was detected only by the funnel plot, are displayed in Fig. 2a. The value of each chart is scaled by their respective control limit resulting in a shared control limit with value 1 to make visual comparisons. We can see that over the whole study period, failures were happening at this specific hospital at a steady rate. None of the charts go back to zero after the initial failure. The binary CUSUM initially departs from zero later than the survival CUSUMs due to the 90 day delay in observing the outcome. The BK- and CGR-CUSUM jump upwards at the same times (at failure times), but with different jump sizes. The BK-CUSUM always jumps up by θ whereas the CGR-CUSUM jumps upwards depending on the current value of $\widehat{\theta}(t)$. For the same reason both charts also drift downwards at different rates. This difference is particularly evident towards the end of the study, where the CGR-CUSUM takes significantly lower values than the BK- and binary CUSUM charts. From these charts we conclude that even though the proportion of failures over the study period was high at this specific hospital, the failures were spread out over the total study duration and most failure times were acceptable.

To investigate why some charts produce signals faster than others, we take a look at hospital 33 in Fig. 2b. This hospital was detected by all charts, with the CGR-CUSUM leading in detection time (243 days), followed by the BK-CUSUM (300 days), binary CUSUM (369 days) and finally the funnel plot (3 years - end of study). Due to its variable jump size, the CGR-CUSUM crosses its control limit after the fourth death is observed. The BK-CUSUM does so only after the fifth death, causing a delay in detection. The same reason holds for the discrepancy between the BK- and binary CUSUM, combined with the 90 day delay. At the time of detection, the maximum likelihood estimate for the CGR-CUSUM was $\widehat{\theta}(t) = 6$ (maximal allowed value), meaning that patients were failing at an extremely rapid rate compared to the national average in those three years. Due to the choice of an alternative failure rate of $e^{\theta} = 2$ in the BK- and binary CUSUM, we therefore experienced delays in detection using those charts.

The maximum likelihood estimator $\widehat{\theta}(t)$ can be a double edged sword if not used properly. An example can be seen in Fig. 2c showing Hospital 53. The rapid consecutive failures in the beginning of the study inflate the estimated value of the hazard ratio, causing a possibly premature detection by the CGR-CUSUM. After the initial spike in all CUSUM charts, the values of all charts rapidly drift downwards, indicating good performance in that part of the study. For this study, the choice of restricting $e^{\widehat{\theta}(t)} \leq 6$ might not have been optimal.

Finally, we take a look at Hospital 31 in Fig. 2d; the BK- and binary CUSUM both cross their control limits but the CGR-CUSUM does not. This could indicate that even though many consecutive failures occur, patients experiencing failures had poor prognostic factors or were failing at acceptable times and therefore should not contribute much to the increase in the value of the chart.

3.2 Hospital stay

The number of hospitals signalled by each method with a significantly lower rate/proportion of patient discharge is displayed in a Venn diagram in Fig. 5b. Detection times in days since the start of the study for all hospitals can be found in Additional file 4. Contrary to the results for postoperative survival, both survival charts detect way more hospitals than the binary charts with the CGR-CUSUM detecting more than half (37) of all hospitals. The binary CUSUM detects almost exclusively the same hospitals as the funnel plot, with the funnel plot agreeing on many detections with the BK-CUSUM. A total of 14 hospitals were detected by both survival charts, but not by the binary charts, stressing that these methods test very different hypotheses. The BK-CUSUM seems to yield the most “balanced” detections, yielding only 2 exclusive detections and overlapping with at least one other method on the rest of its signals.

Hospital 70 in Fig. 3a was detected only by the funnel plot. We can see in the values of the binary CUSUM that over the whole study period quite a few patients were not discharged 21 days after surgery. The survival charts only clearly show when a patient is discharged, indicated by an upward jump. The downward slope indicates how many patients are at the hospital at that point in time, but does not allow for easy comparisons between time periods. This hospital had a period of slow discharges between January 2019 and March 2020, followed by a reasonable discharge rate afterwards.

In Fig. 3b we can see that Hospital 14 had two periods when many patients were staying at the hospital, but were not being discharged at an acceptable rate. In the periods between September and November 2019, a large proportion of patients were being discharged slower than usual. The binary chart does not display any deterioration in the period, as all these patients were most likely being discharged within the 21 day window. It looks like something similar happened in the period after November 2020, with the binary CUSUM finally signalling at the end of the study period. This example shows how the use of binary vs survival outcomes can influence the conclusions.

Interestingly, Hospital 33 in Fig. 3c was detected only by the CGR-CUSUM, with both the BK- and binary CUSUM nearly exceeding their control limit at an early stage of the study. The CGR-CUSUM produces a signal, but only in the second half of the study period. The maximum likelihood estimate of the CGR-CUSUM seems to converge to a value around $\exp(\hat{\theta}(t)) = 0.7$ throughout the three years under consideration. The binary and BK-CUSUM are looking for a halving ($\exp(\theta) = 0.5$) of the discharge rate and therefore do not find sufficient evidence to produce a signal. This raises the question of whether 0.7 times the baseline rate should be deemed acceptable or not.

Finally, the binary CUSUM follows a similar progression as the CGR-CUSUM for Hospital 64 in Fig. 3d. Both charts drift downwards throughout the whole study duration, indicating a persistent delay in patient discharge. The BK-CUSUM on the other hand stays close to zero, suddenly drifting downwards at multiple occasions and finally producing a signal at the end of the study. The CGR-CUSUM estimates a discharge rate of about 0.74 the national average, leading to delays in detection with the BK-CUSUM. It is surprising that the binary CUSUM, which also looks for a halving of failure rate, follows a similar trend to the CGR-CUSUM. This is likely due to a combination of the choice of cut-off at 21 days post surgery and looking for a halving of discharge rate. Whereas the discharge rate looking at overall hospital stay does not seem to be halved, it is possible that the 21-day post surgery discharge rate is very close to half the national rate. This example clearly shows that the choice of cut-off as well as value of θ greatly affects the outcome.

4 Discussion

We applied and compared four methods for the continuous inspection of the quality of care after colorectal surgery in the Netherlands, exploring the funnel plot and three methods based on CUSUM statistics. For the clinical outcome we looked at both survival after surgery as well as (prolonged) hospital stay. We found that survival charts outperformed binary charts with respect to the detection time of deviations at the considered hospitals. We discussed the differences between methods and highlighted some pitfalls that can occur.

The control limits used for signalling a decrease in the quality of care at a hospital are usually determined by a simulation study, where either a single value is used for all hospitals (Steiner et al. 2000) or the hospitals are grouped into categories by size (Biswas and Kalbfleisch 2008) with each group using a separate control limit. When grouping hospitals, the difference between the smallest and largest hospital within a group may be substantial. In contrast to previous studies, we determined control limits for the CUSUM charts using a novel approach where each hospital has a unique control limit, which can be considered "fair" with respect to other hospitals. A drawback of this method is that it requires more computing time.

We found that the use of the funnel plot as continuous inspection scheme is unsuitable, not only yielding detections which are almost certainly false, but also not providing any insights into the reasons why certain hospitals were detected and others were not. Since the funnel plot can only be constructed once the study period is over, it logically resulted in very slow detections of deviations as well. Using the same information, the binary CUSUM chart was able to detect deviations much faster, and provided insights into when deviations at a hospital began and ended. The considered BK- and CGR-CUSUM charts boasted even quicker detection times, but due to the use of a survival outcome yielded very different detections, especially for hospital stay.

We used lower CUSUM charts to detect a decrease in the quality of a care regarding hospital stay, where long stay times are considered unfavourable. This contrast with the usual application of lower CUSUMs for the detection of an increase in the quality of care. Survival CUSUM charts for this outcome yielded signals for over half the considered hospitals, indicating that either deviations were present at many hospitals or that the charts were too sensitive for the desired inspection procedure. As inspecting the duration of hospital stay is more useful for managerial matters, the use of a binary CUSUM might be appropriate here. Even though the choice of a threshold (21 days in this case) can influence the final conclusions, it can be based on practical considerations such as preventing overcrowding and cutting down on costs. Before making decisions of this nature, it is important to consider the objective of the continuous monitoring procedure.

The binary and BK-CUSUM require prior knowledge about the expected increase in failure rate, in the form of a choice for the odds or hazard ratio e^{θ}. We believe that many researchers use CUSUM charts with $e^{\theta} = 2$ simply based on previous research. However, to the best of our knowledge there is no evidence that such a value is appropriate for many, if not any practical application in medicine. On the other hand, there are many simulation studies showing that misspecifying this parameter can lead to delays in detections as well as an increase in false detections. Even though the CGR-CUSUM attempts to solve this problem by using a maximum likelihood estimator, the authors recommend limiting the allowed range to at most 6 times the baseline rate to limit the amount of false detections. This choice is also arbitrary and might be unsuitable for the problem at hand, although it is less likely to influence the study results if the true increase in failure rate at the considered hospitals is not (much) larger than the chosen value.

A good understanding of the hypotheses being tested by each of the charts is vital for a successful inspection procedure. Whereas the difference between the funnel plot and the CUSUM charts is evident, the differences between the CUSUM charts are not as apparent. The binary and CGR-CUSUM test the same alternative hypothesis that the failure rate has increased starting with the surgery of a certain patient, with the CGR-CUSUM using survival outcomes (instead of binary) to test this. The BK-CUSUM is used to test the hypothesis that the failure rate has increased for all patients suddenly, at a certain point in time. This may not sound like a big difference, but can significantly influence the conclusions, especially if one of the two assumptions is more likely to be true.

An important matter to consider when using control charts is that a signal does not imply deviations at the hospital in question. There are many possible reasons for an in-control hospital to be signalled, with a false detection being one of the most obvious. Signals should therefore not be used to draw causal conclusions, but should be used with an additional evaluation of past performance at a specific hospital to determine potential deviations. CUSUM charts provide visual and exact information on when deviations may have started and in which time periods they were exacerbated.

As the quality of care improves, the occurrence of adverse events can become exceedingly rare, resulting in what is called a high-quality process. In such cases standard control charts such as the CUSUMs considered in this article can no longer adequately be used to monitor the process. Time-between-event charts have been developed specifically for the inspection of high-quality processes (for a review see Ali et al. (2016)). Unfortunately in this setting prognostic factors can no longer be incorporated into the model, but the heterogeneity between patients can in some cases still be modelled using an overdispersion parameter (e.g. Albers (2009)).

5 Conclusions

It is crucial to understand what the assumptions of a control chart are and which hypotheses are being tested before choosing the appropriate method. For the considered CUSUM charts the major pitfalls are the choices of parameters and understanding the hypotheses being tested. It is also important to be aware of the many pitfalls of binary and survival CUSUM charts when interpreting the resulting charts. The use of survival over binary CUSUM charts can yield quicker detections and provides different insights about failure rates at hospitals during the study. The use of survival CUSUM charts may not always be appropriate, for example if the exact survival time of the patient is not relevant for the problem at hand. Constructing multiple CUSUM charts can give additional insights into the behaviour of the process. A comparison between different charts can aid the interpretation of chart values and help to distinguish true from false detections. The choice of a control limit is a complicated issue for the use of control charts. We advise to consider our new approach discussed in this article, which assigns a similar value of the control limit for hospitals that treat a similar number of patients. It eliminates the problem of grouping hospitals and provides a clear reason for the difference between control limits at different hospitals. In the future, survival CUSUM charts should be compared with other survival charts to determine how well they perform with respect to detection time and sensitivity to changes in the failure rate.

A limitation of our study is that we considered hospitals to have a constant stream of patients arriving over the whole study period. If this is not the case, the control limit used by a hospital might not be appropriate over the whole study period. Such a change in the number of treated patients is especially apparent in the considered downward CUSUMs (see Fig. 3) where the slope of the CUSUM chart can spontaneously increase or decrease, leading to possibly premature or delayed detections. A similar issue can arise when the risk distribution of the patient population changes over time. For the second issue dynamic control limits have been proposed by multiple studies (an overview is given in Tighkhorshid et al. (2019)). Dynamic control limits could however be used to solve both of these issues simultaneously.

The use of survival charts for continuously monitoring health care requires the information on the status of patients to come in continuously. Although hospitals usually record information in a timely manner, the aggregation of data across hospitals into a registry happens in a delayed fashion. Measures should be taken to streamline the aggregation of data into registries to improve the quality of care. In the meanwhile, individual hospitals can continuously monitor their own quality compared to some historical average performance measure.

Appendix A: Details on the calculation of the CGR-CUSUM

In this section we sketch how the value of the CGR-CUSUM can be calculated in practice. This procedure is implemented in the R package success (Gomon et al. 2023). We will only cover the procedure for upper-sided CUSUM charts as the procedure for lower-sided CUSUM charts is very similar.

Before we can construct a CUSUM chart for a hospital, we first need to determine what the target performance measure is. For this reason, we determine the cox baseline hazard rate $h_0(t)$ and associated risk-adjustment coefficients $\boldsymbol{\beta}$ using historical data. This allows us to determine the patient specific hazard rate $h_i(t) = h_0(t)\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_i)$. Preferably the historical data is known to have in-control failures, but often all available data is used to determine baseline characteristics instead. In the following steps we assume that we have estimates for these quantities. There are multiple ways to estimate them, we make use of the survival package (Therneau 2020).

Given a data set containing the information on n patients at a single hospital we perform four steps.

Step 1: Order patients according to their surgery calendar time S_i for $i = 1, \dots, n$ and determine the K unique failure times of all patients. To minimise computational burden we usually only construct the chart at the failure times for upper-sided CUSUM charts. This results in K unique construction times at which we want to determine the value of the chart, denoted by t_k for $k = 1, \dots, K$. Let us denote the chronologically sorted union of the unique surgery times and unique failure times by $\{s_b\}$ with $b = 1, \dots, B$ and call these the vital times.

Step 2: For each $b = 1, \dots, B$ and each patient $i = 1, \dots, n$, calculate $\Lambda^i(s_b) = \int_0^{s_b} Y_i(u)h_i(u)du$ and store in a matrix $\boldsymbol{\Lambda}$ of dimensions $n \times B$ so that $\boldsymbol{\Lambda}_{ib} = \Lambda^i(s_b)$. Note that $\Lambda_i(s_b)$ is only non-zero when the patient has had surgery before the vital time ($S_i < s_b$). The rows of this matrix then represent the patients (ordered chronologically in calendar time) and the columns represent the vital times $s_1, \dots, s_B$. The values of the matrix then represent the total individual cumulative hazard built up by a single patient at the considered time. Note that the column sums of this matrix represent the total cumulative hazard at the hospital at the respective calendar time: $\sum_{i=1}^n \boldsymbol{\Lambda}_{ib} = \Lambda(s_b)$.

Step 3: Suppose we want to construct the chart at one of the vital times s_l. We can calculate $\Lambda_{\geq v}(s_l)$ from Eq. (5) using above matrix as follows: $\Lambda_{\geq v}(s_l) = \sum_{i=1}^n \boldsymbol{\Lambda}_{il} - \sum_{i=1}^n \boldsymbol{\Lambda}_{iv}$, where s_v is the vital time corresponding to the surgery time of patient v. The value of $N_{\geq v}(s_b)$ is easily calculated by considering only the failures of patients with surgery time larger than S_v. Having obtained both $\Lambda_{\geq v}(s_b)$ and $N_{\geq v}(s_b)$ we can calculate $\widehat{\theta}(s_b)$ and therefore the value of the CGR-CUSUM by iterating over all patients v with a surgery time smaller than s_l and taking the maximum of the obtained values.

Step 4: Repeat Step 3 for all K construction times considered in Step 1.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10742-023-00315-0.

Acknowledgements Not applicable.

Author's contributions MF wrote a proposal to enquire the data. DG, JS and MS conceived and designed the study. DG performed the statistical analysis. DG, JS, MF, HP and MS interpreted the results. DG and JS drafted the manuscript and all authors critically revised it. All authors read and approved the final version.

Funding None

Data availability The datasets analysed during the current study are not publicly available due to privacy concerns but are available from the Dutch Institute for Clinical Auditing (https://dica.nl) on a research request. The R code used for the analyses performed in this article is available from https://github.com/d-gomon/DCRA_CUSUM. The provided code allows for the reconstruction of all figures in this article and control charts for all hospitals in the DCRA data (see also additional materials).

Declarations

Conflict of interest None.

Ethics approval No ethical approval or informed consent was required under Dutch law.

Consent for publication Not applicable.

References

Albers, W.: Control charts for health care monitoring under overdispersion. Metrika **74**(1), 67–83 (2009). https://doi.org/10.1007/s00184-009-0290-z

Ali, S., Pievatolo, A., Göb, R.: An overview of control charts for high-quality processes. Qual. Reliab. Eng. Int. **32**(7), 2171–2189 (2016). https://doi.org/10.1002/qre.1957

Begun, A., Kulinskaya, E., Macgregor, A.J.: Risk-adjusted cusum control charts for shared frailty survival models with application to hip replacement outcomes: a study using the NJR dataset. BMC Med. Res. Methodol. **19**, 217 (2019). https://doi.org/10.1186/s12874-019-0853-2

Bender, R., Augustin, T., Blettner, M.: Generating survival times to simulate cox proportional hazards models. Stat. Med. **24**, 1713–1723 (2005). https://doi.org/10.1002/sim.2059

Biswas, P., Kalbfleisch, J.D.: A risk-adjusted cusum in continuous time based on the Cox model. Stat. Med. **27**, 3382–3406 (2008). https://doi.org/10.1002/sim.3216

Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc. Ser. B Stat. Methodol. **34**(2), 187–202 (1972). https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Diko, M.D., Goedhart, R., Does, R.J.: A head-to-head comparison of the out-of-control performance of control charts adjusted for parameter estimation. Qual. Eng. **32**(4), 643–652 (2019). https://doi.org/10.1080/08982112.2019.1666140

Fatt Gan, F., Sheng Yuen, J., Knoth, S.: Quicker detection risk-adjusted cumulative sum charting procedures. Stat. Med. **39**(7), 875–889 (2020). https://doi.org/10.1002/sim.8448

Gomon, D., Fiocco, M., Putter, H., Signorelli, M.: SUrvival Control Chart EStimation Software in R: the success package. arXiv (2023). https://doi.org/10.48550/ARXIV.2302.07658

Gomon, D., Putter, H., Nelissen, R.G., Van Der Pas, S.: Cgr-cusum: a continuous time generalized rapid response cumulative sum chart. Biostatistics (2022). https://doi.org/10.1093/biostatistics/kxac041

Griffen, D., Callahan, C.D., Markwell, S., de la Cruz, J., Milbrandt, J.C., Harvey, T.: Application of statistical process control to physician-specific emergency department patient satisfaction scores: A novel use of the funnel plot. Acad. Emerg. Med. **19**(3), 348–355 (2012). https://doi.org/10.1111/j.1553-2712.2012.01304.x

Grigg, O.A.: The STRAND chart: a survival time control chart. Stat. Med. **38**, 1651–1661 (2018). https://doi.org/10.1002/sim.8065

Grigg, O., Farewell, V.: An overview of risk-adjusted charts. J. R. Stat. Soc. Ser. A Stat. Soc. **167**(3), 523–539 (2004). https://doi.org/10.1111/j.1467-985x.2004.0apm2.x

Grigg, O.A., Farewell, V.T., Spiegelhalter, D.J.: Use of risk-adjusted cusum and rsprtcharts for monitoring in medical contexts. Stat. Methods Med. Res. **12**(2), 147–170 (2003). https://doi.org/10.1177/096228020301200205

Jiang, W., Shu, L., Zhao, H., Tsui, K.-L.: Cusum procedures for health care surveillance. Qual. Reliab. Eng. Int. **29**(6), 883–897 (2012). https://doi.org/10.1002/qre.1444

Kolfschoten, N.E., Marang van de Mheen, P.J., Gooiker, G.A., Eddes, E.H., Kievit, J., Tollenaar, R.A.E.M., Wouters, M.W.J.M.: Variation in case-mix between hospitals treating colorectal cancer patients in the netherlands. Eur. J. Surg. Oncol. **37**(11), 956–963 (2011). https://doi.org/10.1016/j.ejso.2011.08.137

Mahmoud, M.A., Woodall, W.H., Davis, R.E.: Performance comparison of some likelihood ratio-based statistical surveillance methods. J. Appl. Stat. **35**(7), 783–798 (2008). https://doi.org/10.1080/02664760802005878

Page, E.S.: Continuous inspection schemes. Biometrika **41**, 100–115 (1954). https://doi.org/10.2307/2333009

Sego, L.H., Reynolds, M.R., Woodall, W.H.: Risk-adjusted monitoring of survival times. Stat. Med. **28**, 1386–1401 (2009). https://doi.org/10.1002/sim.3546

Sievert, C.: Interactive Web-Based Data Visualization with R, Plotly, and Shiny. Chapman and Hall/CRC, Boca Raton (2020)

Spiegelhalter, D.J.: Funnel plots for comparing institutional performance. Stat. Med. **24**, 1185–1202 (2005). https://doi.org/10.1002/sim.1970

Steiner, S.H., Jones, M.: Risk-adjusted survival time monitoring with an updating exponentially weighted moving average (EWMA) control chart. Stat. Med. **29**, 444–454 (2009). https://doi.org/10.1002/sim.3788

Steiner, S.H., Cook, R.J., Farewell, V.T., Treasure, T.: Monitoring surgical performance using risk-adjusted cumulative sum charts. Biostatistics **1**(4), 441–452 (2000). https://doi.org/10.1093/biostatistics/1.4.441

Therneau, T.M.: A Package for Survival Analysis in R. (2020). R package version 3.2-7 https://CRAN.R-project.org/package=survival

Tighkhorshid, E., Amiri, A., Amirkhani, F.: A risk-adjusted ewma chart with dynamic probability control limits for monitoring survival time. Commun. Stat. Simul. Comput. **51**(3), 1333–1354 (2019). https://doi.org/10.1080/03610918.2019.1667393

Tsui, K.-L., Chiu, W., Gierlich, P., Goldsman, D., Liu, X., Maschek, T.: A review of healthcare, public health, and syndromic surveillance. Qual Eng. **20**(4), 435–450 (2008). https://doi.org/10.1080/08982110802334138

Tsui, K.-L., Wong, S.Y., Jiang, W., Lin, C.-J.: Recent research and developments in temporal and spatiotemporal surveillance for public health. IEEE Trans. Reliab. **60**(1), 49–58 (2011). https://doi.org/10.1109/tr.2010.2104192

Van Leersum, N.J., Snijders, H.S., Henneman, D., Kolfschoten, N.E., Gooiker, G.A., ten Berge, M.G., Eddes, E.H., Wouters, M.W.J.M., Tollenaar, R.A.E.M., Bemelman, W.A., van Dam, R.M., Elferink, M.A., Karsten, T.M., van Krieken, J.H.J.M., Lemmens, V.E.P.P., Rutten, H.J.T., Manusama, E.R., van de Velde, C.J.H., Meijerink, W.J.H.J., Wiggers, T., van der Harst, E., Dekker, J.W.T., Boerma, D.: The dutch surgical colorectal audit. Eur. J. Surg. Oncol. **39**(10), 1063–1070 (2013). https://doi.org/10.1016/j.ejso.2013.05.008

Verburg, I.W., Holman, R., Peek, N., Abu-Hanna, A., de Keizer, N.F.: Guidelines on constructing funnel plots for quality indicators: a case study on mortality in intensive care unit patients. Stat. Methods Med. Res. **27**(11), 3350–3366 (2017). https://doi.org/10.1177/0962280217700169

Warps, A.K., Detering, R., Tollenaar, R.A.E.M., Tanis, P.J., Dekker, J.W.T.: Textbook outcome after rectal cancer surgery as a composite measure for quality of care: a population-based study. Eur. J. Surg. Oncol. **47**(11), 2821–2829 (2021). https://doi.org/10.1016/j.ejso.2021.05.045

Willik, E.M., Zwet, E.W., Hoekstra, T., Ittersum, F.J., Hemmelder, M.H., Zoccali, C., Jager, K.J., Dekker, F.W., Meuleman, Y.: Funnel plots of patient-reported outcomes to evaluate health-care quality: Basic principles, pitfalls and considerations. Nephrology (Carlton) **26**(2), 95–104 (2020). https://doi.org/10.1111/nep.13761

Wood, S.N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. **73**(1), 3–36 (2011). https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S.N.: Generalized Additive Models: An Introduction with R. CRC Press/Taylor & Francis Group, Boca Raton (2017)

Woodall, W.H.: The use of control charts in health-care and public-health surveillance. J. Qual. Technol. **38**(2), 89–104 (2006). https://doi.org/10.1080/00224065.2006.11918593

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Daniel Gomon[1] · Julie Sijmons[2,3] · Hein Putter[1,4] · Jan Willem Dekker[5] · Rob Tollenaar[2,7] · Michel Wouters[6,7] · Pieter Tanis[3,8] · Marta Fiocco[1,4] · Mirko Signorelli[1]

✉ Daniel Gomon
d.gomon@math.leidenuniv.nl

[1] Mathematical Institute, Leiden University, Leiden, The Netherlands

[2] Dutch Institute for Clinical Auditing, Leiden, The Netherlands

[3] Department of Surgery, Amsterdam University Medical Centre, Amsterdam, The Netherlands

[4] Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, The Netherlands

[5] Department of Surgery, Reinier de Graaf Groep, Delft, The Netherlands

[6] Department of Surgery, Dutch Cancer Institute, Amsterdam, The Netherlands

[7] Department of Surgery, Leiden University Medical Centre, Leiden, The Netherlands

[8] Department of Surgical Oncology and Gastrointestinal Surgery, Erasmus MC, Rotterdam, The Netherlands