Statistical Process Monitoring to Improve Quality Assurance of Inpatient Care

Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.)
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von
Lena Hubig

aus
Engen im Hegau

am
05. März 2020

Supervisor:	Prof. Dr. Ulrich Mansmann
Second expert:	Prof. Dr. Ralf Holle
Dean:	Prof. Dr. med. dent. Reinhard Hickel
Date of oral defense:	21 September 2020

Contents

List of Figures

List of Tables

List of Abbreviations

ANOS	Average Number of Observations to Signal
ARL	Average Run Length
BAQ	Bavarian Agency for Quality Assurance (*German: Bayerische Arbeitsgemeinschaft für Qualitätssicherung*)
CRAM	Cumulative Risk-Adjusted Mortality
CRAN	Comprehensive R Archive Network
CUSUM	Cumulative Sum
DPCL	Dynamic Probability Control Limit
EQA	External Quality Assurance
EWMA	Exponentially Weighted Moving Average
FDR	False Discovery Rate
FSP	False Signal Probability
G-BA	Federal Joint Committee (*German: Gemeinsamer Bundesausschuss*)
GSCUSUM	Group Sequential Cumulative Sum
IC-ARL	In-Control Average Run Length
IHV	Indicator specific Hospital Volume
IQTIG	Federal Institute for Quality Assurance and Transparency in Health Care (*German: Institut for Qualitätssicherung und Transparenz im Gesundheitswesen*)
QSKH-RL	Directive on Measures concerning the Quality Assurance in Hospitals (*German: Richtlinie über Maßnahmen der Qualitätssicherung in Krankenhäusern*)
RA-CUSUM	Risk-Adjusted Cumulative Sum
RA-GSCUSUM	Risk-adjusted Group Sequential Cumulative Sum
SGB	Social Security Code (*German: Sozialgesetzbuch*)
SHI	Statutory Health Insurance (*German: Gesetzliche Krankenversicherung, GKV*)
SPC	Statistical Process Control
SPM	Statistical Process Monitoring
SPRT	Sequential Probability Ratio Test
ST-CUSUM	Standard Cumulative Sum
TSP	True Signal Probability
VLAD	Variable Life Adjusted Display

List of Performance Indicators

11724 **Carotid Stenosis Surgery:** Ratio of observed to expected cases or severe stroke or death under open surgery. (risk-adjusted)

51828 **Neonatology:** Surgically treated necrotizing enterocolitis in small premature infants

54030 **Trauma Surgery**: Preoperative stay over 24 hours for patients with proximal femur fracture

Abstract

Quality assurance in German hospitals has recently become a focus of policy makers and practicioners alike, though its methods based on annual averages still lack in terms of timeliness, accuracy, reliability and evaluability. When monitoring health care performance, Statistical Process Monitoring (SPM) tools have been widely applied to detect quality shifts. However, the use of SPM in German external quality assurance (EQA) is not straightforward, as the monitored hospitals and processes differ greatly and data collection is based not on a sequential, but quarterly rhythm.

This thesis first recapitulates the use of Bernoulli log-likelihood CUSUM charts. It then introduces the construction of CUSUM charts for a predefined false signal probability and evaluates the signalling characteristics of CUSUM charts for different monitoring schemes and process scenarios within the framework of German EQA. This first part explains the influence of case risk mix, hospital volume, baseline failure probability and risk-adjustment on the construction and performance of CUSUM charts, and demonstrates the application of CUSUM charts for fair performance evaluation of inpatient care.

Second, it introduces an extension to traditional CUSUM charts, the Group Sequential CUSUM (GSCUSUM) chart. SPM methods rely on a regular and accurate data collection, which is unrealistic for most hospital settings. The extension enables the use of SPM methods when only aggregated binary performance over irregular time periods and of irregular length are available. A simulation study proves that the GSCUSUM chart is a good approximation for the standard CUSUM chart, and is equivalent to the standard CUSUM chart when the full sequence is observed.

Finally, the CUSUM and GSCUSUM charts are applied to hospital performance data and compared to traditional evaluation methods of the EQA. We find that control charts support the interpretation of data to find performance changes in a much more clearer way. Nevertheless, good data documentation is still of great importance.

Methods of SPM can be a valuable extension to standard performance evaluation in German hospitals. Areas for which control charts may be first implemented could be worst-case processes of high failure probability and the evaluation of process interventions.

1 Introduction

Experts believe that every year 400,000-800,000 patients experience a negligent adverse event in German hospitals and 20,000 patients die a negligent death.[1] Few of these cases are attributable to individuals like Harold Shipman in the UK or Niels Högel in Germany, who were convicted for murdering patients under their care. More frequently, causes for negligent events are overworked and understaffed health care providers and systematic process failures.[2]

Monitoring the performance of health care providers and while doing so also identifying quality deficits is tremendously important to avoid morbidity and mortality inflicted by the caring profession. In Germany, the first monitoring of inpatient care was introduced inBavaria in the 1970s in form of perinatal registries. The monitoring of care was made mandatory in the 1990s, but remained self-regulated until the 2000s.[3] The focus of these evaluations was the comparison among peers, and results were not shared with the public. In the mid-2000s, the Federal Joint Committee (G-BA), which is the highest decision-making body in the German Statutory Health Insurance (SHI) system, took over responsibility and is since coordinating quality assurance measures. Quality assurance is based on two procedures: an external quality assurance (EQA) and an internal quality management. Both of these measures are mandatory for all inpatient and outpatient health care providers treating patients under SHI.

In 2014, the next step was taken by the legislator to make the monitoring of quality of care more transparent and rigorous by instructing the G-BA to found an independent, scientific institute to conduct the quality assurance procedure, the Federal Institute for Quality Assurance and Transparency in Healthcare (IQTIG). Two years later, the Law reforming the Structures of Hospital Care (Krankenhausstrukturgesetz - KHSG) was introduced.[4] For the first time, this law allows the use of quality of care as a decisive factor in hospital planning and hospital financing. Hospital departments, which do not provide acceptable quality of care, can be shut down and quality deductions or quality supplements can be used as punishments or stipulates to provide exceptional care.

All these changes demand a reliable and robust method to evaluate hospital performance and to identify and signal quality deficits.

To this end, Statistical Process Monitoring (SPM) is a method worth investigating. It was first introduced in the 1930s for the monitoring of industrial production processes,[5*] but has since been adapted and applied to different health care settings. Tools of SPM have been used to monitor the performance of individual physicians[8–15] and health care providers,[16–19] assess the learning curve of trainees,[20–24] monitor infectious diseases,[25–29] and to manage diseases in individuals.[30,31]

Being used and enhanced across all different disciplines and areas of research, SPM instruments can be quite sophisticated and are a flexible tool to monitor the performance of processes. Given the right circumstances, they return exact and timely feedback on the quality of a process. The main objective of SPM is to quickly detect process changes by classifying process performance as *in-control* or *out-of-control*, using the graphical representation of control charts. When the process is in-control, the performance varies naturally around an accepted failure probability, which is called *stable system of chance causes* in the framework of SPM.[32] These chance causes of variability are unavoidable and part of the natural process. On the other hand, if other causes of variability exists that are assignable and special, the process is deemed to be out-of-control. In industrial production control, these causes may be malfunctioning machines, human error or defective materials – in a hospital similar causes are imaginable.

This thesis assesses the application of SPM methods to external quality assurance of inpatient care in Germany and gives guidance on its application. While there have been some projects of introducing SPM to monitor large health care settings, German EQA poses unique challenges that are addressed in the main part of this thesis:

1. The EQA setting in Germany is very diverse. 205 performance indicators of 28 medical faculties were monitored in 2018. EQA is mandatory for all German hospitals treating patients covered by the SHI, which vary vastly in case risk mix and patient numbers. A method eligible to be used in EQA should be applicable to most processes and be equally able to classify between acceptable and unacceptable performance for fair performance evaluation. Questions arising from these specific needs are addressed in Chapter 3. Parts of this chapter have been published in our first paper.[33]
2. The data quality of German EQA is not optimal. The only available date value is the date of documentation. Performance data are documented by the treating physicians, who often document multiple procedures at the same time. Thus, the inherent sequence of events is lost and simple SPM methods fail, as they

*When introduced by Shewhart, he called the method *Statistical Process Control (SPC)*, which is still the more frequently used term. It is however mostly associated with simple control chart techniques, and implies that some sort of control action is taking place. Woodall and others proposed to use the term *monitoring* instead of *control*,[6,7] which appropriatly reflects tools for continuous monitoring of a process.

rely on regular data transmission in order to guarantee consistent control chart performance. Unfortunately, a better date variable (date of admission/discharge) is not available due to data protection laws. An extension to standard SPM methods is presented in Chapter 4, which deals with this kind of group-sequential data.

Chapter 2 describes the EQA procedure as regulated by §136 SGB V and lays out important shortcomings. Furthermore, it introduces the motivating example and the data set that is the basis for all analyses in this thesis.

Chapter 5 compares all presented methods of SPM to the standard performance evaluation of EQA. This comparison shows factors influencing the signalling of a performance deficit and points out the importance of accurate data documentation.

Accompanying this thesis, software for constructing and evaluating CUSUM charts are available as an open source R package on the Comprehensive R Archive Network (CRAN)* and development versions are available on *github*.† This package has been downloaded from CRAN more than 8000 times and we have received user reports from practicioners around the world. In the Appendix, two instructional vignettes on how to use the package to calculate and evaluate CUSUM control charts are presented (Appendix 1). Furthermore, an R shiny app is provided that illustrates the use of CUSUM charts.‡

*https://CRAN.R-project.org/package=cusum
†https://github.com/lhubig/cusum
‡https://shiny.lenahubig.de/cusum

2 External quality assurance of inpatient care

2.1 External quality assurance procedure

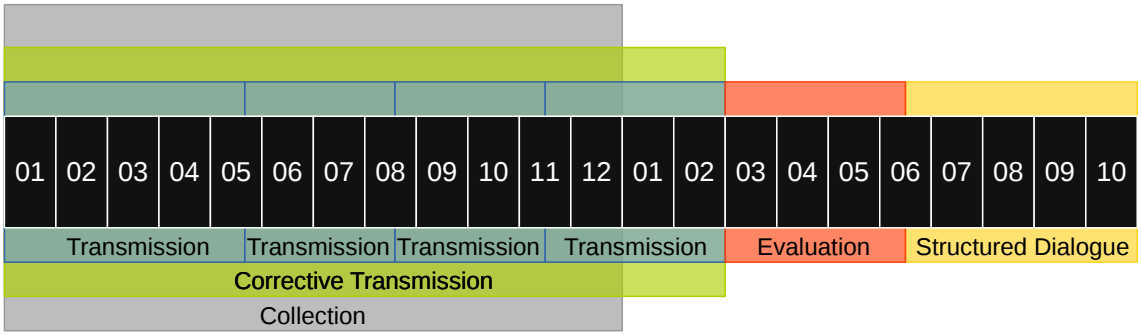


Fig. 2.1. Timeline of quality assurance process following QSKH-RL.[34] The reporting period, for which data is collected, covers one year (grey box). Data is transmitted towards the end of each quarter before 15 May, 15 August, 15 November, or 28 February (dark green boxes), and corrective data transmissions (light green box) are possible throughout the year. Data evaluation (red box) starts when the complete annual data is available, followed by the Structured Dialogue (yellow box) if a quality deficit is suspected.

External quality assurance (EQA) of inpatient care is regulated by the Directive on Measures concerning the Quality Assurance in Hospitals (QSKH-RL).[34] According to the directive, each patient's treatment is documented based on a set of nationally standardised performance indicators for selected interventions. Performance data of the previous quarter are submitted to the central agency, the IQTIG, and to the corresponding state offices by 15 May, 15 August, 15 November and 28 February, and corrective data transmissions from all quarters are accepted until 28 February (Figure 2.1).* Most of the performance data are submitted to the corresponding state offices, who then transmit the data to the IQTIG (indirect procedure). Performance data for

*This schedule was introduced for the reporting year of 2019. Previously, all performance data had to be submitted annually by 28 February, with the possibility to submit data throughout the year for interim analyses and data checks.

few interventions, which have small case numbers and few reporting hospitals, are submitted directly to the IQTIG (direct procedure). The whole process is accompanied via a data validation process to ensure the correct and complete documentation of patients' treatments.

Even though the directive requires a quarterly data transmission, the main analysis is the annual evaluation of hospital performance data. Until 15 June of the following year, hospitals must receive a performance evaluation, comparing their results to that of their peers. Arithmetic deviations are defined for each performance indicator, where a target range is set by the IQTIG.* Target ranges can either be defined by a fixed value (fixed reference range), or by the distribution of the results of all providers (percentile reference range). If the annual aggregated failure rate of a hospital and performance indicator is in the target range, it is considered acceptable performance. The signalling of an arithmetic deviation does not consider random error due to hospital volume.

If a deviation is detected, a so-called *Structured Dialogue* must be initiated. Here, hospitals must provide a statement on the suspected cause of deviation. If the explanation for deviation is not compelling, further interventions can be considered. These range from meetings of experts with hospital representives, over audits of the affected departments, to target agreements.

2.2 Challenges of external quality assurance

Quality in German hospitals as measured by the EQA has continuously improved over the last years, which was the result of analyses carried out by Rückle and Stausberg,[35] and Lack and Gerhardinger.[36] Still, there is room for improvement, as the current EQA procedure is lacking in four areas: timeliness, accuracy, statistical reliability and evaluability.

Timeliness Although since 2019 the data ought to be transmitted quarterly and quarterly results are provided to the hospitals, the main analysis and the basis for intervention is still the annual evaluation. As a consequence, over a year may pass before quality deficits are investigated and interventions are considered.

Accuracy By only evaluating aggregated performance data, more intricate performance changes are missed. Trends, seasonal effects or general runs of conflicting performance are masked by the average. Interventions may benefit from identifying different patterns of performance changes, and targeted actions may help prevent further deviations in following years.

* *German: Referenzbereich*

Statistical reliability Hospital volume is not automatically considered when assessing hospital performance. The state offices and IQTIG provide additional statistical analyses, which include confidence intervals for hospital results and enables the additional evaluation of a *statistical deviation*. This evaluation is not mandatory and not the decisive factor for a quality deficit. Furthermore, deviations that are the result of one single event are excluded from further analyses and interventions.

Evaluability It is currently not possible to evaluate interventions over short periods of time within the EQA framework. As a result, it is also not possible to attribute positive long-term trends to particular interventions. Possible process improvements can only be noticed when the data of the following year is analysed, and then it is difficult to link the intervention to the positive change.

2.3 Motivating example

To illustrate the use of SPM, all methods were applied to real performance data from Bavarian hospitals over the period of 2016–2017, made available by the Bavarian Agency for Quality Assurance (BAQ). Three performance indicators were chosen as examples to test and evaluate all methods (Table 2.1). The indicators were developed by the IQTIG, and the exact specifications and algorithms are published on the website of the IQTIG.[37]

Indicator 11724 is risk-adjusted and monitors in-hospital complication or death after open carotid stenosis surgery. The risk model is estimated by the IQTIG and updated annually.[38] For 2016, the explanatory variables were given as: age, indication group, preoperative degree of disability, and ASA classification. The hospital result is calculated as the ratio of numbers of observed cases to numbers of expected cases. The patient individual risk for complications or death ranged between 0.24% to 40.98%, with a median of 0.83% across 2016 and 2017.

Indicator 51838 monitors the cases of surgically treated necrotizing enterocolitis in small premature infants, a serious intestinal infection often leading to death.[39] This indicator was chosen because of its low failure rate: Only 1.07% cases were recorded in 2016 in Bavarian hospitals.

Indicator 54030 measures the cases of extended preoperative stay of patients with proximal femur fracture, which repeatedly has a high failure probability (20.35% in 2016). Rapid surgery within 24 hours may prevent severe complications such as thrombosis, pulmonary embolism or pressure ulcers.[40]

Figure 2.2 and Figure 2.3 break down the individual hospital results for these indicators in 2016 and 2017 by hospital volume, i.e. number of patients to which the respective

Table 2.1: Performance indicators selected to illustrate and evaluate control chart performance.

Number	**Risk-adjusted**	**Description**	**Target range** (2016)
11724	Yes	**Carotid Stenosis Surgery**: Ratio of observed to expected cases of severe stroke or death under open surgery	≤ 4.58
51838	No	**Neonatology**: Surgically treated necrotizing enterocolitis in small premature infants	not determined
54030	No	**Trauma surgery**: Preoperative stay over 24 hours for patients with proximal femur fracture	$\leq 15\%$

indicator is applicable. For indicator 11724 and 51838, many hospitals reported no cases, and even large hospitals did not record a single case in a year. The figures also show that the majority of hospital had less than 100 patients per year: In 2016 and 2017 hospitals recorded on average 50 and 48 patients per performance indicator respectively.

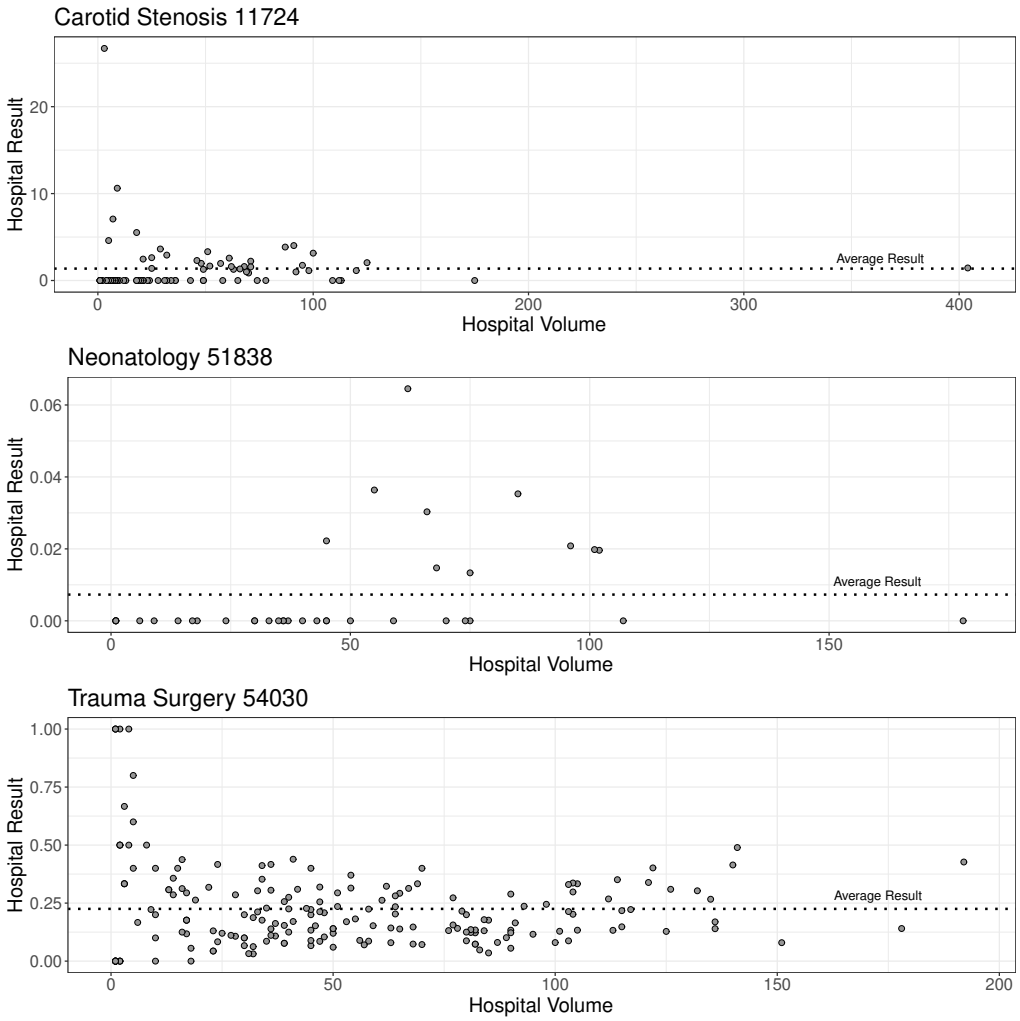


Fig. 2.2. Annual hospital results of the selected performance indicators displayed by hospital volume in Bavaria, 2016.

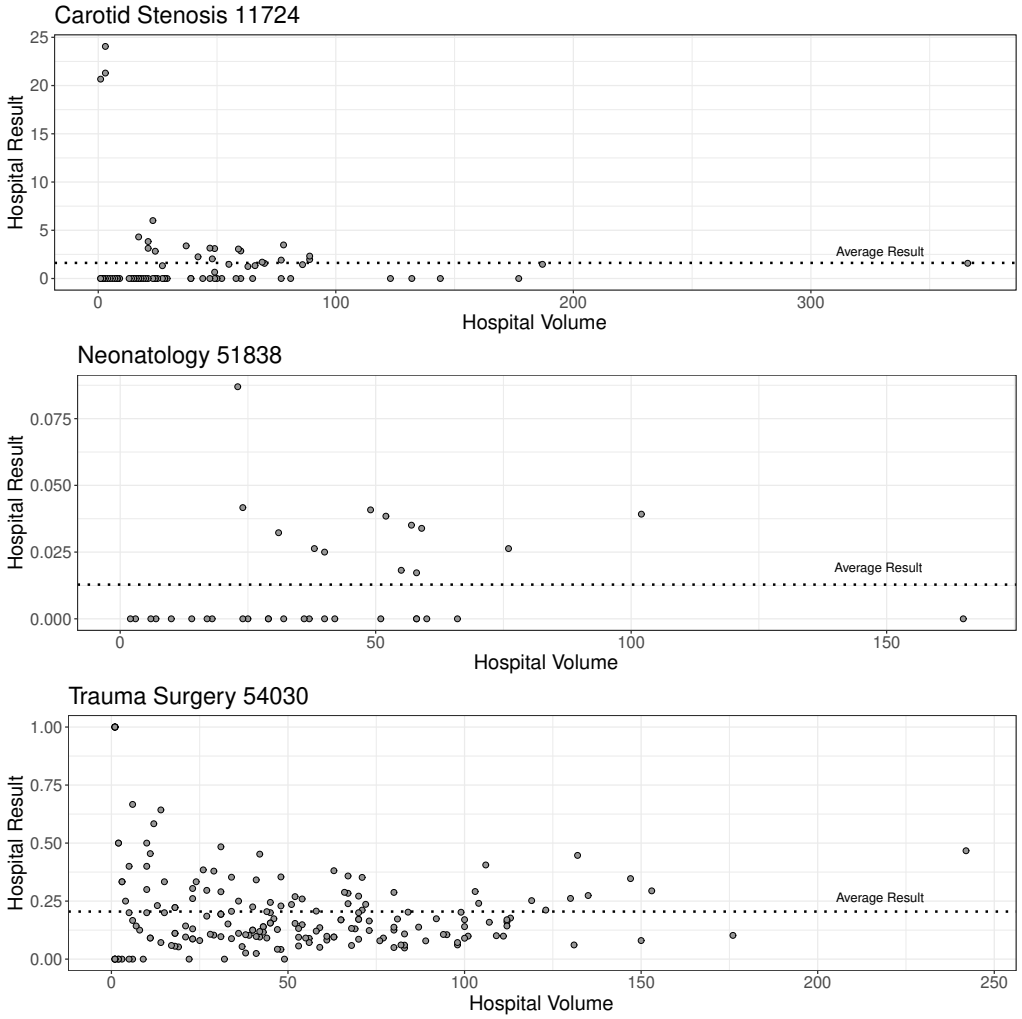


Fig. 2.3. Annual hospital results of the selected performance indicators displayed by hospital volume in Bavaria, 2017.

3 Cumulative Sum Chart

First introduced in 1954 by Page,[41] Cumulative Sum (CUSUM) charts are one of the most popular and at the same time sophisticated control charts.[42–45] As there remains confusion about the correct definition of CUSUM charts, this chapter starts with a short differentiation between the Sequential Probability Ratio Test (SPRT), the Variable Life Adjusted Display (VLAD), and subsequently the CUSUM.

Most frequently, CUSUM charts are confused with the **Sequential Probability Ratio Test (SPRT)** introduced by Spiegelhalter et al.[46] and Grigg et al.,[47] which is also known as the CUSUM log-likelihood ratio test. It looks similar to the CUSUM chart, as in that it consists of two horizontal thresholds, but it is not restricted to non-negative or non-positive values. The SPRT is a sequential hypothesis test, where the limits define a threshold to reject the null hypothesis, and as long as the observations are inside the limits, the hypothesis test is continued. Crossing a threshold signals the acceptance of the alternative hypothesis, which reflects a deviating performance. Woodall et al.[48] specifically discourage the use of the SPRT and the risk-adjusted SPRT, the RSPRT. As these charts are building up credit during periods of good performance, they are less likely to detect process deteriorations.

The CUSUM is also frequently confused with the **Variable Life Adjusted Display (VLAD)**, proposed by Lovegrove et al.,[49] or Cumulative Risk-Adjusted Mortality (CRAM), by Poloniecki et al.,[50] that is also known as the E–O CUSUM. The VLAD monitors the difference between the expected and observed events. Its main disadvantage is the difficulty of setting control limits. In order to detect a performance change, control limits of the VLAD have to widen over the monitoring period, due to an increase in variance with greater sample size. Different methods have been proposed to construct control limits, e.g. the Rocket Tail plot by Sherlaw-Johnson,[51] or the updated V-mask by Wittenberg et al.[52] Additionally, like the SPRT, VLAD charts are prone to build up credit, as they are not restricted to one side of zero. Because of this, the exclusive use of VLAD charts is generally discouraged, and it is recommended to show VLAD charts and generate signals via CUSUM statistics,[53] as VLAD charts are more easy to interpret than CUSUM charts.

We use the **Tabular CUSUM** chart as described by Montgomery in his introduction to SPM and CUSUM charts,[32] and by Steiner et al.[54] Tabular CUSUM charts consist

of the continuous monitoring of two CUSUM statistics, the upper C detecting process deteriorations and the lower Z signalling process improvements, both of which are restricted to one side of zero, enabling faster signalling in case of a process change as no credit is building up.

We decided to evaluate and propose the tabular CUSUM chart for binary events due to the following reasons:

- The CUSUM chart is optimal in detecting process changes.[53,55,56]
- The CUSUM chart is very flexible regarding different monitoring scenarios. It is possible to construct CUSUM charts for different performance indicators, including risk-adjusted and non-risk-adjusted, with different failure probabilities.[54]
- While the calculation and background of CUSUM charts may not be as simple as the VLAD chart, interpretation of a signal is straightforward.
- With the CUSUM chart it is possible to detect process deteriorations as well as process improvements. For now, only the CUSUM chart restricted to non-negative values that is detecting process deteriorations is of interest, as current EQA is focused on detecting quality deficits. Future application may of course benefit from this feature of the CUSUM.
- CUSUM charts are quite popular and well researched. They are not a niche control chart and there is a rich literature on applications and enhancements (see also Section 3.1.4)

This Chapter introduces the general framework of CUSUM charts, its construction, influencing factors and potential enhancements in Section 3.1. Section 3.2 presents a simulation study which evaluates chart performance in form of a power analysis, and Section 3.3 applies the presented methods to hospital performance data. Large parts of this chapter (Secs. 3.1 through 3.5 with the exception of 3.1.4) have been published in our first paper.[33]

3.1 Construction of CUSUM charts

3.1.1 Definition and graphical representation

CUSUM charts for monitoring process performance for a deterioration in quality over time are defined as:[41]

$$C_t = \max(0, C_{t-1} + W_t), \quad t = 1, 2, 3, ... \tag{3.1}$$

The dichotomous outcome of observation y equals 0 for every success and 1 for every adverse event. Observations are plotted in sequence of their temporal occurrence. Depending on the outcome, the CUSUM decreases or remains at zero for every success, and increases for every adverse event. The magnitudes of increase and decrease are denoted by CUSUM weights W_t. Following Steiner et al. the weights W_t for the Standard CUSUM (ST-CUSUM) are:[54]

$$W_t = \begin{cases} \log\left(\frac{1-c_A}{1-c_0}\right) & \text{if } y_t = 0 \\ \log\left(\frac{c_A}{c_0}\right) & \text{if } y_t = 1 \end{cases}, \tag{3.2}$$

where c_0 is the baseline failure probability and c_A the smallest unacceptable failure probability, which is the change in performance that is detected. CUSUM weights may be individualised for patient risk in the risk-adjusted CUSUM (RA-CUSUM). Here, the weights are:[54]

$$W_t = \begin{cases} \log\left(\frac{1}{1-p_t+R_A p_t}\right) & \text{if } y_t = 0 \\ \log\left(\frac{R_A}{1-p_t+R_A p_t}\right) & \text{if } y_t = 1 \end{cases}, \tag{3.3}$$

where p_t represents the individual patient risk score. The baseline failure probability is no longer constant, but tailored to patients' risk. The risk-adjusted CUSUM monitors for a change in risk specified by an odds ratio change from R_0 to R_A, with R_A greater than one indicating process deteriorations.

The hereafter ommitted CUSUM chart to monitor process improvements is constructed on the same prinicple. It differs from CUSUM charts detecting process deterioriations only by being mirrored around zero, and hence restricted to non-positive values: $Z_t = \min(0, Z_{t-1} + W_t)$.

The Tabular CUSUM is usually presented in a CUSUM chart as shown in Figure 3.1, where either both CUSUM statistics, C and Z, or only one, is plotted. CUSUM statistics are plotted in their sequence of observations, as well as the horizontal control limits, that signal a performance change.

3.1.2 Factors influencing CUSUM chart performance

Several factors influence the characteristics and performance of CUSUM charts. Some factors may be regarded as control switches of the monitoring schemes, as they are configurable and directly influence control charts. Other factors are mostly fixed by the

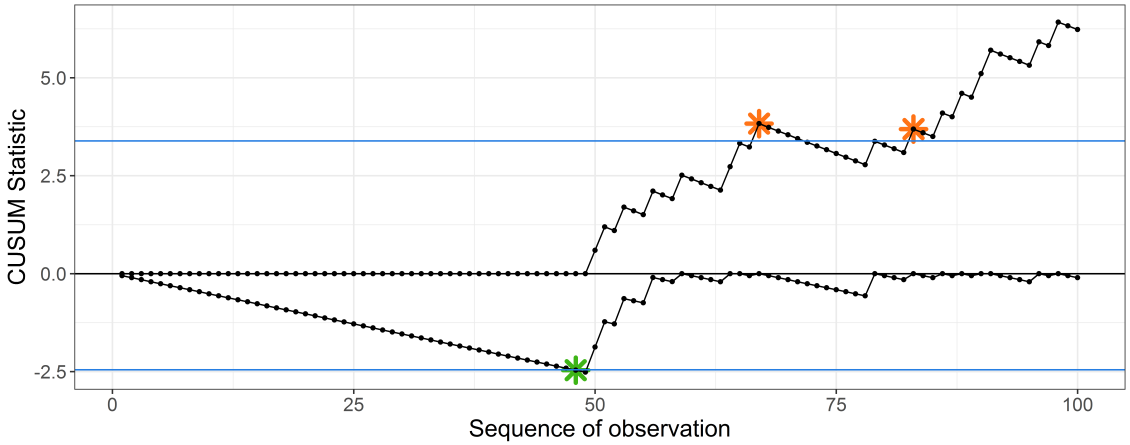


Fig. 3.1. Example of Tabular CUSUM chart for simulated data, where the first half of the process is in-control, and the second half of the process is out-of-control. The CUSUM statistic $C \geq 0$ signals negative performance changes (increase in failure rate). The lower CUSUM statistic $Z \leq 0$ signals positive performance changes (decrease in failure rate). Blue horizontal lines show the control limits, crossing of which generates a signal (orange/green stars).

process that is monitored. Most of these factors are also relevant when applying other types of performance monitoring or SPM. Additionally, other types of variations exist that may influence the performance of CUSUM charts, but they are not accounted for. These may be unknown or random factors that are not measured or difficult to quantify, e.g. the quality of the data.

Performance indicator: Performance indicators quantify a process output, indicating quality of care. For each performance indicator, the subset of patients covered by this indicator is specified. The performance indicator establishes the baseline failure probability c_0 or the risk-adjustment model for the patients' risk scores p_t. Additionally, the performance indicator should be considered when setting up a monitoring scheme due to the implications of the process at hand on detecting performance deteriorations.

Hospital Volume: Hospital volume is here defined as the annual number of patients per performance indicator and hospital. It is a major source of variation between hospitals and possibly also within hospitals across years, and it is considered for fair performance evaluation in the control limit simulation as the sample size n. As the hospital volume directly influences the control limit, it has a considerable effect on CUSUM performance.

Case risk mix: Adjusting for individual patient risk is necessary when comparing outcomes, but there is often some uncertainty about the validity of the risk adjustment model. When possible, previous experience of the process can be used to estimate the case risk distribution. The estimation of case risk mix is used in the simulation of the control limit, where outcome data is simulated on

the estimated risk population.

Detection level δ: Detectable changes in performance are determined by an odds ratio multiplier δ. In the ST-CUSUM, this change of δ defines the alternative failure probability c_A, which influences the CUSUM weights W_t in Eq. 3.2. For the RA-CUSUM, δ is equal to R_A in Eq. 3.3. Values of δ greater than one detect process deteriorations, while values less than one detect process improvements.

False Signal Probability: The False Signal Probability (FSP) is defined as the type 1 error of the CUSUM chart. It is the probability of a CUSUM signal within the monitoring of a process when the process is truly in control. Here it is applied as the defining parameter to construct CUSUM charts in the simulation of control limits.

3.1.3 Setting the CUSUM control limit

The CUSUM chart signals a performance change when the CUSUM statistic exceeds a pre-defined control limit. The process should then be investigated for quality deficits and monitoring can restart by resetting the current CUSUM statistic.[32]

Shewhart proposed setting control limits for his control chart, the $\bar{x}$-chart, to 3σ, where σ is the standard deviation of an in-control process. This method is still used in CUSUM chart,[19,57] but depends on exact estimation of previous performance.

Most commonly, control limits are constructed based on the Average Run Length (ARL),[12,58] which is the average time to first signal.* First, an appropriate in-control ARL (IC-ARL) is chosen, which is the ARL when the monitored process performance is as expected. Then, one estimates the IC-ARL of possible control limits and iteratively identifies the control limit that results in the desired IC-ARL. One can approximate the IC-ARL using Markov-Chain approximation, as proposed by Brook and Evans,[60] and improved by Knoth et al. for the RA-CUSUM.[61]

Setting the control limit based on a desired ARL has the following issues: The concept of ARL is difficult to understand and adapt for non-specialists of SPM. One has to decide how long a process is expected to run on average and deduce what kind of ARL is appropriate. If the ARL is used, there is no direct estimate of a false signal probability, a parameter essential for estimating the signalling characteristics of a control chart. The use of ARL is further hindered by the exponentially skewed distribution of control chart run lengths. This results in more short run lengths than expected and a higher probability for a false signal.

*Sometimes, one might also come across the Average Number of Observations to Signal (ANOS),[59] which is the appropriate term for ARL when observations are not taken regularly. The procedure for both methods is the same.

Instead of using the ARL, we decided to construct control limits based on FSP directly. Its interpretation is intuitive and clinicians are already familiar with the concept of false positive from diagnostic tests. The control limit is set by calculating possible CUSUM statistics for a process of a specific failure probability and estimating the probability for crossing a control limit.

As the signal probability approaches 100% with increasing run length, these parameters have to be estimated for a fixed sample size (n). For very small sample sizes, it is possible to estimate the exact FSP of possible control limits, by calculating all possible CUSUM paths. For larger sample sizes, we propose the following algorithm to select a control limit that will result in a specific FSP:[33]

1. Simulate a sufficiently large number of in-control sequential outcome data for $t = 1, 2, \ldots, n$, with baseline failure probability or, if applicable, individual risk probabilities drawn from the population.
2. Unrestricted CUSUM runs are calculated for these simulated sequences. This means the CUSUM charts do not include a control limit and are not reset.
3. The maximum CUSUM statistics (C_t) are collected from each CUSUM run.
4. The desired control limit for a sequence of size n is the $(1 - \text{FSP})$-percentile of the maximum CUSUM statistics.

3.1.4 CUSUM Chart enhancements

Let us at this point deviate from Ref. [33] to enumerate some additional CUSUM Chart enhancements, which may be of interest and could be applied to individualise or improve the proposed basic design. They range from simple extensions, which are already available, to more complex designs, which would need further research before implementing them in German EQA.

Fast Initial Response Scheme At the beginning of monitoring, the state of performance is unknown. Starting the CUSUM statistic at $C_0 > 0$ may result in fast signals if the process is out-of-control. If the process however is truly in control, the CUSUM statistics likely converge to zero.[62] The fast initial response results in more early signals, and results in a higher false signal probability. If used, a signal at the beginning of monitoring should be treated more carefully.

Supplementary Control Limits Supplementary signalling rules were already proposed in the Western Electric Handbook in the year 1956 for the Shewhart $\bar{x}$-chart, which are unnatural patterns of deviations:[63]

- One data points falls beyond the 3σ-control limit
- Two out of three consecutive points fall beyond the 2σ-control limit
- Four out of five consecutive points fall beyond the 1σ-control limit

Following these ideas, additional control limits may be added to the CUSUM chart, warning of the possibility of a process change. This provides the ability to react more quickly and prevent a true CUSUM signal. Additional signalling rules however increase the probability for a false signal, hence these signals should naturally be treated less seriously than signals resulting from the original CUSUM control limits.

Dynamic Probability Control Limits (DPCL) DPCL for the RA-CUSUM chart were introduced by Zhang et al. as a flexible alternative to constant control limits.[64] They are able to control the false signal probability during monitoring and adjust for case risk mix at every observation, which minimises the risk of false estimation of case risk mix from Phase I. They do however require constant calculation of risk mix and are complex to construct and interpret.

Multivariate Monitoring Monitoring several performance indicators simultaneously might beg the question if there is additional information gained by taking the relationship between indicators into account. Tang et al. proposed a method of SPM that allows the monitoring of more than two outcomes,[65] and a review on different multivariate SPM methods was provided by Bersimis et al.[66] These approaches might be interesting for performance indicators of the EQA, which can be summarised in a index, or which measure similar things, e.g. preoperative stay for femur fracture and preoperative stay for hip or knee endoprothesis.

Monitoring of Multiple Data Streams When monitoring multiple data streams, the risk of a false discovery increases. Previous work by Benjamini and Kling[67] and by Grigg and Spiegelhalter[68] introduced controlling the False Discovery Rate (FDR) by applying strategies from multiple testing to normally distributed data, and Mei proposed a scalable global monitoring scheme for concurrent data streams.[69] A method to control the FDR of Bernoulli log-likelihood CUSUM chart is currently not available.

3.2 Performance of CUSUM charts

3.2.1 Simulation design

Returning to Ref. [33], we simulated hospital performance data to assess the effect of various influencing factors on the False Signal Probability (FSP) and True Signal Probability (TSP) of ST-CUSUM and RA-CUSUM charts. Figure 3.2 illustrates how the described factor influence the construction and simulation of CUSUM charts.

CUSUM runs are simulated for the three previously described performance indicators from EQA. The baseline failure probabilities for the non-risk-adjusted performance indicators were set to the national overall average failure rate of 2016 and 2017

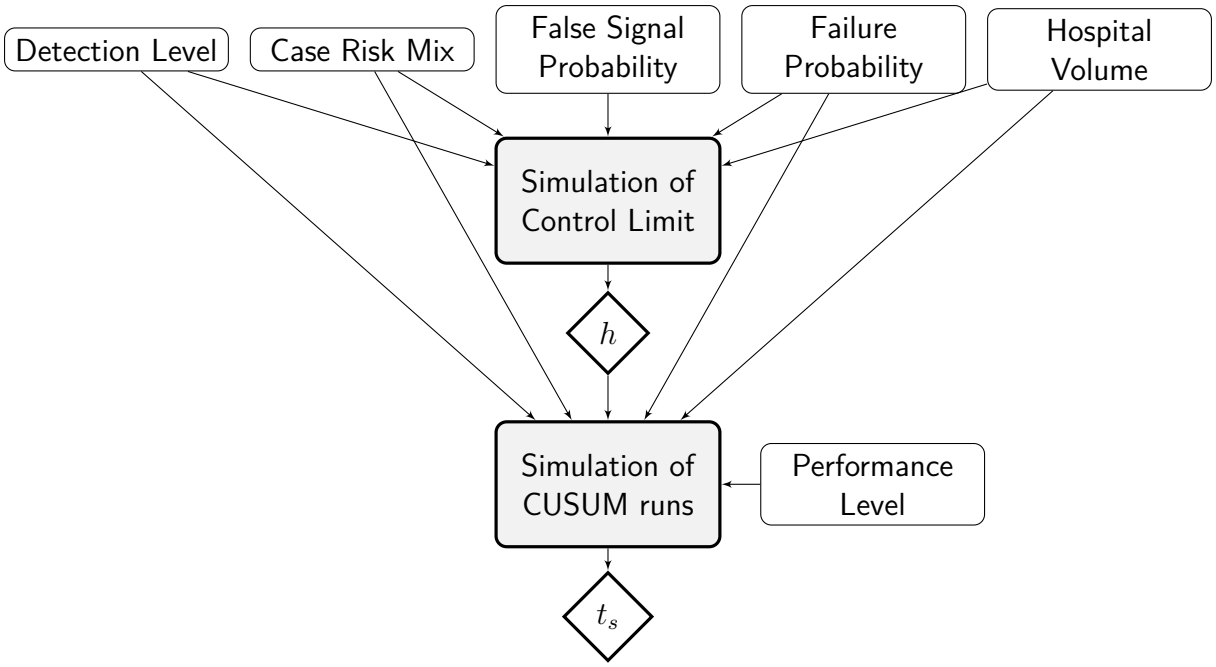


Fig. 3.2. Factors influencing simulation of control limits (h) and time to signal (t_s) in CUSUM runs simulation.

(51838: $c_0 = 1.25\%$; 54030: $c_0 = 19.21\%$). For the risk-adjusted indicator 11724, we re-sampled risk scores with replacement from the total hospital population of 2016 and 2017. Additionally, we created artificial subpopulations based on case risk mix. For a high risk population, risk scores were sampled from the risk population of the upper 25th percentile ($\geq 1.04\%$). A low risk population was considered with risk scores sampled from the risk population of the lower 25th percentile ($\leq 0.56\%$).

Three hospital volumes were derived for small, medium and large hospitals. The volume was estimated by taking the mean of the hospital volume percentiles across all performance indicators. The mean of hospitals below the 25th percentile ($n_s = 7$) was used as an estimate for small hospitals, the mean between the 25th and 75th percentile ($n_m = 42$) for medium hospitals, and the mean above the 75th percentile ($n_l = 105$) for large hospitals.

First, 100,000 in-control CUSUM runs were simulated to estimate control limits h (Figure 3.2). We simulated control limits for FSP of 0.1%, 0.5%, 1% and 5%, in accordance with typical values of type 1 error rates. The CUSUM was set to detect deteriorations with $\delta > 1$. The detection level of a doubling (2) of odds was considered as well as one step below (1.5) and one (2.5) and two (3) steps above.

In a second step, 2,000 in-control and out-of-control CUSUM runs were simulated to assess how well the specific CUSUM chart differentiates between good and poor performance. From these runs, we collected the run length to signal t_s, where the CUSUM statistic first exceeds the control limit. Finally, the signal rates are calculated

as the proportion of CUSUM runs, where the run length to signal was smaller than the hospital volume.

3.2.2 Simulation results

For every hospital volume, performance indicator and risk population, sixteen control charts were constructed for varying FSP and detection levels δ. Control limits were wider when FSP was small, detection level was high, baseline failure probability or case risk mix was high and hospital volume was large.

Figures 3.3 (a) and 3.4 (a) show the percentage of in-control CUSUM runs that signalled a process change as signal rates. Here, performance was as expected and signal rates should not exceed the predefined FSP of the control chart. Mostly, signal rates of in-control simulations were close to the desired FSP, demonstrating successful simulation of control limits. For two scenarios, the achieved in-control signal rate deviated from the desired FSP: For small hospital volumes of indicator 51838 (Figure 3.3 (a), bottom left), the CUSUM limit equals the CUSUM weight of an adverse event, which results in a higher false signal rate of $\approx 15\%$. For small hospital volume and low risk population of indicator 11724 (Figure 3.4 (a), top left), the control limit was set to zero, and the RA-CUSUM signalled at every observation. Hence, the in-control and out-of-control signal rates for this scenario were 100%. For these scenarios the exact estimation of FSP failed, because there are only finitely many possible CUSUM control limits due to the discrete nature of the CUSUM chart. When dealing with scenarios that require a careful estimation of tight control limits, it may be reasonable to chose a lower FSP and in turn also accept a lower TSP.

Signal rates for out-of-control CUSUM runs (Figures 3.3 (b), 3.4 (b)) represent the correctly identified deteriorations and ideally should be close to 100%. Large hospital volumes and higher failure probability resulted in a higher TSP. Control chart of indicator 54030 achieved 99.25% for the highest FSP and detection level (Figure 3.3 (b), top right). Yet, most CUSUM runs had smaller TSP; particularly CUSUM runs for small hospital volumes did not trigger a signal in the majority of CUSUM runs within one observation period.

3.3 Application of CUSUM charts

CUSUM charts with FSP-simulated control limits are applied to real data from EQA of inpatient care from the years 2016 and 2017 provided by the Bavarian Agency of Quality Assurance (BAQ).

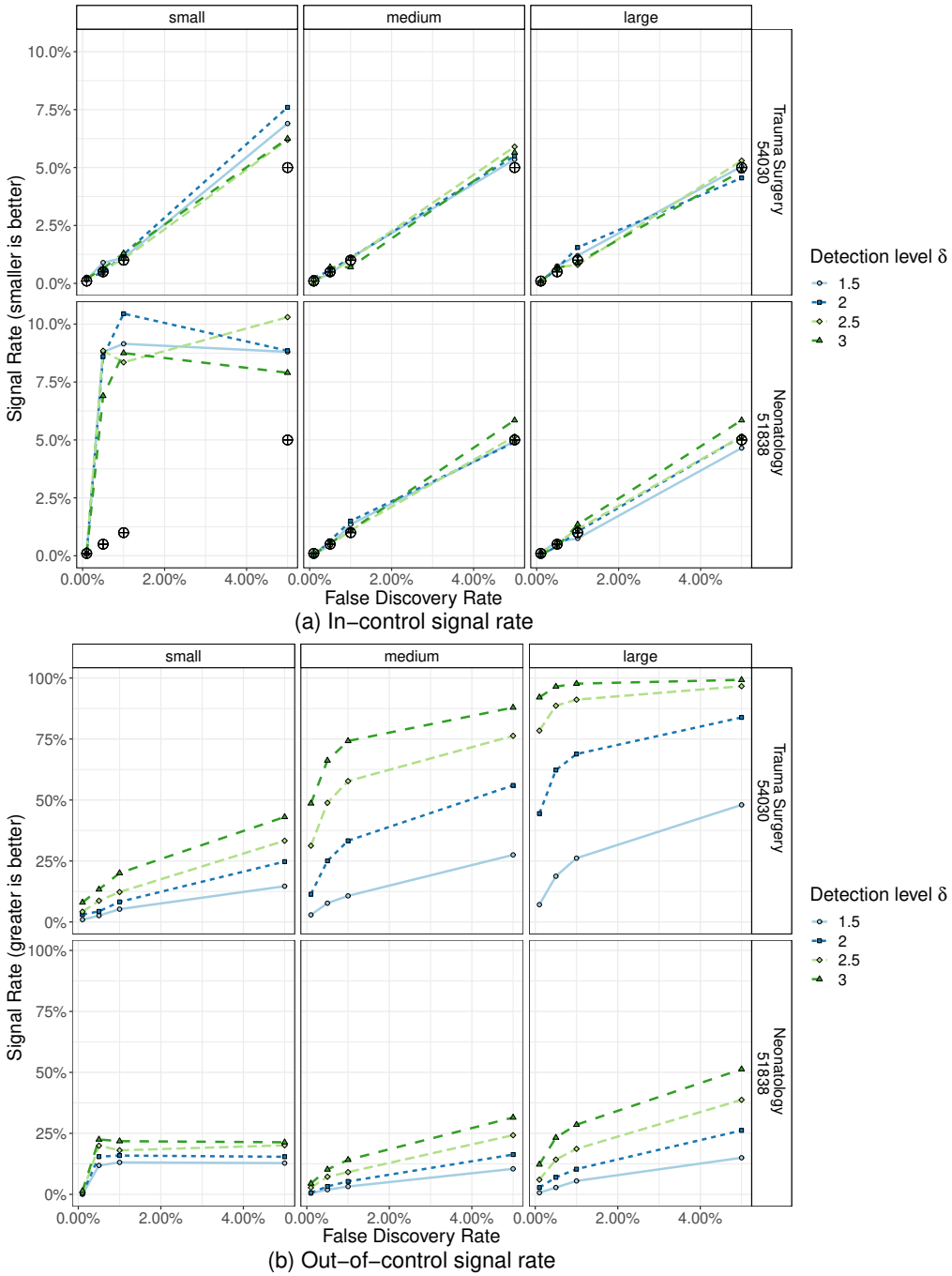
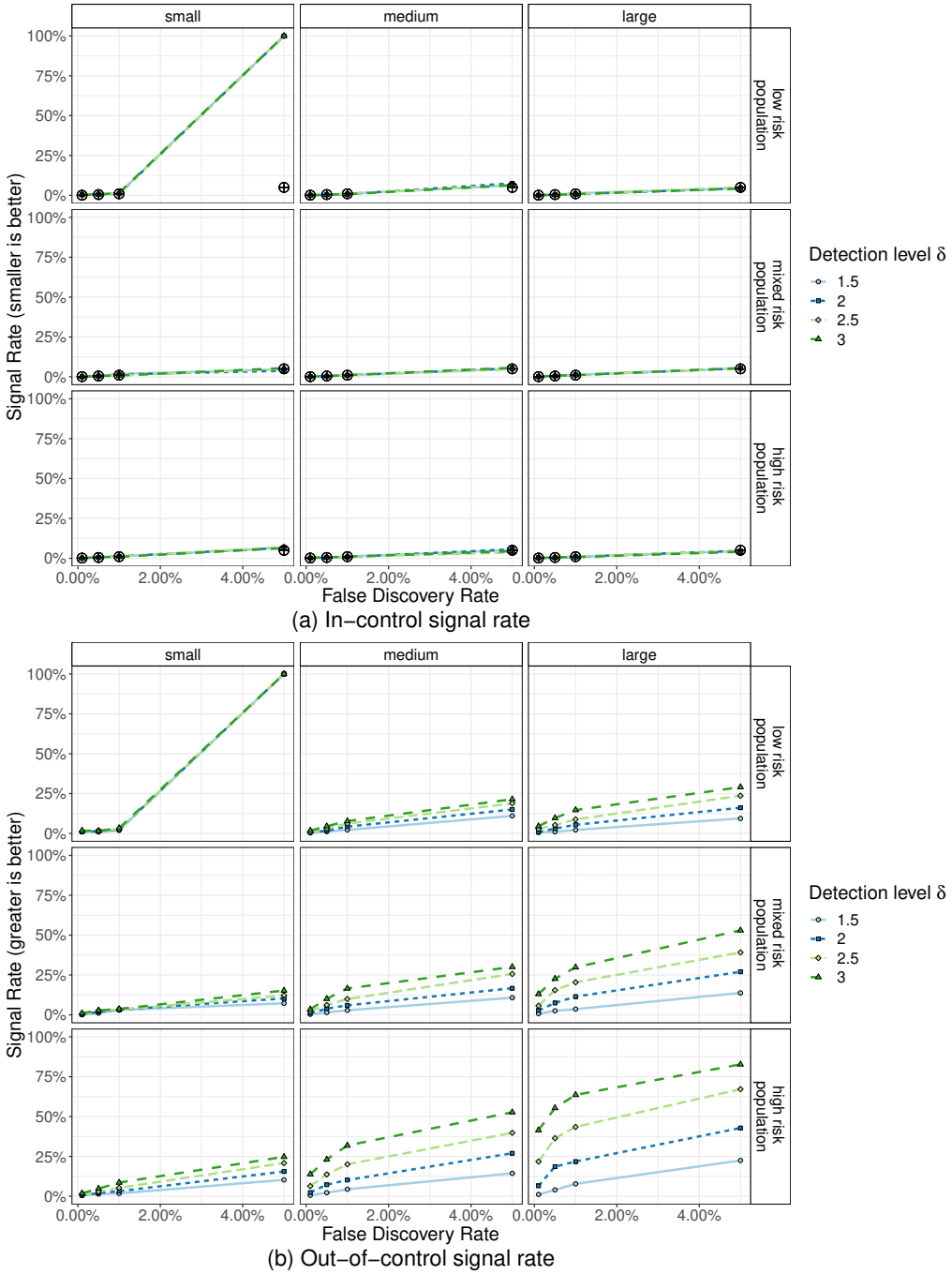


Fig. 3.3. Percentage of ST-CUSUM charts signalling a process deterioration (signal rate) from 2,000 simulated in-control (top) and out-of-control (bottom) ST-CUSUM runs. The desired FSP is marked by black symbols.



(a) In–control signal rate

(b) Out–of–control signal rate

Fig. 3.4. Percentage of RA-CUSUM charts signalling a process deterioration (signal rate) from 2,000 simulated in-control (top) and out-of-control (bottom) for risk-adjusted indicator 11724. RA-CUSUM runs were simulated for mixed, low and high risk populations. The desired FSP is marked by black symbols.

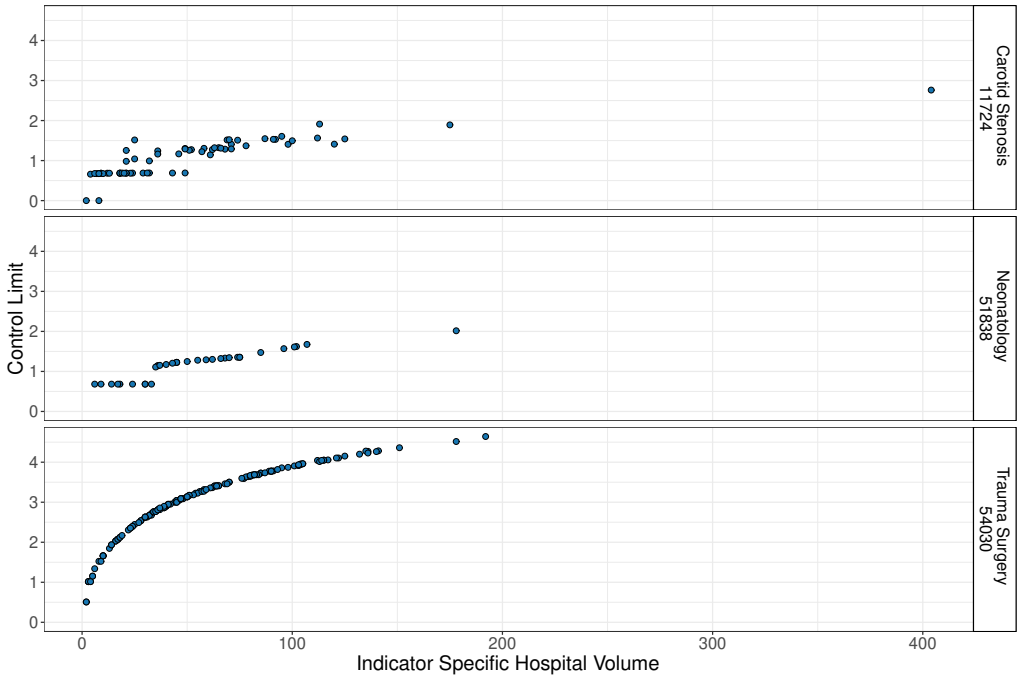


Fig. 3.5. Control limits for hospital performance data of EQA in Bavaria. Control limits were estimated on performance data of 2016 and simulated for $\delta = 2$ and FSP=5%.

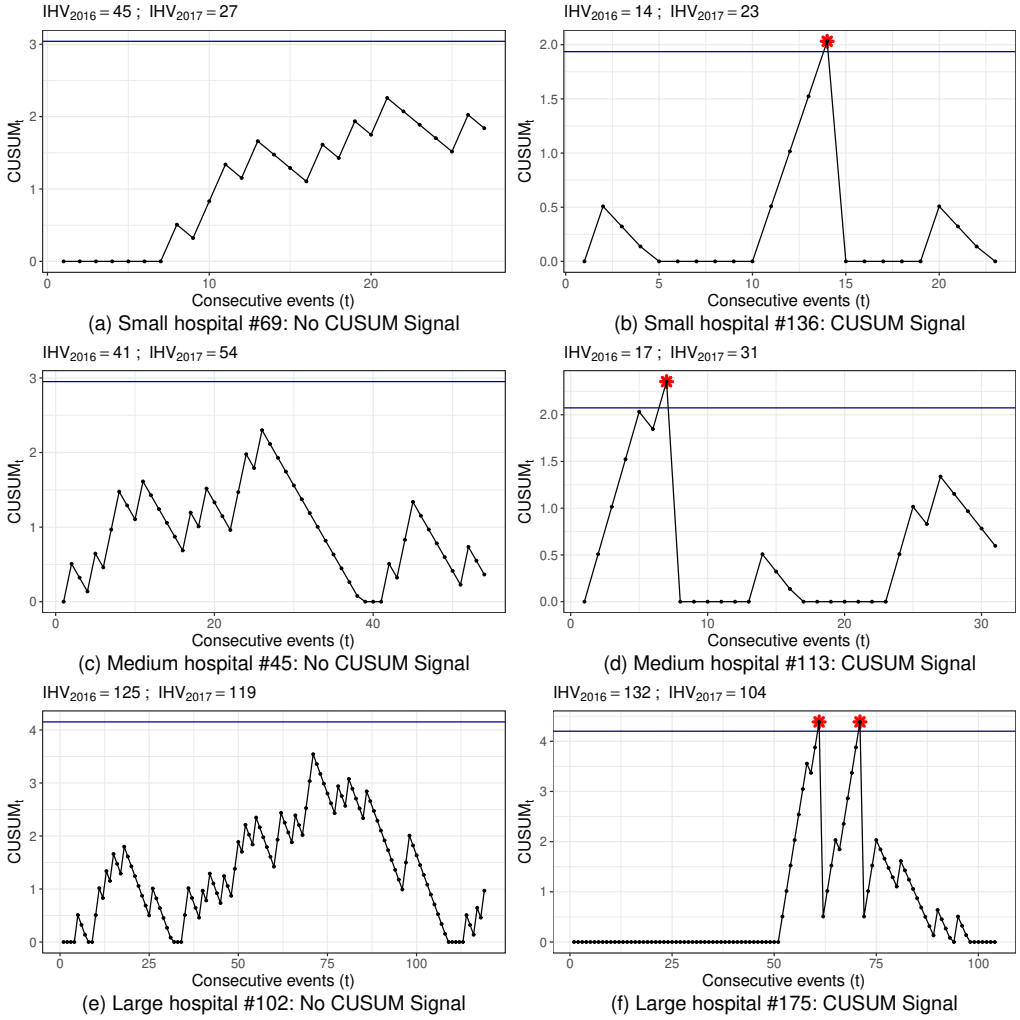
Performance data from 2016 is used to estimate baseline failure probability and case risk mix to construct CUSUM charts for performance data of 2017, though the monitoring period extends from 1 March 2017 to 28 February 2018, becauce documentation and transmission deadline is 28 February for the previous year with the reporting year shifted by two months.

CUSUM charts were constructed by simulating the control limit for a FSP of 5%. We set the detection level to $\delta = 2$ and constructed control charts for hospitals with hospital volume > 1 in 2016 and 2017.

We initiated all CUSUM runs with $C_0 = 0$ and reset C_t to zero after every signal, which is applicable if an investigation after a signal takes place and appropriately identifies underlying issues.[32]

Simulated control limits of ST-CUSUM charts for indicators 54030 and 51838 increased with increasing hospital volume to ensure a constant FSP during one observation period (Figure 3.5). Control limits of the RA-CUSUM chart for indicator 11724 increased as well, but adjustment of the different case risk mixes influenced variability of the control limits.

Of the 261 hospitals' CUSUM charts, 34 processes triggered a signal and were identified as out-of-control. Overall, 86.21% of the hospitals were classified as in-control (Table 3.1).

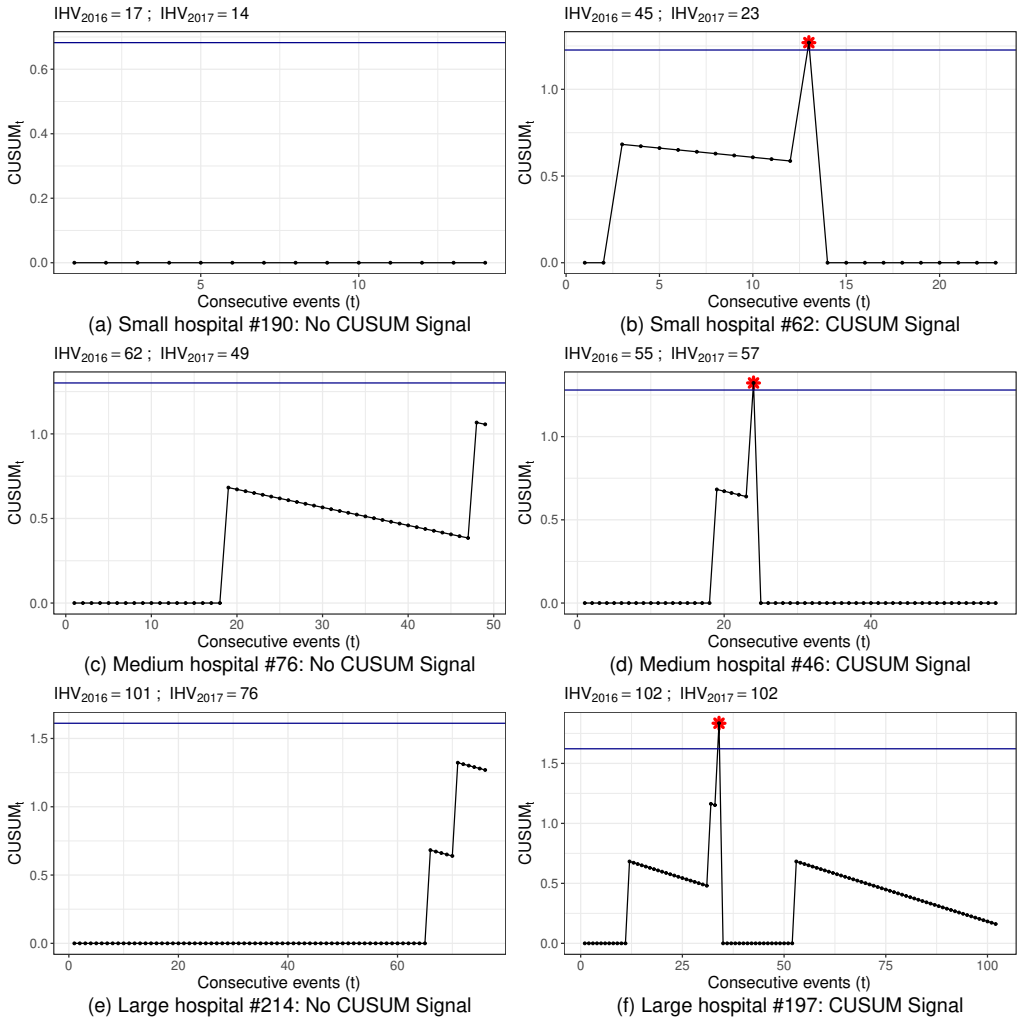


(a) Small hospital #69: No CUSUM Signal
(b) Small hospital #136: CUSUM Signal
(c) Medium hospital #45: No CUSUM Signal
(d) Medium hospital #113: CUSUM Signal
(e) Large hospital #102: No CUSUM Signal
(f) Large hospital #175: CUSUM Signal

Fig. 3.6. Trauma Surgery 54030. Selected ST-CUSUM plots for individual hospital annual performance data of 2017. IHV denotes indicator specific hospital volume.

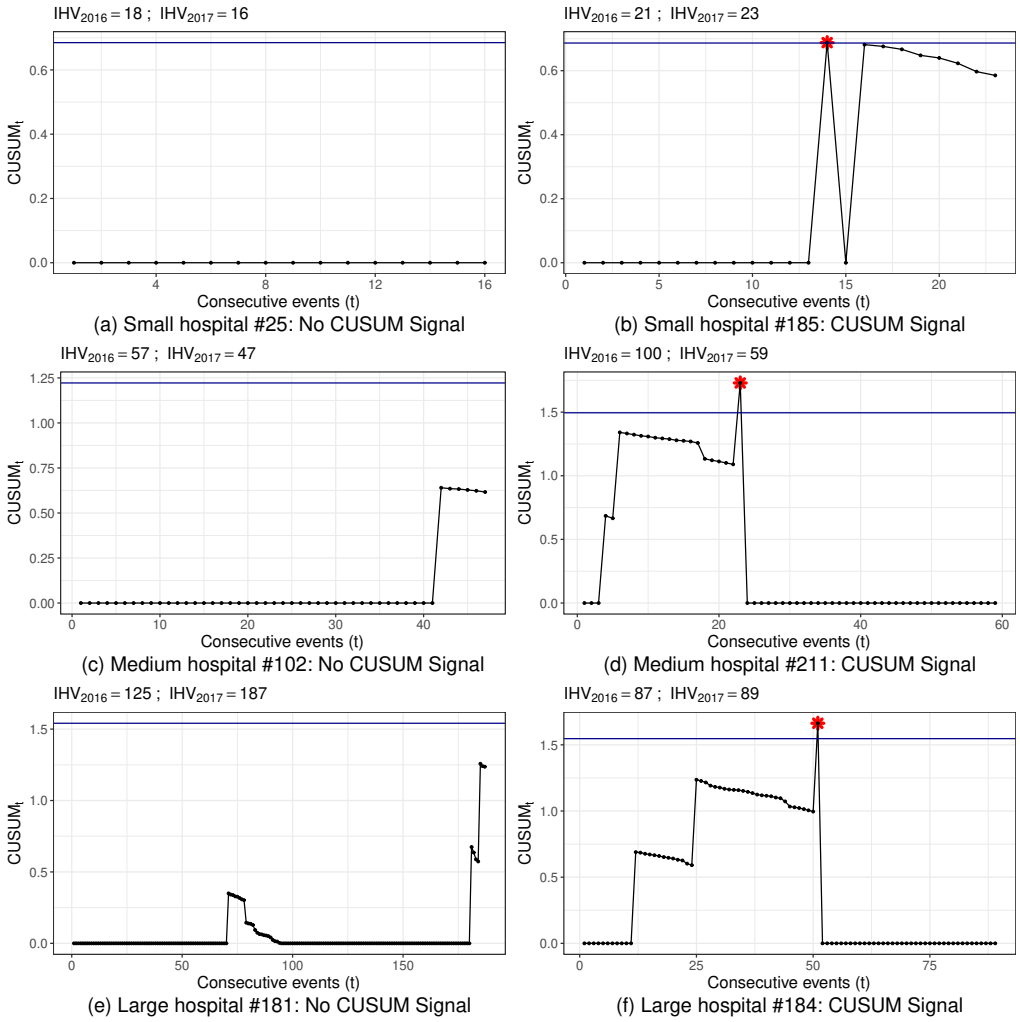
Table 3.1: Percentage of hospitals with CUSUM signals per performance indicator in Bavaria in 2017. Two of the control charts for indicator 11724 had to be discarded due to incorrect control limit (Signals: NA).

Signals	54030 ($n = 163$)	51838 ($n = 34$)	11724 ($n = 64$)
0	85.89%	85.29%	88.00%
1	9.82%	14.71%	9.00%
2	1.84%	0.00%	0.00%
3+	2.45%	0.00%	0.00%
NA	0.00%	0.00%	3.00%



(a) Small hospital #190: No CUSUM Signal
(b) Small hospital #62: CUSUM Signal
(c) Medium hospital #76: No CUSUM Signal
(d) Medium hospital #46: CUSUM Signal
(e) Large hospital #214: No CUSUM Signal
(f) Large hospital #197: CUSUM Signal

Fig. 3.7. Neonatology 51838. Selected ST-CUSUM plots for individual hospital annual performance data of 2017. IHV denotes indicator specific hospital volume.



(a) Small hospital #25: No CUSUM Signal
(b) Small hospital #185: CUSUM Signal
(c) Medium hospital #102: No CUSUM Signal
(d) Medium hospital #211: CUSUM Signal
(e) Large hospital #181: No CUSUM Signal
(f) Large hospital #184: CUSUM Signal

Fig. 3.8. Carotid Stenosis 11724.Selected RA-CUSUM plots for individual hospital annual performance data of 2017. IHV denotes indicator specific hospital volume.

As the positive CUSUM weights $W_t(y = 0)$, which decrease the CUSUM, were smaller for indicators 51838 and 11724 than for indicator 54030, adverse events were more difficult to compensate by good performance (e.g. Figure 3.8 (f)). For this reason, in-control CUSUM charts of indicators 51838 and 11724 generally allowed for no more than two adverse events. Still, out-of-control processes of indicators 51838 and 11724 had at most one signal.

CUSUM charts of indicator 54030 triggered multiple signals in several hospital processes. These hospitals most likely had a persistent quality deficit for this indicator and were not able to control the process during the entire monitoring period. For some hospital processes, it became clear that the quality deficit was only at one specific time due to a clustering of adverse events (Figure 3.6 (f)). These insights help to locate causes of quality deficits and lead subsequent investigations.

The hospital example also illustrates the influence of hospital volume on control charts. Larger hospital volume lead to wider control limits, allowing more adverse events within a year. Large hospital #102 (Figure 3.6 (e)) was categorised as in-control for indicator 54030, although a third of the observations were adverse events. Hospital #113 (Figure 3.6 (d)) had 29% adverse events for indicator 54030 and triggered a signal. This is partly due to the shorter sequence of adverse events and the smaller hospital volume. However, this hospital also had a substantial increase in volume from 2016 to 2017, so that the control limit was probably lower than necessary.

3.4 Concluding remarks

Controlling the FSP worked well for sufficiently large hospital volumes and high baseline failure probability. In monitoring schemes of small hospital volumes, it often remains impossible to adjust the control limit to fit a specific FSP, as these control charts are not as flexible as control charts for larger volumes.

Small hospitals present an issue in SPM, as corresponding CUSUM charts are difficult to construct and evaluate. In our simulation, it is quite possible that no failure was simulated for small hospital volume processes ($n_s = 7$), especially for indicators with a small failure probability such as for indicator 51838 ($c_0 = 1.25\%$). Detecting a doubling or tripling of odds with a small failure probability and small hospital volume is difficult, as even with doubled or tripled odds, the probability to observe no adverse event is still large. Taking this example, 92% of $n_s = 7$ observations show no adverse events at failure probability c_0 compared to 84% at doubled odds – i.e., in 84% of all possible sets of $n_s = 7$ patients, no difference between the in-control and out-of-control state is observable. As most control charts required at least two adverse events to signal, signals

became very unlikely. The hospitals' CUSUM charts in the example showed that small hospitals may still benefit from an individual investigation based on the CUSUM chart as differences in performance are fairly well illustrated. Hospital volume may be increased by extending the data to cover multiple years, if the achievable FSP is not acceptable.

Current German regulations require that in cases of an extremely adverse clinical outcome written explanations have to be furnished by the medical staff in every such instance. This strategy does not rule out the use of control charts for indicators with low baseline failure probability and we suggest that individual investigations of adverse events should accompany CUSUM charts for these indicators. The monitoring of rare events is a common issue in SPM and Woodall and Driscoll gave a comprehensive review on this topic.[70] In this context, our example ($c_0 = 1.25\%$) is not yet regarded as rare, as the methods discussed here consider failure probabilities that are ten or hundred times smaller.

As CUSUM charts are based on performance data of the previous year, they may be subject to uncertainty of these estimations. Monitoring across different years presents the additional challenge that specifications of performance indicators may change due to clinical recommendations of national advisory panels, and thus indicators may not always be comparable across different monitoring periods. Additionally, hospital volume and case risk mix vary across years, which affects the signal characteristics of the CUSUM scheme. It has been shown that wrong expectations of risk mix or wrong model specifications can have a significant impact on CUSUM runs .[61,71,72]

In the example, we reset the CUSUM after every signal to gain a sense of frequency of signals. However, according to the theoretical background of SPM in industrial process contro, this is only appropriate if the process is investigated and brought back in control, which is naturally more complex in hospitals. Additionally, when the CUSUM restarts with the same control limit as before, the FSP and TSP may be lower than anticipated, as the hospital volume decreases. If resetting the CUSUM to zero is not reasonable, resetting it to any value between zero and the control limit is also an option. This was already proposed by Lucas and Crosier in 1982,[62] and results in faster subsequent signals.

4 Group Sequential Cumulative Sum Chart

In the quality control of industrial production process, samples are taken regularly during the production process to inspect quality. In health care settings, it cannot be expected that a regular weekly sampling of performance data results in similar sample size due to patient fluctuation in the weekly hospital volume. Additionally, the exact sequence of patients is unknown, as patients are treated simultaneously or switch sequence during their inpatient stay.

As a result, the implementation of standard SPM methods is greatly hindered in common hospital settings. Standard SPM tools rely on meticulous data collection with two options: Observing the whole sequence of events, where each points reflects one observation ($n = 1$), or collecting regular samples of equal size ($n > 1$), to guarantee a consistent control chart performance.[32] The data set described and used here consists of b irregular samples $i = 1, 2, \ldots, b$ of size $n_i \geq 1$.

We introduce the Group Sequential CUSUM (GSCUSUM) chart for processes, where only aggregated binary performance data of unequal sample size are available. This approach can also be applied to other control charts, like the VLAD,[49,50] or the Exponentially Weighted Moving Average (EWMA).[73] It is also possible to extend the GSCUSUM to risk-adjustment or other enhancements described in Section 3.1.4.

This chapter provides details on the construction of GSCUSUM charts and evaluates their performance. Section 4.1 explains the construction for non-risk-adjusted as well as risk-adjusted indicators. Section 4.2 assesses the performance of GSCUSUM charts in a simulation study. In Section 4.3 the GSCUSUM chart is applied to performance data of EQA.

4.1 Construction of Group Sequential CUSUM charts

Consider the simulated data in Table 4.1. 50 observations, of which 10 are events or failures, are grouped in 10 samples of unequal size.

Table 4.1: Simulated group-sequential example data. The data consists of sequential blocks b, that contain n observations, of which m are events ($y = 1$).

Block b	Number of observations n	Number of events m	Cumulative n
1	4	0	4
2	5	2	9
3	6	0	15
4	8	1	23
5	5	2	28
6	4	0	32
7	5	0	37
8	1	1	38
9	7	3	45
10	5	1	50

For this data set, common SPM methods fail because of vastly different sample sizes. Particular sample sizes of size $n = 1$ (Block $b = 8$) greatly distort the result.

First, we describe the estimation of GSCUSUM charts based on data from the example (Table 4.1), which is also illustrated in Figure 4.1. At the end, we present the general algorithm to construct GSCUSUM and risk-adjusted GSCUSUM (RA-GSCUSUM) charts. Software to calculate GSCUSUM and RA-GSCUSUM charts is provided in the R package `cusum`[74] and presented in the Appendix 1.

The GSCUSUM charts permutes through possible sequences of observations, in order to estimate possible CUSUM paths, considering the uncertainty of he unobserved sequence of observation. In our example, the GSCUSUM initialises with $C_0 = 0$. Using data from the first block, new sequences are created by permuting the four observations. As the first block has no events, there is only one possible sequence of observations: The CUSUM statistic remains at zero.

The data of the second block, which ends at $t = 9$, are used in a similar way. Here we have five observations, of which two are events. Thus there are ten possible sequences and ten possible CUSUM paths (Figure 4.1).

For the third block, there are five possible CUSUM paths to continue. From each of these path, the last CUSUM statistic is used to initialise CUSUM paths, and new sequences are generated to estimate new possible CUSUM distributions. All following blocks are handled in a similar iterative way: the final CUSUM distribution of the previous block is used to calculate new possible CUSUM paths and estimate distributions of CUSUM statistics.

The general algorithm to construct GSCUSUM charts is as follows:

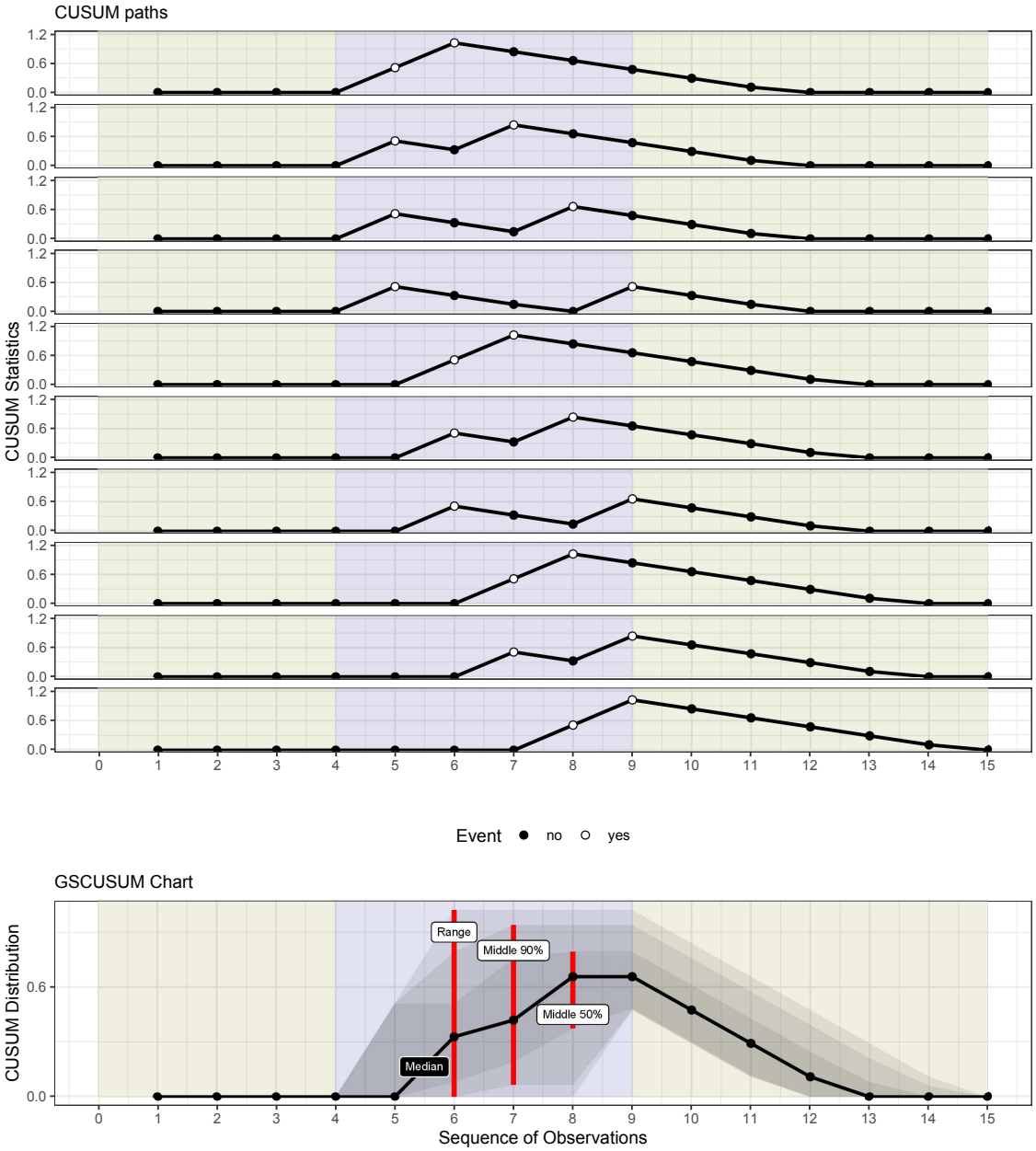


Fig. 4.1. First three data blocks of example Table 4.1. Data blocks are coloured in alternating yellow and blue.
The top plot shows all possible CUSUM paths that result from generated new sequences.The bottom plot is the resulting CUSUM distribution, illustrated in a GSCUSUM chart. The main black path represents the median of the CUSUM distribution, and the middle 50%, middle 90% and total range are shaded in grey.

1. First block data Y_i for $i = 1, \ldots, t$
2. Repeat multiple times:
   a) Shuffle sequence of data: Y_i^*
   b) Initialise CUSUM with $C_0 = 0$
   c) Calculate CUSUM statistics on Y_i^*: $C_{Y_i}^*$
   d) Estimate distribution of CUSUM statistics
3. Second block data Y_j for $j = t + 1, \ldots t'$
4. Repeat for all C_t:
   a) Repeat multiple times:
      i. Shuffle sequence of data: Y_j^*
      ii. Initialise CUSUM with C_t
      iii. Calculate CUSUM statistics on Y_j^*: $C_{Y_j}^*$
   b) Estimate distribution of CUSUM statistics
5. Repeat 4. for all following blocks, always considering all final CUSUM statistics of the previous block as initialisation of the CUSUM

This algorithm can be used to construct non-risk-adjusted GSCUSUM and risk-adjusted GSCUSUM (RA-GSCUSUM) charts. In the risk-adjusted case, performance data include individual patient risks for adverse event $p_i \in [0, 1]$, that are considered when estimating the RA-CUSUM statistic as described in Chapter 3. Notice that the possible number of different sequences increases in the risk-adjusted case. Each observation has to be permuted regardless of the outcomes in a block, as each RA-CUSUM weight is individual and each sequence of observation can result in a different RA-CUSUM path. The number of possible sequences is thus the factorial of the number observations in one block, and it may not be feasible to permute through all possible sequences beyond a certain block size.

Yet oftentimes, the GSCUSUM is simplified. When all CUSUM weights within a block are the same, the sequence of observations is irrelevant. When no adverse events occur, the CUSUM distribution returns to zero, mimicking a restart of the CUSUM and all initialisations of the GSCUSUM at the beginning of a new data block are zero. Additionally, when data blocks are small, the GSCUSUM converges to the traditional CUSUM, as expected.

In CUSUM charts, a signal is triggered when the CUSUM statistic crosses the control limit. Instead of only one CUSUM statistic per observation, the GSCUSUM gives all possible CUSUM statistics, so the simple decision rule of the traditional CUSUM chart fails. Still, the GSCUSUM may return a probability of signalling for every point in time by aggregating all CUSUM statistics and estimating the rate γ of statistics greater than the control limit. We use this rate γ as a benchmark when assessing GSCUSUM chart performance, as we need a definite signal of good or poor performance. Smaller values

of γ return more false signals and larger γ may miss important changes. To balance the risk of false signals and false negatives we set γ to 0.5.

Nevertheless, we abstain from plotting signals in the GSCUSUM plots, as these signals are not inherent to the GSCUSUM chart and the process evaluation is not as straightforward as a standard signal might suggest.

4.2 Performance of Group Sequential CUSUM charts

The performance of GSCUSUM charts was evaluated in a simulation study and compared to the traditional CUSUM chart. We analysed three scenarios:

1. The fully observed CUSUM chart as reference
2. The outlier CUSUM chart, which groups all adverse events $y_i = 1$ as close as the data blocks allow together in order to simulate a worst case scenario
3. The GSCUSUM chart using the information of grouped outcomes in blocks of different sample size

Data sets for hospital volumes between 10 and 100 were simulated, denoting the number of patients documented in a hospital during a monitoring period. Data blocks were generated by simulating the outcomes in groups of unequal sample sizes. Data quality was simulated for three levels: bad, fair and good data quality. The number of blocks n_b per process was used as a measure of data quality, e.g. many blocks imply rather small blocks overall and thus better data quality; with few blocks the opposite is implied. The number of blocks n_b were defined in relation to hospital volume n:

- Bad data quality: $n_b = n * 0.75$
- Fair data quality: $n_b = n * 0.5$
- Good data quality: $n_b = n * 0.25$

In-control performance was simulated with an odds-multiplier of 1 and out-of-control performance was simulated with an odds-multiplier of 2. All charts were set up to detect a performance change from $c_0 = 20.35\%$ to $c_A = 32.22\%$, which corresponds to a doubling of odds for failure($\delta = 2$). This baseline failure probability c_0 is deduced from the performance indicator 54030 of the motivating example and is the Bavarian average failure rate of all hospitals in the year 2016.

The GSCUSUM charts' control limits were set like the traditional CUSUM control limits to a false signal probability of 5%. A total of 10'000 data sets were generated per scenario on which the CUSUM, the outlier CUSUM and the GSCUSUM were calculated.

Data were simulated using R 3.6.0.[75]

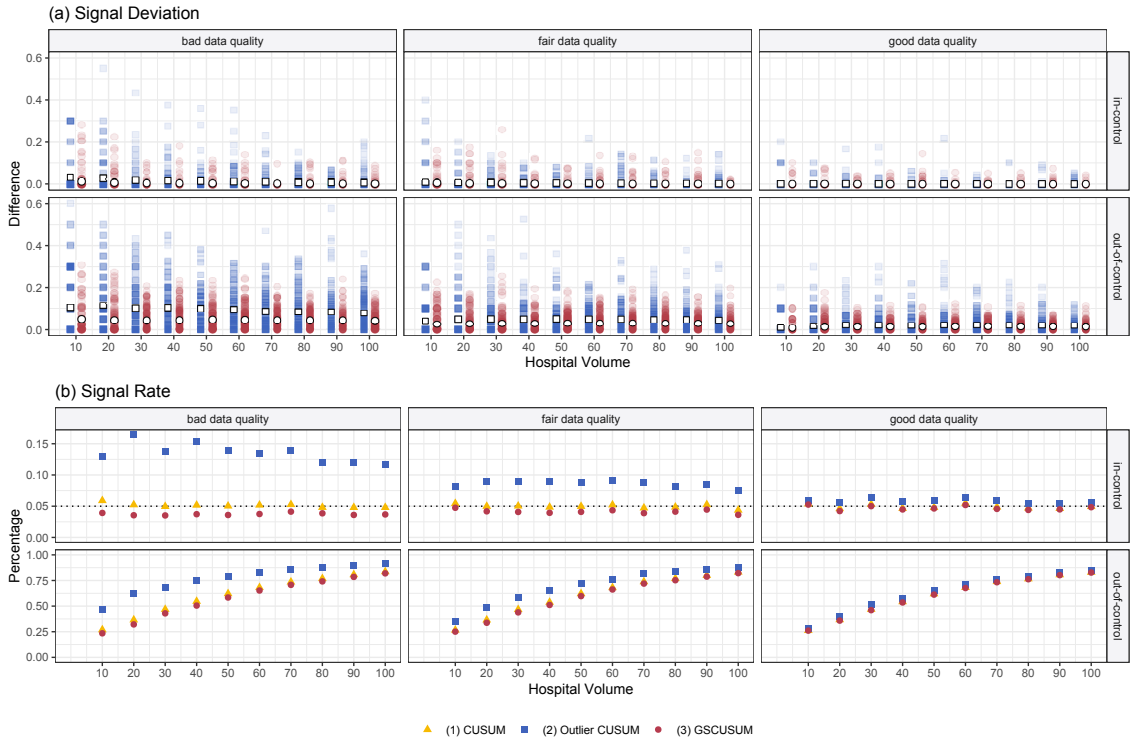


Fig. 4.2. Performance of (3) GSCUSUM compared to (1) CUSUM and (2) Outlier CUSUM chart. The top plots show deviations from signals of reference CUSUM to GSCUSUM and outlier CUSUM (10'000 random results per method are plotted; white points refer to average value). Bottom plot show signal rates of in- and out-of-control performance (dashed line refers to desired false signal probability level of 5%).

The primary estimands of the simulation study were the signal probabilities of the in-control and in the out-of-control processes. Additional, the deviation from the reference CUSUM was estimated for the outlier and the GSCUSUM in the form of deviation from a signal at every observation. While the signal rate indicates the signalling of control charts over the whole run, the signal deviation is a marker for delayed or early signals that are a result of the loss of information about the true sequence.

Figure 4.2 (a) displays the deviation in signalling time of the outlier-CUSUM and GSCUSUM to the reference CUSUM. Signal deviation was greater in the out-of-control scenarios than the in-control scenarios. For all cases, the outlier-CUSUM had a greater deviation in signalling time than the GSCUSUM. Hospital volume does not seem to influence the deviation greatly. As expected, the effects of the different analysis methods became insignificant with better, i.e. more continuous, data documentation. Signal probabilities of the in- and out-of-control processes are shown in Figure 4.2 (b). Signalling rates of the GSCUSUM were close to the true CUSUM; in parts the GSCUSUM was more conservative than the reference CUSUM. In contrast, the outlier CUSUM signalled vastly more when data quality was worse.

4.3 Application of Group Sequential CUSUM charts

We illustrate the characteristics of the GSCUSUM with hospital performance data of the EQA in Bavaria, Germany. The non-risk-adjusted GSCUSUM is shown on indicator 54030, and the risk-adjusted GSCUSUM on indicator 11724, as introduced before.

Baseline failure probability, case risk mix and hospital volumes were estimated on data from 2016. CUSUM and GSCUSUM charts were constructed to detect a doubling of odds from the Bavarian average of all hospitals in the year 2016 ($c_0 = 20.35\%$; $c_A = 32.22\%$). For the risk-adjusted example, case risk mix was estimated individually for each hospital on their data from 2016. Control limits were set to a false signal probability of 5%. We assumed monitoring began for 2017.

Data were grouped in samples based on the date of documentation recorded in the data set. Additionally, we constructed traditional CUSUM charts where patients with the same date of documentation were randomly rearranged in order to receive a pseudo-sequence of observations.

Exemplary control charts of three hospital processes for the two performance indicators illustrate the GSCUSUM (Figure 4.3) and RA-GSCUSUM (Figure 4.4). Figure 4.3 (a) shows an example of an out-of-control process. The pseudo-sequential CUSUM charts as well as the GSCUSUM charts classified the process as out of control. In Figure 4.4 (a), the RA-CUSUM signals, and also the median of the RA-CUSUM distribution in the RA-GSCUSUM chart crossed the control limit. CUSUM statistics of the second examples (Figure 4.3 (b) and Figure 4.4 (b)) did not exceed the control limit, and also the GSCUSUM did not give cause to suspect poor performance.

In Example 3 the interpretation of the charts differs. The CUSUM chart signalled at observation 86 (Figure 4.3 (c)), though the median of the GSCUSUM did not cross the control limit at this point. In the later run one could have suspected a process change, though the statistic seemed to recede below the control limit at the end of the monitoring period. In this case it remains unclear if a signal was warranted at observation 86 or if at this point no deviation had occurred. The RA-GSCUSUM of Example 3 (Figure 4.4 (c)) clearly shows that the signal in the RA-CUSUM chart was the result of an outlier, as the majority of sequences did not cross the control limit.

4.4 Concluding remarks

The GSCUSUM chart proposed here poses a valuable and practical extension to the standard Bernoulli log-likelihood CUSUM chart to more realistic hospital performance data settings. GSCUSUM is a good approximation of the true CUSUM statistic when

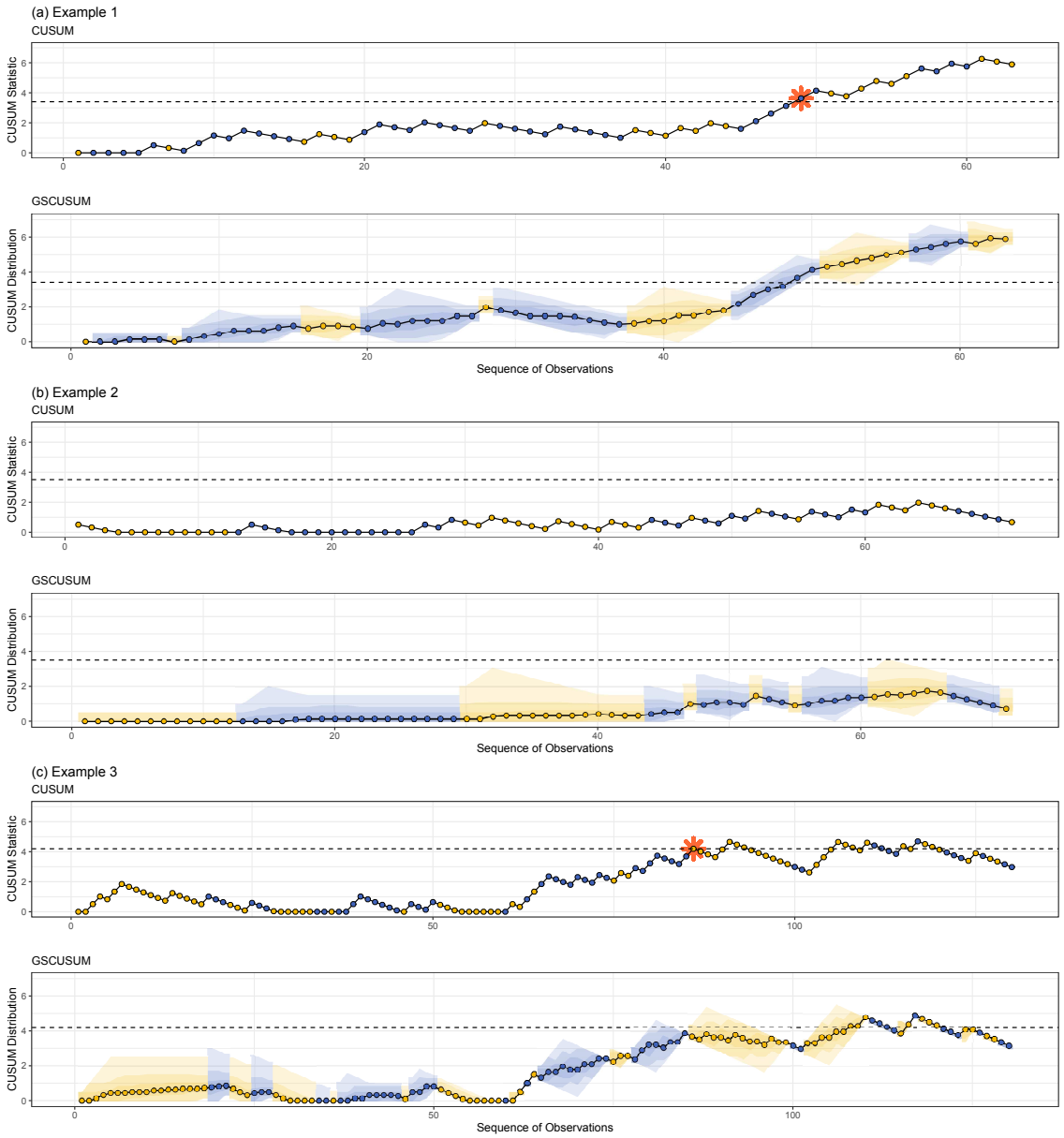


Fig. 4.3. Three examples of pseudo-sequence CUSUM (top) and GSCUSUM (bottom) for hospital performance data of Bavarian external quality assurance in 2017 (Indicator 54030). Data blocks are coloured in alternating order. In the GSCUSUM plots, the median of the CUSUM distribution is shown as the main path, shaded areas present middle 50%, middle 90% and total range.

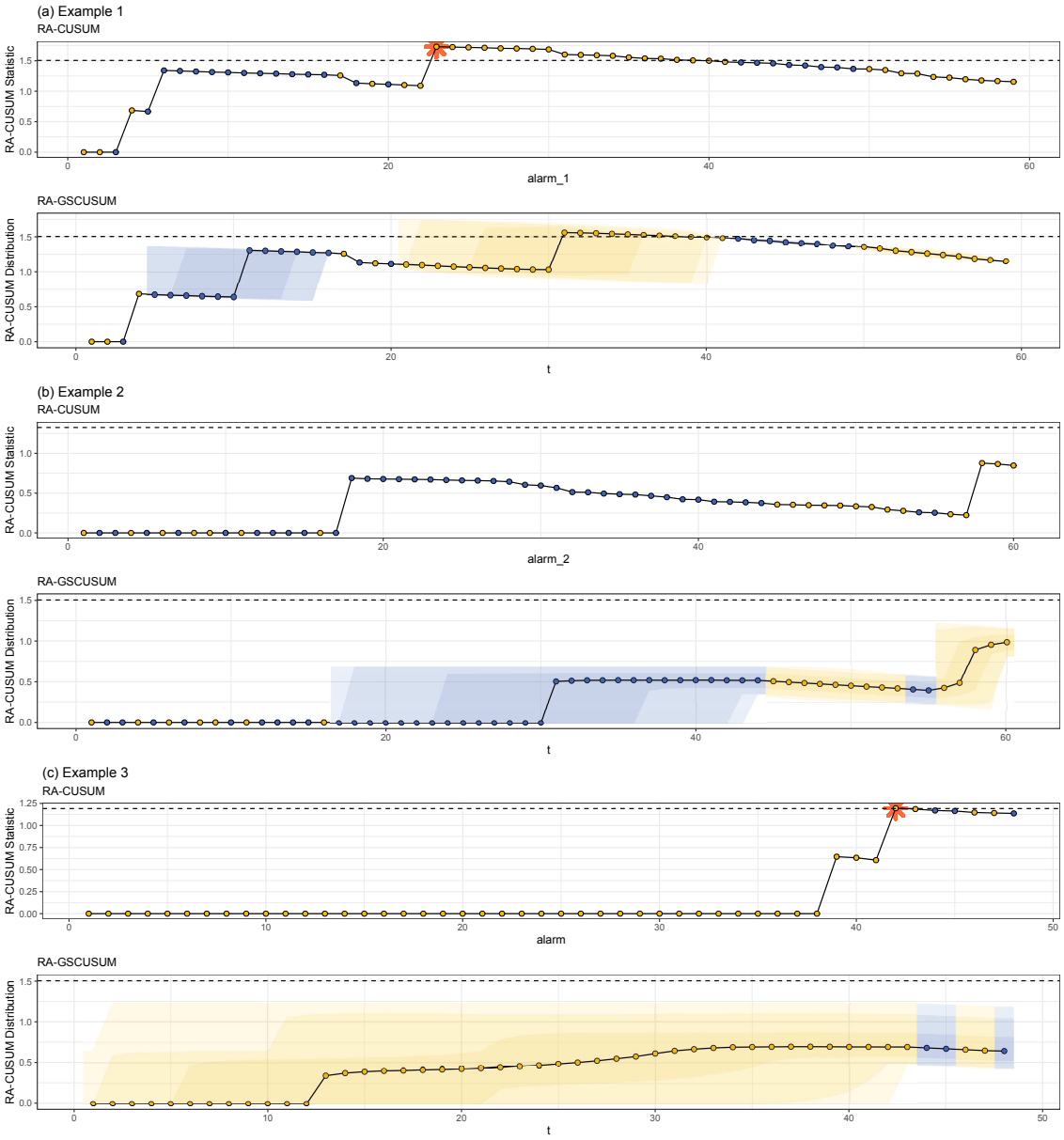


Fig. 4.4. Three examples of pseudo-sequence RA-CUSUM (top) and RA-GSCUSUM (bottom) for hospital performance data of Bavarian external quality assurance in 2017 (Indicator 11724). Data blocks are coloured in alternating order. In the RA-GSCUSUM plots, the median of the RA-CUSUM distribution is shown as the main path, shaded areas present middle 50%, middle 90% and total range.

the complete sequence of observations is not observed: The GSCUSUM and the fully observed CUSUM chart signalled almost simultaneously and signal rates were comparable. We illustrated the concept on the Bernoulli log-likelihood CUSUM chart for non-risk-adjusted and risk-adjusted processes; it is also possible to apply the method to other types of control charts that monitor binary sequences.

Our simulation study and the hospital example showed that the naïve approach of randomly assigning a sequence may change the signalling of CUSUM charts and result in wrong conclusions about the process performance, as it is possible to select an atypical outlier sequence. The GSCUSUM, in contrasts, studies all possible sequences that fit the grouped data, thus giving an accurate probability of a signal for each point in time.

The software implementation of GSCUSUM and RA-GSCUSUM is not straightforward and can be computationally expensive if not programmed optimally. We therefore provide R-Code with a fast C++ implementation to calculate GSCUSUM and RA-GSCUSUM charts in the R-package `cusum`.[74]

A major limitation of the GSCUSUM chart is the loss of the inherent signal that is available in the CUSUM chart. We have proposed an alternative signalling when a proportion γ of CUSUM statistics cross the control limit, though a more conservative choice may be to signal once all possible CUSUM statistics exceed the control limit. Naturally, different proportions of exceeding CUSUM paths (γ) may also be interpreted as different types of signals. For example, 25% of paths crossing may lead to a warning signal whereas 75% of paths crossing cause an alarm signal.

We assumed that observations in data blocks are randomly distributed, but it is also plausible that failures are more likely to occur clustered due to a common source of error. Our approach tends to avoid false signals, as the equal distribution of observations enables compensating for failures before the next occurs. If there are grounds to believe that this assumption is violated in an application, it is useful to evaluate a GSCUSUM that takes clustering of failures into account, i.e. the outlier CUSUM.

To conclude, the GSCUSUM chart enables the use of SPM in applications where using a CUSUM chart was previously not possible due to a lack of data quality. When data quality is improved, the GSCUSUM chart also converges smoothly to the CUSUM chart and hence provides for a seamless handling of the transition period until optimal data quality is achieved.

5 Application to hospital performance data

This chapter applies both methods, presented in Chapter 3 and 4 respectively, the CUSUM and the GSCUSUM, to hospital performance data for the three introduced performance indicators 11724, 51838 and 54030, and compares the evaluation results to the evaluation conducted by EQA. The objective is to find differences and similarities in evaluations and show how control charts can assist conventional EQA evaluation methods.

5.1 Performance evaluation methods

In-control parameters were estimated on hospital performance data of 2016, and performance data from 2017 were monitored and evaluated. Hospitals with fewer than two observations per year and performance indicator were excluded from analyses.

Following the evaluation process used by the existing EQA, annual hospital results were calculated by building the average failure rates for non-risk-adjusted performance indicators and ratios of observed to expected events for the risk-adjusted performance indicator. Upon these results, hospital processes were either classified as acceptable, arithmetically deviating or statistically deviating. Arithmetically deviating is defined as an absolute deviation from the target value or national average. Statistically deviating also considers sample size and is defined as the target value being outside the range of the confidence interval of the hospital result.

Control charts were constructed for a FSP of 5% and an detection level of $\delta = 2$. For indicator 54030, the specified target value was used as baseline failure probability, and for indicator 51838 the overall average failure rate of the year 2016 was used, as no target value is specified for this indicator. All simulations necessary to construct CUSUM and GSCUSUM charts were performed 100,000 times.

Analyses were done using R 3.6.0,[75] and the `cusum` R package.[74] Statistical deviation were calculated using the R package `IQTIGpvci`, provided by the IQTIG.[76] R Code

to evaluate hospital performance using CUSUM, GSCUSUM and conventional EQA is provided in the Appendix 2. The result is presented as an upset plot, which was created using the R package `UpSetR`.[77]

5.2 Performance evaluation results

261 hospital processes were included in the analyses. The upset plot in Figure 5.1 shows that 159 (60.9%) were classified by all methods as acceptable performance. Conventional EQA methods classified 57 processes as arithmetically deviating and 39 as statistically deviating. SPM signalled 53 CUSUM signals and 30 GSCUSUM signals. 27 (10.3%) hospital processes were statistically deviating and were also signalled by the CUSUM and GSCUSUM charts. Hospitals should be informed about a potential quality deficit in these processes and some form of intervention should occur. Figure 5.9 shows two of the 27 hospital processes, which shows the out-of-control state of the process during the complete monitoring period.

15 (5.4%) additional processes were arithmetically deviating and triggered CUSUM and/or GSCUSUM signals and could be considered for intervention. The hospital examples in Figure 5.8 suggest that some of these processes were able to mask a poor run for the conventional EQA evaluation. Thus, they were only classified as arithmetically deviating, but triggered both CUSUM and GSCUSUM signals.

While an arithmetical deviation may be due to chance, statistical deviations are more robust. Yet, seven statistically deviating processes were not detected by either control chart. This might have been because events were evenly distributed over the whole monitoring period (Figure 5.4), which did not cause the CUSUM to break out and cross the control limit. The second example in Figure 5.4 (right) also shows that the CUSUM depends on exact estimation of hospital volume to construct control charts. This hospital decreased in hospital volume from 2016 to 2017, resulting in a wide control limit.

Six (2.3%) processes were only signalled by the two control charts, all of these in processes of Indicator 11724 (Figure 5.5). This is mainly because the target value is set very high (4.58), and hospital processes with an increased ratio of observed to expected are not detected by the EQA.

Better data documentation was needed in processes where the CUSUM and the GSCUSUM differ in signalling (Figures 5.6–5.7). Acting upon a signal is hindered because the validity of a signal is unclear.

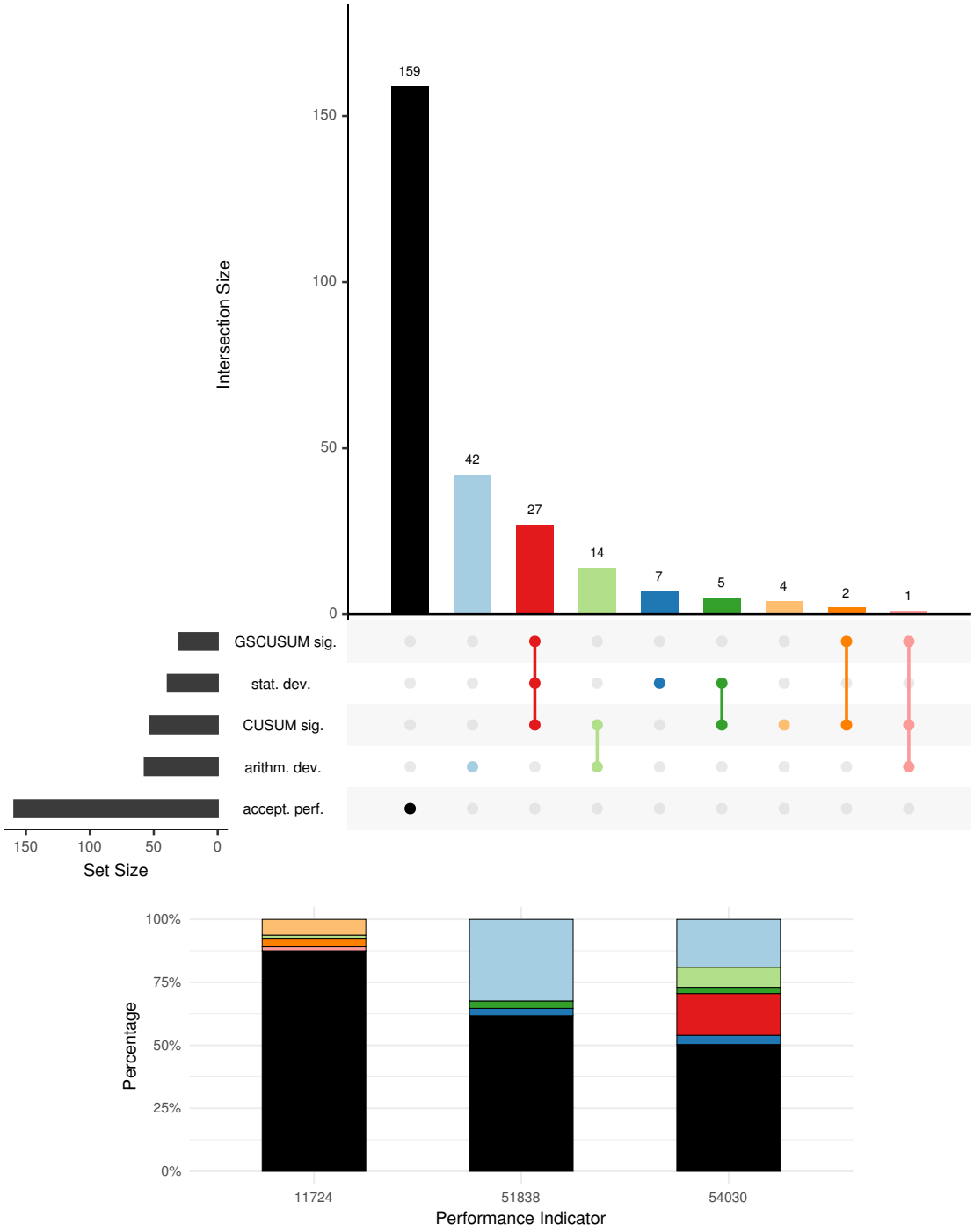


Fig. 5.1. The top figure shows all possible combinations of evaluation results (top) as well as the total counts of each evaluation (left) summed for all three performance indicators. The middle plot acts as a guide to illustrate the combinations. For example, the red bar (third from left) shows the total number of hospital processes that were classified as statistically deviating and as out-of-control by the GSCUSUM and CUSUM chart.
The bottom figure shows the occurrence of combinations of evaluation results per performance indicator. Colouring of the subsets is derived from the top plot.

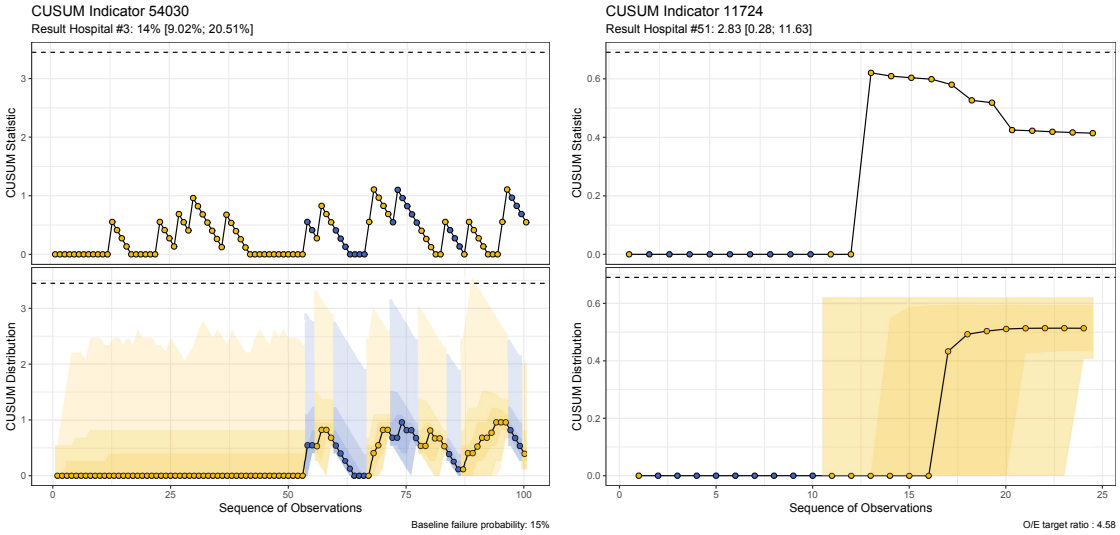


Fig. 5.2. Example hospital control charts that were classified by all methods as acceptable performance. The hospital results (left: 14%; right: 2.83) are smaller than the reference value (left: $\leq$15%; right: $\leq$4.58) and their confidence interval includes the reference value. Both hospitals are thus classified as acceptable performance by the EQA, and also the control charts give no cause to suspect a deviation.

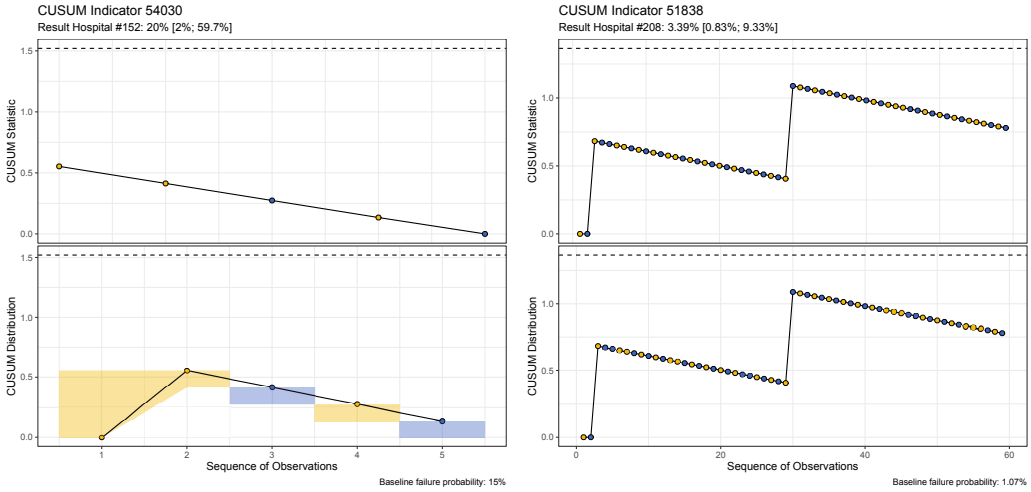


Fig. 5.3. Example hospital control charts that were only classified by the EQA for arithmetic deviation.Both hospital results (left: 20%; right: 3.39%) are greater than the baseline failure probability (left: $\leq$15%; right: $\leq$1.07%), but as their confidence interval includes the reference value, they are classified as arithmetically deviating. Both CUSUM and GSCUSUM do not signal.

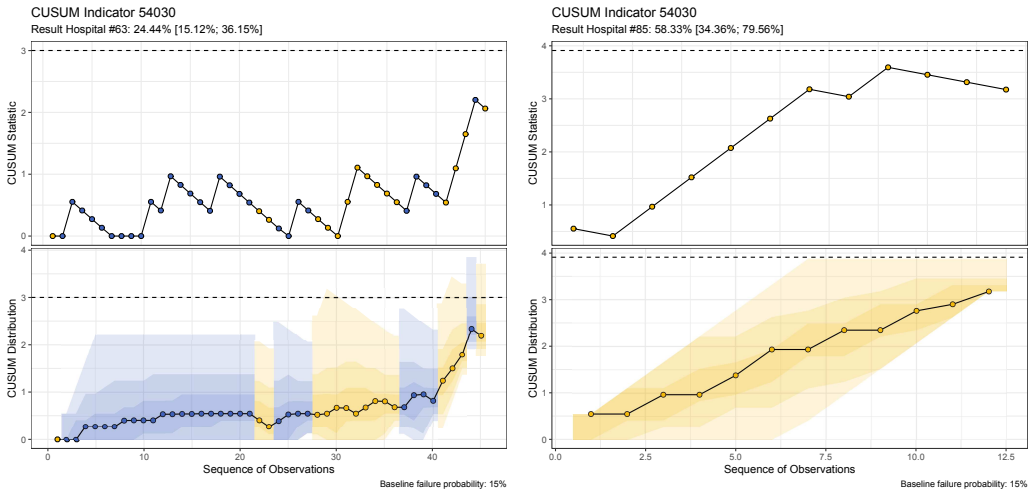


Fig. 5.4. Example hospital control charts that were only classified by the EQA for statistic deviation. The hospital results (left: 24.44%; right: 58.33%) are greater than the baseline failure probability (both: 15%), and the confidence intervals do not contain the reference value, thus being classified as statistically deviating by EQA. The CUSUM does not signal, though this may be due to poor data documentation, as the GSCUSUM suggests.

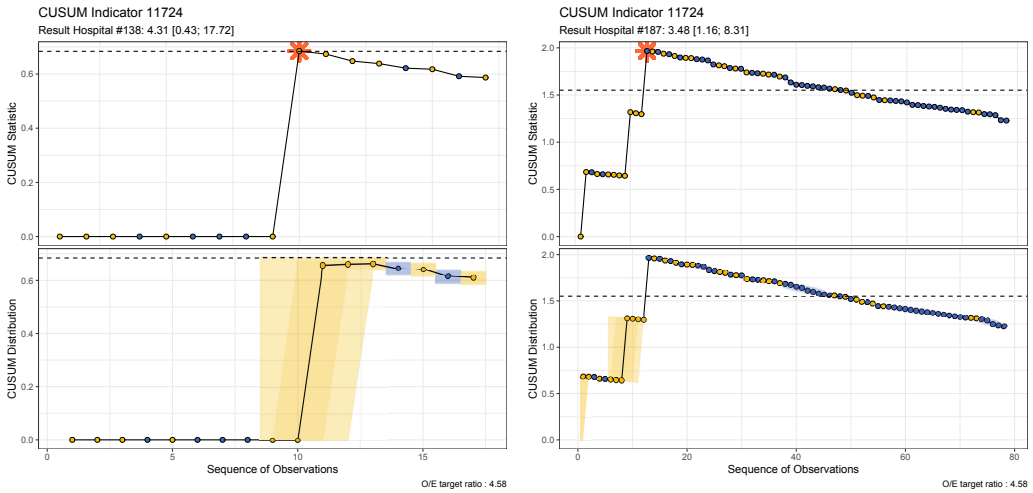


Fig. 5.5. Example hospital control charts that were only classified by CUSUM and/or GSCUSUM as out-of-control. The hospital result (left: 4.31; right: 3.48) is smaller than the target ratio (4.58), and is classified by EQA as acceptable performance. As the control charts are set to detect a doubling of odds, they trigger a signal.

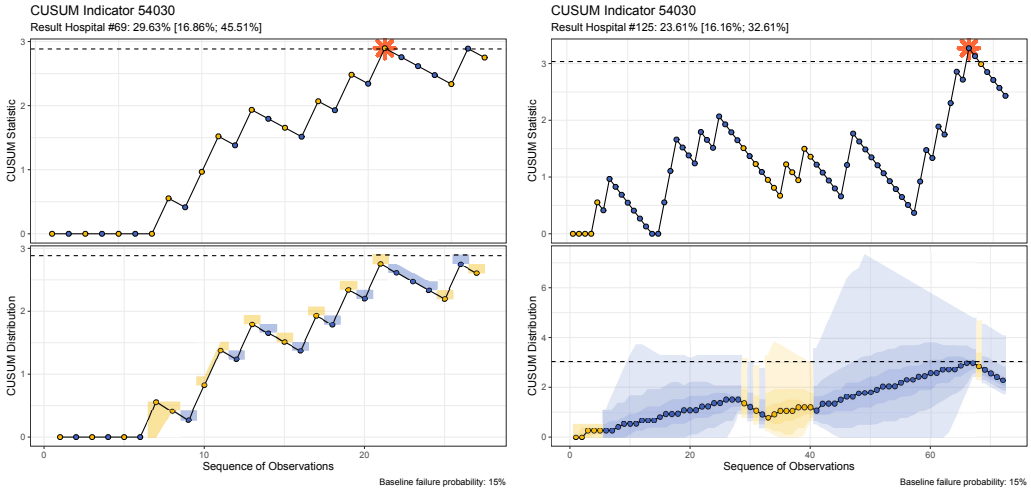


Fig. 5.6. Example hospital control charts that were classified by the EQA for statistic deviation and the CUSUM as out-of-control. Both hospital results (left: 29.63%; right: 23.61%) and their confidence intervals are greater than the baseline failure probability (both: 15%). The GSCUSUM failed to signal due to poor data quality.

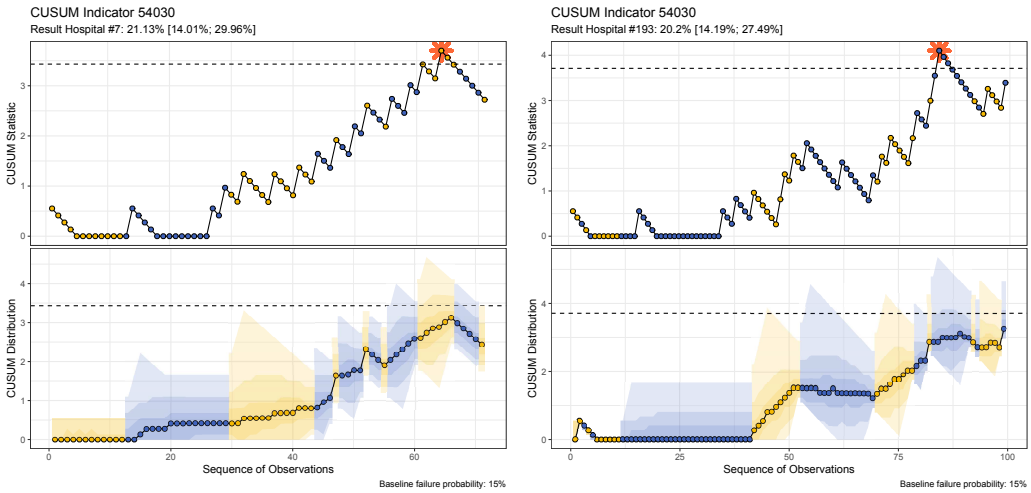


Fig. 5.7. Example hospital control charts that were classified by the EQA for arithmetic deviation and the CUSUM as out-of-control. Both hospital results (left: 29.63%; right: 23.61%) and their confidence intervals are greater than the baseline failure probability (both: 15%). Although the CUSUM signalled, the GSCUSUM failed to do so. For the right hospital, this is due to poor data documentation. For the left hospital, it is likely that the hospital volume decreased from the previous year and the control limit was set too high, resulting in a late CUSUM signal.

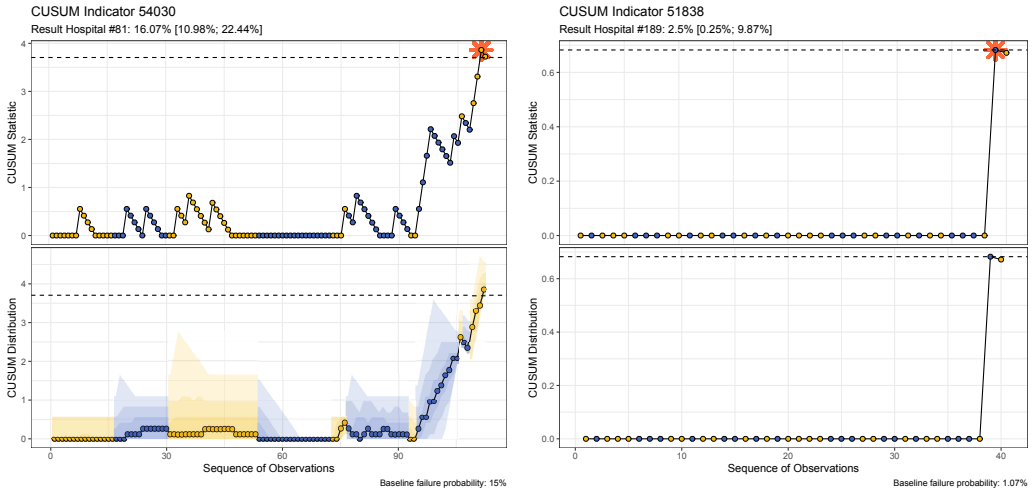


Fig. 5.8. Example hospital control charts that were classified by the EQA for arithmetic deviation and the CUSUM and GSCUSUM as out-of-control.Both hospital results (left: 16.07%; right: 2.5%) are greater than the baseline failure probability (left: 15%; right: 1.07%), but the confidence intervals contain the baseline failure probability. Both processes start out with acceptable performance, but deviate at the end. Therefore the processes are signalled by the GSCUSUM and CUSUM, but the bad run is masked in the EQA evaluation by a run of good performance.

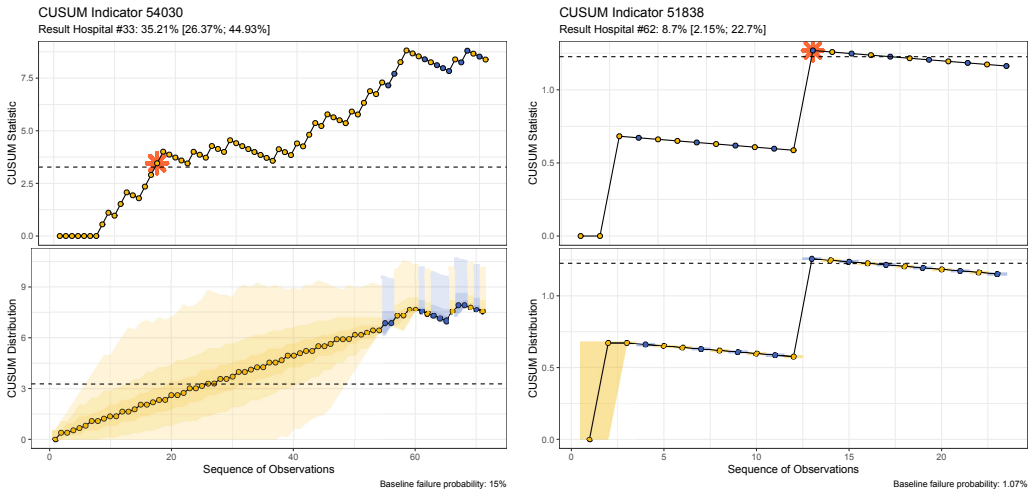


Fig. 5.9. Example hospital control charts that were classified by the EQA for statistic deviation and the CUSUM and GSCUSUM as out-of-control. Both hospitals results (left: 35.21%; right: 8.7%) and their confidence intervals are greater than the baseline failure probability (left: 15%; right: 1.07%). In the left chart, the CUSUM clearly shows the out-of-control state of the process. The right CUSUM chart shows the almost exact sequence of observations at the time of the signal, so there is clarity for when the signal happened.

5.3 Concluding remarks

While the final results of the evaluation methods do not differ greatly, the CUSUM enhances the evaluation procedure and provides additional and more timely information about the process over shorter periods of time.

Furthermore, by visualising the process, the CUSUM helps to contextualise failures and identify potential periods of poor performance. Completely deviating CUSUM charts (Figure 5.9) also show that the process is fundamentally out-of-control and needs to change. Processes which change from in-control to out-of-control during an observation period (Figure 5.8, left) are also visualised straightforwardly. Moreover, in such a case, it becomes easy to identify the moment in time at which the process changed. Naturally, this will make it more feasible for the health care provider to bring the process back under control as compared to the singular yearly average provided by current EQA methods.

The main conclusion when comparing the CUSUM and GSCUSUM chart must be the importance of regular data documentation. When the control charts and EQA evaluation differ, it is important to have a definite signal and know whether the suspected quality deficit is due to a random sequence or a true difference in the classification of performance. Often, when EQA signals, the GSCUSUM shows that a signal could have been warranted if the sequence was different (Figure 5.4).

6 Conclusion

Augmenting established EQA with concurrent SPM may well help to improve the explanatory power of quality assurance measures. Furthermore, SPM control charts may detect quality deficits without a delay, which in turn provides the opportunity to intervene and improve performance before any deterioriations are detected using conventional EQA methods. Moreover, control charts provide a visualisation of the process, presenting events in their temporal context, indicating trends or seasonal effects. CUSUM charts can also be of assistance in the evaluation of interventions and will facilitate showcasing best practice examples.[33]

To be statistically rigorous, the whole sequence of observation should be observed. So far, the date of documentation remains an unsatistfactory surrogate parameter in place of a precise time stamp. While the GSCUSUM chart enables the use of SPM in this kind of setting where the complete sequence of observation is not known, signals are less accurate and not as interpretable as for the standard CUSUM chart. In the future, even further advances in processing electronic health records may help to approximate real time bed-side performance evaluation. For the time being, performance monitoring is still constrained by unnecessarily complicated and laborious processes of data documentation, transmission, validation and evaluation. Further advances in timely data documentation can be motivated by the prospect of implementing efficient SPM.[33]

Naturally, there exists a trade-off between low FSP and high TSP when setting up a monitoring scheme. Prioritising a low FSP will protect hospitals with good quality of care from false accusations. As all signals require investigation at hospital level, false signals will result in unnecessary draining of resources of monitoring investigators as well as of those investigated. Still, detecting deteriorations should not be disregarded and an adequate balance between FSP and TSP should be sought. A FSP of 5%, which can result in an acceptable TSP, may be a reasonable choice for most scenarios.[33]

Adjustment for case risk mix is necessary for a fair and robust quality assurance. If a risk-model for the particular indicator exists, risk-adjusted CUSUM charts are easy to implement and their performance in our study was similar to the conventional CUSUM charts.[33]

Benefits and issues arising from simultaneously monitoring multiple data streams should be dealt with more thoroughly when implementing CUSUM in German EQA. Multiple indicators of one hospital can provide additional information about the hospital's performance. The global probability for a false discovery (FDR) can, however, increase with multiple data streams. Methods to control FDR of multiple data streams need to be developed and evaluated for their suitability in the monitoring scheme of German EQA.[33]

The recent change from an annual data transmission to quarterly data transmission shows the political interest of a more timely quality deficit detection. Yet, the main analysis is still based on the annual average, which is a weak approximation as pointed out in Section 2.2.

Performance indicators especially eligible for CUSUM analysis are indicators, which are clinically relevant and have been difficult to control over the last years. Furthermore, specifications should remain constant over time in order to allow construction of long term CUSUM charts. SPM can be introduced for individual performance indicators in addition to the standard procedure and it is not necessary to completely overhaul the EQA procedure at once.

Some currently implementable areas of SPM are as follows:

- **The state offices** receive data from indirect procedures on a regular basis and are able to provide continuous monitoring.
  - CUSUM charts could be reported alongside conventional EQA results, but possibly on a more regular basis. These analyses can warn health care providers of possible deviations and support the intervention procedure.
  - CUSUM charts could be used to guide the structured dialogues and help identify sources of performance deficits.
- **The IQTIG** receives the data from the indirect procedure only in aggregate form, so continuous monitoring over the year is not possible. Some post-analysis could still be worthwile:
  - CUSUM charts for process improvements could be used to evaluate measures of interventions, like the Structured Dialogue.
  - Like the state offices, the IQTIG could use CUSUM charts for in-depth analyses of deviating processes in the structured dialogues.

In March 2019, I was invited to present at the Medical Biometry and Statistics Unit of the IQTIG in Berlin. They are looking into the use of SPM for the use cases described here.

Publication

Hubig, L., Lack, N. & Mansmann, U.:
Statistical Process Monitoring to improve Quality Assurance of Inpatient Care. *BMC Health Service Research* 20(1):21. doi: 10.1186/s12913-019-4866-7 (2020).

Bibliography

1. Schrappe, M. *APS-Weißbuch Patientensicherheit* (ed Aktionsbündnis Patientensicherheit e.V. (APS)) (Medizinisch Wissenschaftliche Verlagsgesellschaft mbH & Co. KG, Copenhagen, 2018).
2. Kohn, L., M., C. J. & Donaldson, M. S. *(Institute of Medicine) To Err is Human: Building a Safer Health System* (National Academies Press, Washington (DC), 2000).
3. Busse, R., Nimptsch, U. & Mansky, T. Measuring, Monitoring, And Managing Quality In Germany's Hospitals. *Health Aff* **28,** w294–w304 (2009).
4. Gesetz zur Reform der Strukturen der Krankenhausversorgung (Krankenhausstrukturgesetz–KHSG). *Bundesgesetzblatt Jahrgang 2015* **Teil 1,** 2229–2253 (2015-12-17).
5. Shewhart, W. A. *Economic control of quality of manufactured product* (D. Van Nostrand Company, Inc, 1931).
6. Woodall, W. H. Bridging the Gap between Theory and Practice in Basic Statistical Process Monitoring. *Qual Eng* **29,** 2–15 (2017).
7. Steiner, S. H. & MacKay, J. Discussion of "Controversies and contradictions in statistical process control". *J Qual Tech* **32,** 370–372 (2000).
8. De Leval, M. R., François, K, Bull, C, Brawn, W & Spiegelhalter, D. Analysis of a cluster of surgical failures: Application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg* **107,** 914–924 (1994).
9. Steiner, S. H., Cook, R. J. & Farewell, V. T. Monitoring paired binary surgical outcomes using cumulative sum charts. *Stat Med* **18,** 69–86 (1999).
10. Grunkemeier, G. L., Wu, Y. X. & Furnary, A. P. Cumulative sum techniques for assessing surgical results. *Ann Thorac Surg* **76,** 663–667 (2003).
11. Smith, I. R. *et al.* Use of Graphical Statistical Process Control Tools to Monitor and Improve Outcomes in Cardiac Surgery. *Heart Lung Circ* **22,** 92–99 (2013).
12. Biau, D. J., Landreau, P. & Gravelau, N. Monitoring surgical performance: an application of industrial quality process control to anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc* **18,** 1263–1268 (2010).

13. Harris, J. R. *et al.* Risk-adjusted analysis of early mortality after ruptured abdominal aortic aneurysm repair. *J Vasc Surg* **42,** 387.e1–387.e6 (2005).

14. Snyder, J. J., Salkowski, N. & Wey, A. Developing CUSUM Charts for Monitoring Transplant Outcomes: Varied Goals and Many Possible Paths to Success. *Transplantation* **103,** 1760–1762 (2019).

15. Schuh, A., Canham-Chervak, M. & Jones, B. H. Statistical process control charts for monitoring military injuries. *Inj Prev* **23,** 416–422 (2017).

16. Bottle, A. & Aylin, P. Intelligent Information: A National System for Monitoring Clinical Performance. *Health Serv Res* **43,** 10–31 (2007).

17. Coory, M., Duckett, S. & Sketcher-Baker, K. Using control charts to monitor quality of hospital care with administrative data. *Int J Qual Health Care* **20,** 31–39 (2008).

18. Rasmussen, T. B., Pilgaard Ulrichsen, S. & Nørgaard, M. Use of risk-adjusted CUSUM charts to monitor 30-day mortality in Danish hospitals. en. *Clin Epidemiol* **10,** 445–456 (Apr. 2018).

19. Walker, A., Bacon, S, Croker, R & Goldacre, B. Detecting change in comparison to peers in NHS prescribing data: a novel application of cumulative sum methodology. *BMC Medical Inform Decis Mak* **18,** 62 (2018).

20. Drake, E. J., Coghill, J. & Sneyd, J. R. Defining competence in obstetric epidural anaesthesia for inexperienced trainees. *Br J Anaesth* **114,** 951–957 (2015).

21. Stringer, K. R., Bajenow, S. & Yentis, S. M. Training in airway management. *Anaesthesia* **57,** 967–983 (2002).

22. Chan, G. & Pautler, S. E. Use of cumulative summation (CUSUM) as a tool for early feedback and monitoring in robot-assisted radical prostatectomy outcomes and performance. *Can Urol Assoc J* **13.** doi:10.5489/cuaj.5350 (2018).

23. Roberts, H. W., Wagh, V. B., Sung, J., Ni, M. Z. & O'Brart, D. P. S. Risk-Adjusted CUSUM Analysis of the Learning Curve of Femtosecond Laser Assisted Cataract Surgery. *Curr Eye Res* **44,** 887–895 (2019).

24. Sood, A. *et al.* Application of the Statistical Process Control Method for Prospective Patient Safety Monitoring During the Learning Phase: Robotic Kidney Transplantation with Regional Hypothermia (IDEAL Phase 2a–b). *Eur Urol* **66,** 371–378 (2014).

25. Woodall, W. H. The Use of Control Charts in Health-Care and Public-Health Surveillance. *J Qual Technol* **38,** 89–104 (2006).

26. Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C. & Andrews, N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *J R Stat Soc Ser A Stat Soc* **175,** 49–82 (2011).

27. Tsui, K.-L., Wong, S., Jiang, W. & Lin, C.-J. Recent Research and Developments in Temporal and Spatiotemporal Surveillance for Public Health. *IEEE Trans Reliab* **60,** 49–58 (2011).

28. Jiang, W., Shu, L., Zhao, H. & Tsui, K.-L. CUSUM Procedures for Health Care Surveillance. *Qual Reliab Eng Int* **29,** 883–897 (2013).

29. Kass-Hout, T. A. *et al.* Application of change point analysis to daily influenza-like illness emergency department visits. *J Am Med Inform Assoc* **19,** 1075–81 (2012).

30. Baldewijns, G., Luca, S., Vanrumste, B. & Croonenborghs, T. Developing a system that can automatically detect health changes using transfer times of older adults. *BMC Med Res Methodol* **16.** doi:10.1186/s12874-016-0124-4 (2016).

31. Peeks, F. *et al.* Clinical and biochemical heterogeneity between patients with glycogen storage disease type IA: the added value of CUSUM for metabolic control. *J Inherit Metab Dis* **40,** 695–702 (2017).

32. Montgomery, D. C. *Introduction to statistical quality control* 6th ed. 734 pp. (Wiley, Hoboken, N.J, 2009).

33. Hubig, L., Lack, N. & Mansmann, U. Statistical Process Monitoring to Improve Quality Assurance of Inpatient Care. *BMC Health Serv Res* **20.** doi:10.1186/s12913-019-4866-7 (2020).

34. Gemeinsamer Bundesausschuss. *Richtlinie über Maßnahmen der Qualitätssicherung in Krankenhäusern / QSKH-RL* https://www.g-ba.de/downloads/62-492-1834/QSKH-RL_2019-03-22_iK-2019-06-01.pdf. Accessed: 2020-03-03. July 2019.

35. Rückle, F. & Stausberg, J. Entwicklung der Qualität der stationären Versorgung zwischen 2008 und 2011: Eine Studie aus Daten der externen Qualitätssicherung. *Dtsch Med Wochenschr* **140,** e36–e42 (2015).

36. Lack, N. & Gerhardinger, U. Wirksamkeitsanalyse externer Qualitätssicherungsmaßnahmen anhand von Veränderungen in Qualitätskennzahlen einzelner Krankenhäuser. *Z Evid Fortbild Qual Gesundhwes* **104,** 503–511 (2010).

37. IQTIG. *Methodische Grundlagen* https://iqtig.org/downloads/berichte/2017/IQTIG_Methodische-Grundlagen-V1.0.pdf. Accessed: 2020-03-03. Berlin, 2017.

38. IQTIG. *Karotis-Revaskularisation. Beschreibung der Qualitätsindikatoren für das Erfassungsjahr 2016* https://iqtig.org/downloads/auswertung/2016/10n2karot/QSKH_10n2-KAROT_2016_QIDB_V02_2017-04-26.pdf. Accessed: 2020-03-03. Berlin, 2017.

39. IQTIG. *Neonatologie. Beschreibung der Qualitätsindikatoren für das Erfassungsjahr 2016* `https://iqtig.org/downloads/auswertung/2016/neo/QSKH_NEO_2016_QIDB_V02_2017-04-26.pdf`. Accessed: 2020-03-03. Berlin, 2017.

40. IQTIG. *Hüftgelenknahe Femurfraktur mit osteosynthetischer Versorgung. Beschreibung der Qualitätsindikatoren für das Erfassungsjahr 2016* `https://iqtig.org/downloads/auswertung/2016/17n1hftfrak/QSKH_17n1-HUEFTFRAK_2016_QIDB_V02_2017-04-26.pdf`. Accessed: 2020-03-03. Berlin, 2017.

41. Page, E. S. Continuous Inspection Schemes. *Biometrika* **41,** 100–115 (1954).

42. Ewan, W. D. When and How to Use Cu-Sum charts. *Technometrics* **5,** 1–22 (1963).

43. Hawkins, D. M. Cumulative sum control charting: An underutilized SPC tool. *Qual Eng* **5,** 463–477 (1993).

44. Gan, F. F. An optimal design of CUSUM control charts for binomial counts. *J Appl Stat* **20,** 445–460 (1993).

45. Woodall, W. H. & Adams, B. M. The statistical design of CUSUM chart. *Qual Eng* **5,** 559–570 (1993).

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

47. Grigg, O. A., Farewell, V. & Spiegelhalter, D. Use of risk-adjusted CUSUM and RSPRTcharts for monitoring in medical contexts. *Stat Methods Med Res* **12,** 147–170 (2003).

48. Woodall, W. H., Fogel, S. L. & Steiner, S. H. The Monitoring and Improvement of Surgical-Outcome Quality. *J Qual Technol* **47,** 383–399 (2015).

49. Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C. & Gallivan, S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **350,** 1128–1130 (1997).

50. Poloniecki, J, Valencia, O & Littlejohns, P. Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *BMJ* **316,** 1697–1700 (1998).

51. Sherlaw-Johnson, C., Morton, A., Robinson, M. B. & Hall, A. Real-time monitoring of coronary care mortality: a comparison and combination of two monitoring tools. *Int J Cardiol* **100,** 301–307 (2005).

52. Wittenberg, P., Gan, F. F. & Knoth, S. A simple signaling rule for variable life-adjusted display derived from an equivalent risk-adjusted CUSUM chart. *Stat Med* **37,** 2455–2473 (2018).

53. Steiner, S. H. & Woodall, W. H. Debate: what is the best method to monitor surgical performance? *BMC Surg* **16,** 15 (2016).

54. Steiner, S. H., Cook, R. J., Farewell, V. T. & Treasure, T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* **1,** 441–452 (2000).

55. Cook, D. A., Duke, G., Hart, G. K., Pilcher, D. & Mullany, D. Review of the Application of Risk-adjusted Charts to analyze mortality outcomes in critical care. *Crit Care Resusc* **10,** 239–251 (2008).

56. Hawkins, D. M. & Wu, Q. The CUSUM and the EWMA Head-to-Head. *Qual Eng* **26,** 215–222 (2014).

57. Ung, N. & Wee, L. Cumulative sum method in nonzero fixed action level setup correction strategy: an application of statistical process control for targeted prostate radiotherapy. *Med Phys* **39,** 2746–53 (2012).

58. Kim, H.-J. *et al.* Monitoring the Quality of Robot-Assisted Pedicle Screw Fixation in the Lumbar Spine by Using a Cumulative Summation Test: *Spine* **40,** 87–94 (2015).

59. Reynolds, M. R. & Stoumbos, Z. G. A CUSUM Chart for Monitoring a Proportion When Inspecting Continuously. *Journal of Quality Technology* **31,** 87–108 (1999).

60. Brook, D & Evans, D. A. An approach to the probability distribution of cusum run length. *Biometrika* **59,** 539–549 (1972).

61. Knoth, S., Wittenberg, P. & Gan, F. F. Risk-adjusted CUSUM charts under model error. *Stat Med* **23,** 2206–2218 (2019).

62. Lucas, J. M. & Crosier, R. B. Fast Initial Response for CUSUM Quality-Control Schemes: Give Your CUSUM A Head Start. *Technometrics* **24,** 199–205 (1982).

63. Western Electric. *Statistical Quality Control Handbook* 2nd (Western Electric Co., Inc, Easton, Pennsylvania, 1958).

64. Zhang, X. & Woodall, W. H. Dynamic probability control limits for risk-adjusted Bernoulli CUSUM charts. *Stat Med* **34,** 3336–48 (2017).

65. Tang, X., Gan, F. F. & Zhang, L. Risk-Adjusted Cumulative Sum Charting Procedure Based on Multiresponses. *J Am Stat Assoc* **110,** 16–26 (2015).

66. Bersimis, S, Psarakis, S & Panaretos, J. Multivariate Statistical Process Control Charts: An Overview. *Qual Reliab Eng Int* **23,** 517–543 (2007).

67. Benjamini, Y & Kling, Y. *A Look at Statistical Process Control Through P-Values* Technical Report RP-SOR-99-08 (Tel Aviv University, Tel Aviv, Israel, 1999).

68. Grigg, O. A. & Spiegelhalter, D. J. An Empirical Approximation to the Null Unbounded Steady-State Distribution of the Cumulative Sum Statistic. *Technometrics* **50,** 501–511 (2008).

69. Mei, Y. Efficient scalable schemes for monitoring a large number of data streams. *Biometrika* **97,** 419–433 (2010).

70. Woodall, W. H. & Driscoll, A. R. in *Frontiers in Statistical Quality Control 11* (eds Knoth, S. & Schmid, W.) 15–27 (Springer International Publishing, Cham, 2015).

71. Jones, M. A. & Steiner, S. H. Assessing the effect of estimation error on risk-adjusted CUSUM chart performance. *Int J Qual Health Care* **24,** 176–181 (2012).

72. Tian, W., Sun, H., Zhang, X. & Woodall, W. H. The impact of varying patient populations on the in-control performance of the risk-adjusted CUSUM chart. *Int J Qual Health Care* **27,** 31–36 (2015).

73. Roberts, S. Control Chart Tests Based on Geometric Moving Averages. *Technometrics* **1,** 239–250 (1959).

74. Hubig, L. *cusum: Cumulative Sum (CUSUM) Charts for Monitoring of Hospital Performance* R package version 0.4.1 (2019).

75. R Core Team. *R: A language and environment for statistical computing* R Foundation for Statistical Computing (2019).

76. IQTIG. *IQTIGpvci: R Functions for Hospital Profiling* R package version 1.0.0 (2018).

77. Gehlenborg, N. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets.* R package version 1.4.0 (2016).

Appendix

1 R Package `cusum` vignettes

Following, two instructional vignettes published alongside the R package `cusum`* are presented. They show the main functions and application of the package:

Construct CUSUM charts for hospital performance The first vignette considers the construction and evaluation of ST-CUSUM and RA-CUSUM charts

GSCUSUM charts The second vignette desribes GSCUSUM charts, their construction and graphical representation using the `cusum` and `ggplot2` R packages.

*https://cran.rstudio.com/packages=cusum

Construct CUSUM charts for hospital performance

Overview

This vignette describes CUSUM charts based on a simulated false alarm probability for hospital performance data in the R package **cusum**. This is a practical guide to constructing and evaluating non-risk-adjusted and risk-adjusted CUSUM charts following Steiner et al. (Biostatistics 1.4 (2000), pp. 441-52).

The **cusum** packages takes different factors into account that influence the alarm rate of CUSUM charts. Some are given by the process to be monitored; these factors are:

- number of patients: How many observations do we expect in a monitoring period (e.g. a month/a year)?
- risk-adjustment: Is risk adjustment available? Can we allocate different risks to different observations?
  - if yes:
    - patient risks: What are these risks and how are they distributed?
  - if no:
    - accepted failure probability: What kind of failure rate do we expect on average?

The primary control input when constructing a CUSUM chart is the control limit. The control limit alarms performance deterioration once crossed by the cumulated sum.

The control limit depends on a number of variables:

- the desired target odds multiplier associated to an out-of-control process
- the accepted false alarm probability α
- number of simulations

Motivating example

To illustrate how **cusum** can be used for monitoring, we employ a simple and artificial data set generated to closely follow the performance data of German hospitals for one non-risk-adjusted performance indicator and one risk-adjusted performance indicator in 2016 and 2017.

risk-adj.	Indicator Description	Further explanation (in German)
NO	Ratio of observed to expected cases of severe stroke or death under open carotid stenosis surgery	pdf (p4)
YES	Preoperative stay more than 24 hours for patients with proximal femur fracture	pdf (p23)

Non-risk-adjusted performance indicator

```
data("cusum_example_data", package = "cusum")

head(cusum_example_data, 5)
#>   t     y year
#> 1 1 FALSE 2016
#> 2 2 FALSE 2016
#> 3 3 FALSE 2016
#> 4 4 FALSE 2016
#> 5 5 FALSE 2016
```

Risk-adjusted performance indicator

```
data("racusum_example_data", package = "cusum")

head(racusum_example_data, 5)
#>   t     y   score year
#> 1 1 FALSE 0.00237 2016
#> 2 2 FALSE 0.00237 2016
#> 3 3 FALSE 0.02412 2016
#> 4 4 FALSE 0.01893 2016
#> 5 5 FALSE 0.00725 2016
```

First, CUSUM charts are constructed on performance data from 2016 (Phase I), and then applied and evaluated on performance data from 2017 (Phase II).

```
cusum_example_p1 <- cusum_example_data[cusum_example_data$year == 2016, ]
cusum_example_p2 <- cusum_example_data[cusum_example_data$year == 2017, ]

racusum_example_p1 <- racusum_example_data[racusum_example_data$year == 2016, ]
racusum_example_p2 <- racusum_example_data[racusum_example_data$year == 2017, ]
```

Non-risk-adjusted CUSUM chart

Simulation of CUSUM Control Limits

We get the control limit of our CUSUM chart by simulating for a false alarm probability depending on sample size and accepted failure probability.

We can estimate the accepted failure probability by taking the average of Phase I. Alternatively, we could also define an accepted failure probability.

```
failure_probability <- mean(cusum_example_p1$y)

n_patients <- nrow(cusum_example_p1)
```

Then, control limits can be simulated using *cusum_limit_sim*.

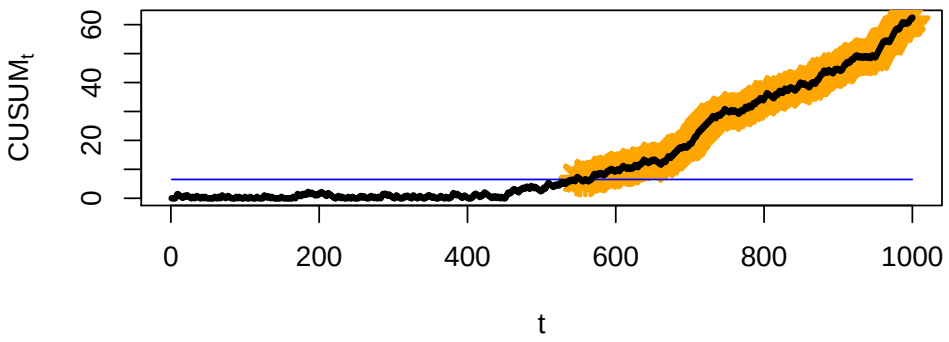
```
cusum_limit <- cusum_limit_sim(failure_probability,
                               n_patients,
                               odds_multiplier = 2,
                               n_simulation = 1000,
                               alpha = 0.05,
                               seed = 2046)
print(cusum_limit)
#> [1] 6.498476
```

Applying CUSUM Charts

CUSUM charts are applied on performance data from 2017 (Phase II) and the control limit *cusum_limit*. It can be calculated using *cusum*.

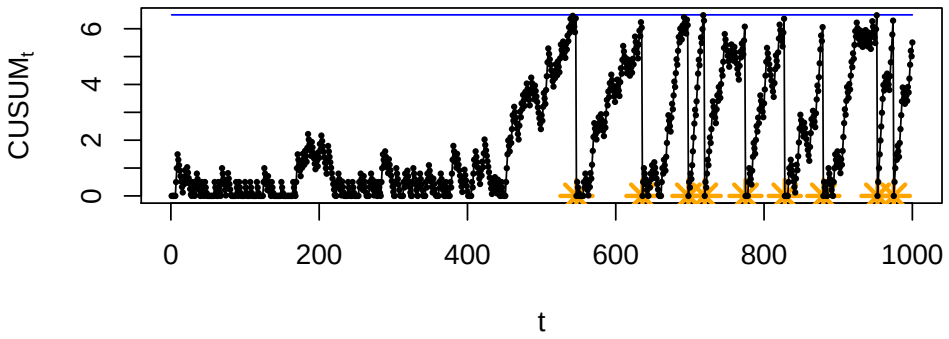
```
patient_outcomes <- cusum_example_p2$y
cusum_cs <- cusum(failure_probability,
                  patient_outcomes,
                  limit = cusum_limit,
```

```
                  odds_multiplier = 2,
                  reset = FALSE)
plot(cusum_cs)
```



Performance is as expected during the first half of monitoring, and then deteriorates. We get a alarm at t=547. If reset==TRUE, the CUSUM resets after each alarm.

```
cusum_cs <- cusum(failure_probability,
                  patient_outcomes,
                  limit = cusum_limit,
                  odds_multiplier = 2,
                  reset = TRUE)
plot(cusum_cs)
```



3

Evaluating CUSUM charts

The false alarm probability of a CUSUM chart can be simulated using *cusum_alpha_sim* given a predefined control limit.

```
n_patients <- nrow(cusum_example_p2)

cusum_alpha <- cusum_alpha_sim(failure_probability,
                               n_patients,
                               odds_multiplier = 2,
                               n_simulation = 1000,
                               limit = cusum_limit,
                               seed = 2046)
print(cusum_alpha)
#> [1] 0.05
```

We see that *cusum_alpha* equals our previously defined false alarm probability of 0.05.

Risk-adjusted CUSUM chart

Simulation of RA-CUSUM Control Limits

Control limits of RA-CUSUM charts are simulated for a false alarm probability depending on sample size and risk distribution.

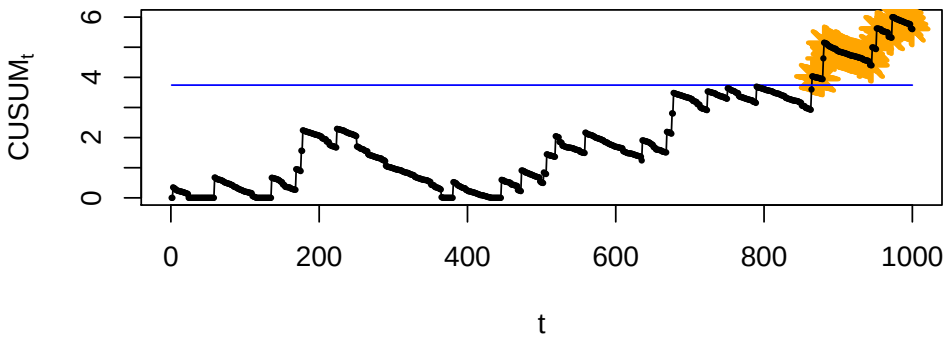
RA-CUSUM Control limits can be simulated using *racusum_limit_sim*.

```
patient_risks <- racusum_example_p1$score

racusum_limit <- racusum_limit_sim(patient_risks,
                                   odds_multiplier = 2,
                                   n_simulation = 1000,
                                   alpha = 0.05,
                                   seed = 2046)
print(racusum_limit)
#> [1] 3.742861
```

Applying RA-CUSUM charts

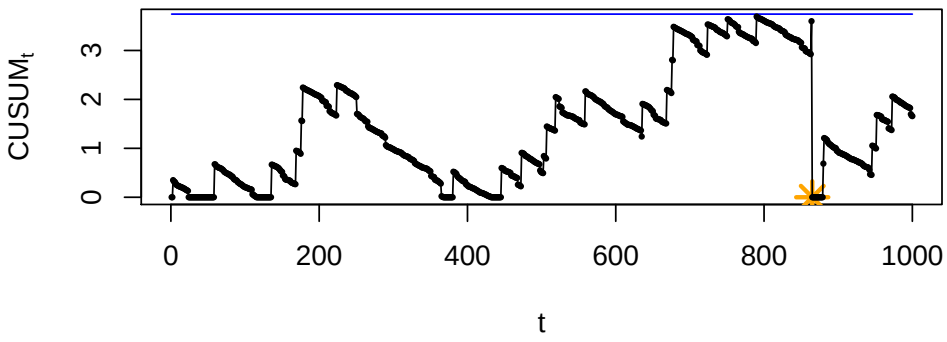
RA-CUSUM chart are applied on performance data from 2017 (Phase II) and the control limit *racusum_limit*. It can be calculated using *racusum*.

```
patient_risks <- racusum_example_p2$score
patient_outcomes <- racusum_example_p2$y

racusum_cs <- racusum(patient_risks,
                      patient_outcomes,
                      limit = racusum_limit,
                      odds_multiplier = 2,
                      reset = FALSE)
plot(racusum_cs)
```



Performance is as expected during the first half of monitoring, and then deteriorates. We get a alarm at t=865. If reset==TRUE, the CUSUM resets after each alarm.

```
racusum_cs <- racusum(patient_risks,
                      patient_outcomes,
                      limit = racusum_limit,
                      odds_multiplier = 2,
                      reset = TRUE)
plot(racusum_cs)
```



Evaluating RA-CUSUM charts

The false alarm probability of a CUSUM chart can be simulated using *cusum_alpha_sim*.

```
racusum_alpha <- racusum_alpha_sim(patient_risks,
                                   odds_multiplier = 2,
```

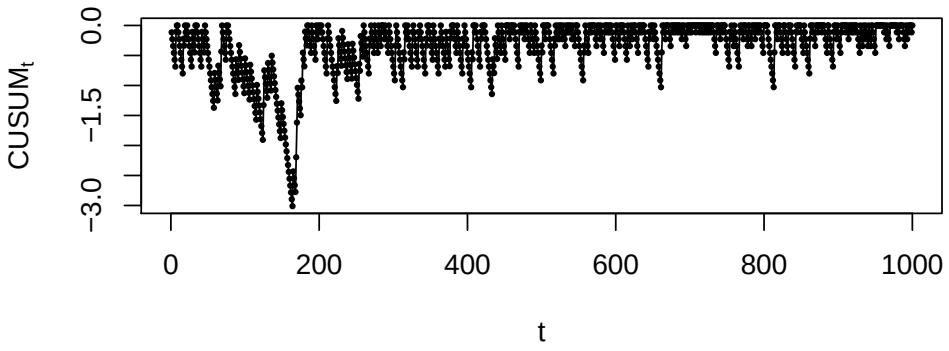
```
                                    n_simulation = 1000,
                                    limit = racusum_limit,
                                    seed = 2046)
print(racusum_alpha)
#> [1] 0.058
```

We see that *racusum_alpha* is similar to our previously defined false alarm probability of 0.05. Deviation is possible due to a slight change in risk population.

CUSUM Chart for process improvement

CUSUM charts for detecting process improvements can be constructed similarly, but the CUSUM statistic is restricted to non-positive values.

```
cusum_limit_improve <- cusum_limit_sim(failure_probability,
                                       n_patients,
                                       odds_multiplier = .5,
                                       n_simulation = 1000,
                                       alpha = 0.5,seed = 2046)

cusum_cs_improve <- cusum(failure_probability,
                          patient_outcomes = cusum_example_p2$y,
                          limit = cusum_limit_improve,
                          odds_multiplier = .5)

plot(cusum_cs_improve)
```



GSCUSUM charts

Overview

This vignette describes how to use GSCUSUM charts, an extension to standard CUSUM charts for binary performance data grouped in samples of unequal size.

Data preparation

Following information has to be available:

- patient-individual outcomes
- block-identifier in continuous sequence (can be obtained for example with `dplyr::group_indices()`)
- (patient-individual risk scores / risk of adverse event/failure)

These information are collected in a numeric matrix.

Non-risk-adjusted example data:

```
head(gscusum_example_data)
#> # A tibble: 6 x 4
#>       t y      year block_identifier
#>   <int> <lgl> <dbl>            <int>
#> 1     1 FALSE  2016                1
#> 2     2 FALSE  2016                1
#> 3     3 FALSE  2016                1
#> 4     4 FALSE  2016                1
#> 5     5 FALSE  2016                1
#> 6     6 FALSE  2016                1
```

Risk-adjusted example data:

```
head(ragscusum_example_data)
#> # A tibble: 6 x 5
#>       t y       score  year block_identifier
#>   <int> <lgl>   <dbl> <dbl>            <int>
#> 1     1 FALSE 0.00829  2016                1
#> 2     2 FALSE 0.00237  2016                1
#> 3     3 FALSE 0.00926  2016                1
#> 4     4 FALSE 0.00394  2016                1
#> 5     5 FALSE 0.0241   2016                1
#> 6     6 FALSE 0.00557  2016                1
```

Non-risk-adjusted GSCUSUM chart

Like in the standard CUSUM chart (see vignette for CUSUM charts), parameters have to be estimated in order to set up the charts,

```
failure_probability <- mean(gscusum_example_data$y[gscusum_example_data$year == 2016])

n_patients <- nrow(gscusum_example_data[gscusum_example_data$year == 2016,])
```

and control limits have to be estimated:

```
cusum_limit <- cusum_limit_sim(failure_probability,
                               n_patients,
                               odds_multiplier = 2,
                               n_simulation = 1000,
                               alpha = 0.05,
                               seed = 2046)


print(cusum_limit)
#> [1] 4.91128
```

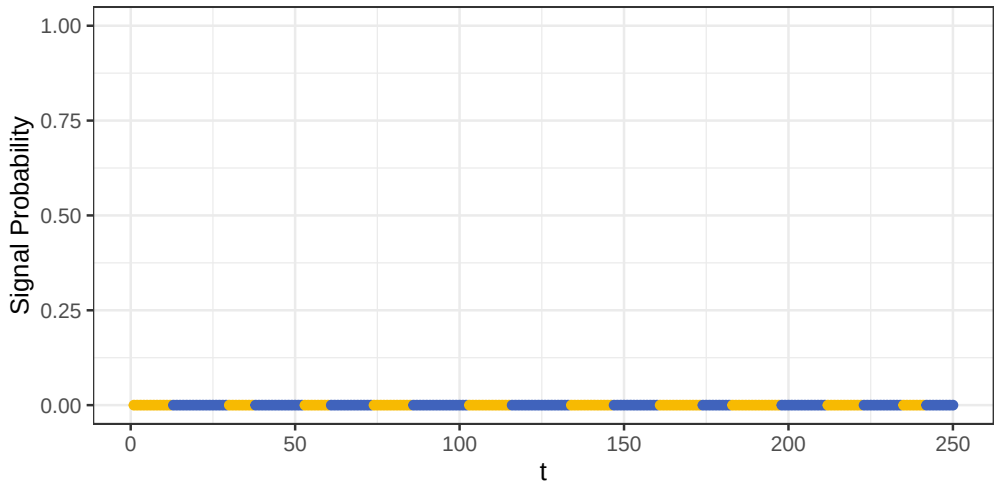
GSCUSUM charts are constructed on performance data from 2017.

```
gscusum_data <- gscusum_example_data[gscusum_example_data$year == 2017,]

input_outcomes <- matrix(c(gscusum_data$y, gscusum_data$block_identifier), ncol = 2)


gcs <- gscusum(input_outcomes = input_outcomes,
               failure_probability = failure_probability,
               odds_multiplier = 2,
               limit = cusum_limit,
               max_num_shuffles = 1000,
               quantiles = c(0.,0.05,0.25,0.5,0.75,.95,1))
```

This function returns the signal probability, average CUSUM values and quantiles of the CUSUM distribution specified in the function call.

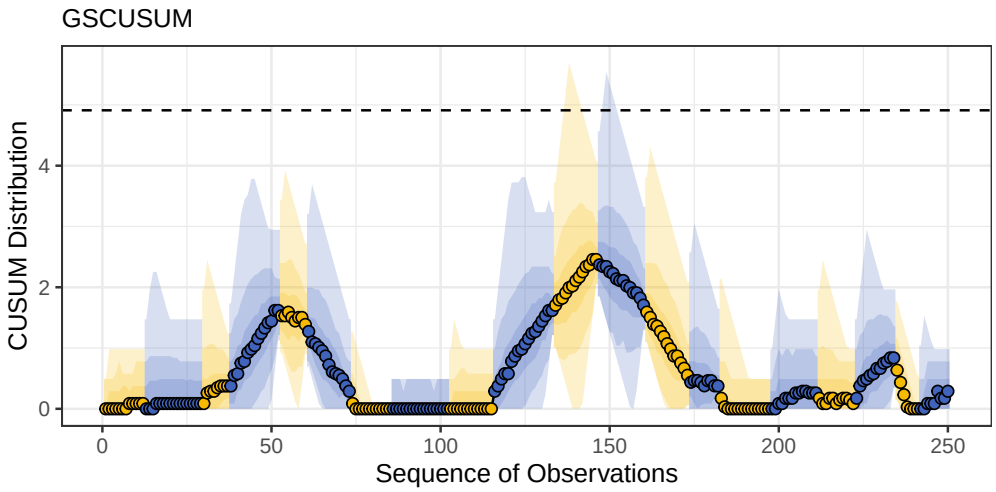
```
gcs <- as.data.frame(gcs)
names(gcs) <- c("sig_prob", "avg", "min", "q05", "q25", "median","q75","q95","max")
head(gcs)
#>   sig_prob        avg min q05 q25 median       q75       q95       max
#> 1        0 0.08494781   0   0   0      0 0.0000000 0.4910278 0.4910278
#> 2        0 0.13276181   0   0   0      0 0.2889099 0.4910278 0.9820557
#> 3        0 0.15132578   0   0   0      0 0.2889099 0.4910278 0.9820557
#> 4        0 0.15889569   0   0   0      0 0.2889099 0.5778198 0.9820557
#> 5        0 0.17285135   0   0   0      0 0.2889099 0.7799377 0.9820557
#> 6        0 0.18675470   0   0   0      0 0.3757019 0.5778198 0.9820557
```

```
gcs$block_identifier <- input_outcomes[,2]
gcs$t <- seq(1,nrow(gcs))

col1 <- "#f7ba02"
col2 <- "#4063bc"
palette <- rep(c(col1, col2), 300)

ggplot() +
  geom_line(data = gcs, aes(x = t, y = sig_prob)) +
  geom_point(data = gcs, aes(x = t, y = sig_prob, col = as.factor(block_identifier) )) +
  scale_color_manual(guide=FALSE, values = palette) +
  scale_y_continuous(name = "Signal Probability", limits = c(0,1))+
  theme_bw()
```



The complete run can be plotted with:

```
nblock <- max(gcs$block_identifier)

p <- ggplot(gcs)

for ( i in 1: nblock){
  dblock <- gcs[gcs$block_identifier == i,]
  col <- ifelse(i %% 2 == 0,col2,col1)
  dblock_before <- dblock[1,]
  dblock_before$t <- dblock_before$t - .5
  dblock_after <- dblock[nrow(dblock),]
  dblock_after$t <- dblock_after$t + .5
  dblock_n <- rbind(dblock, dblock_before, dblock_after)

  p <- p +
    geom_ribbon(data = dblock_n, aes(x = t, ymin = min, ymax = max), fill = col, alpha = 0.2) +
    geom_ribbon(data = dblock_n, aes(x = t, ymin = q05, ymax = q95), fill = col, alpha = 0.2) +
    geom_ribbon(data = dblock_n, aes(x = t, ymin = q25, ymax = q75), fill = col, alpha = 0.2)

}

p <- p +
  geom_line(data = gcs, aes(x = t, y = median)) +
  geom_point(data = gcs, aes( x = t, y = median, fill = as.factor(block_identifier)), size=2, pch = 21)·
  geom_hline(aes(yintercept = cusum_limit), linetype = 2) +
  theme_bw() +
  scale_y_continuous(name = "CUSUM Distribution") +
  scale_x_continuous(name = "Sequence of Observations") +
  scale_fill_manual(values = palette, guide = FALSE) +
  labs(subtitle = "GSCUSUM")
p
```



Risk-adjusted GSCUSUM chart

Like in the standard RA-CUSUM chart (see vignette for CUSUM charts), parameters are estimated in order to set up the charts,

```
n_patients <- nrow(ragscusum_example_data[ragscusum_example_data$year == 2016,])
```

and control limits are set:

```
racusum_limit <- racusum_limit_sim(patient_risks = ragscusum_example_data$score[ragscusum_example_data$
                            odds_multiplier = 2,
                            n_simulation = 1000,
                            alpha = 0.05,
                            seed = 2046)


print(racusum_limit)
#> [1] 2.403465
```

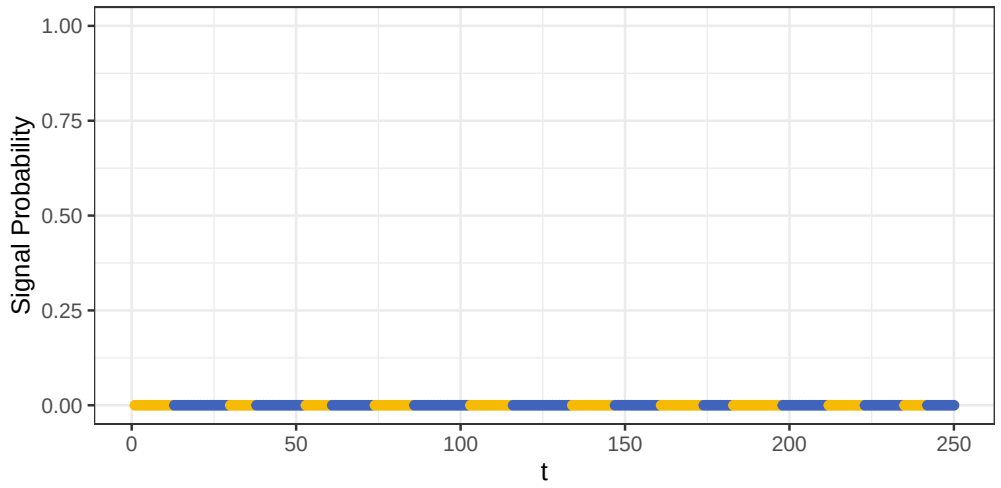
GSCUSUM charts are constructed on performance data from 2017.

```
ragscusum_data <- ragscusum_example_data[ragscusum_example_data$year == 2017,]

input_outcomes <- matrix(c(gscusum_data$y, gscusum_data$block_identifier), ncol = 2)


gcs <- gscusum(input_outcomes = input_outcomes,
               failure_probability = failure_probability,
               odds_multiplier = 2,
               limit = cusum_limit,
               max_num_shuffles = 1000,
               quantiles = c(0.,0.05,0.25,0.5,0.75,.95,1))
```

This function returns the signal probability, average CUSUM values and quantiles of the CUSUM distribution specified in the function call.

```
gcs <- as.data.frame(gcs)
names(gcs) <- c("sig_prob", "avg", "min", "q05", "q25", "median","q75","q95","max")
head(gcs)
#>   sig_prob        avg min q05 q25 median       q75       q95       max
#> 1        0 0.06776184   0   0   0      0 0.0000000 0.4910278 0.4910278
#> 2        0 0.12496124   0   0   0      0 0.2889099 0.4910278 0.9820557
#> 3        0 0.15704095   0   0   0      0 0.2889099 0.4910278 0.9820557
#> 4        0 0.16123550   0   0   0      0 0.2889099 0.5778198 0.9820557
#> 5        0 0.17368954   0   0   0      0 0.2889099 0.5778198 0.9820557
#> 6        0 0.18926572   0   0   0      0 0.3757019 0.5778198 0.9820557
```

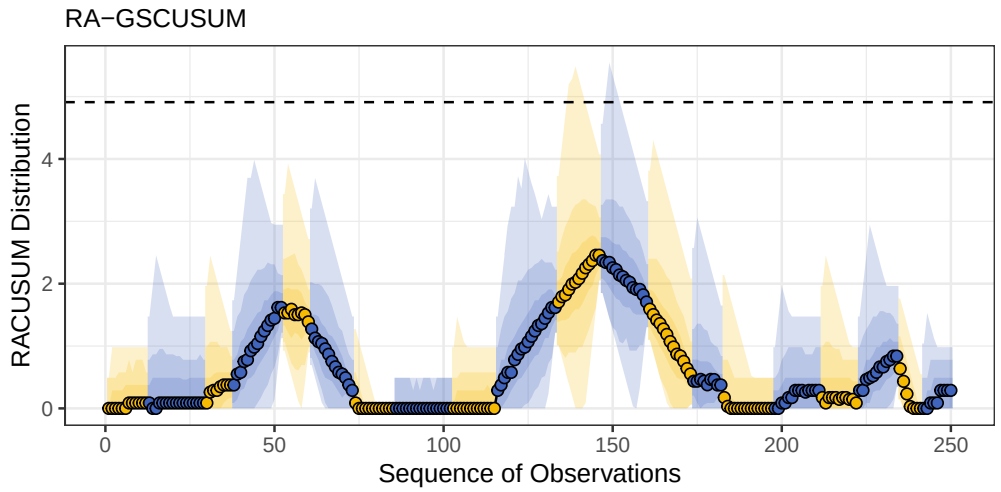
```
gcs$block_identifier <- input_outcomes[,2]
gcs$t <- seq(1,nrow(gcs))

col1 <- "#f7ba02"
col2 <- "#4063bc"
palette <- rep(c(col1, col2), 300)

ggplot() +
  geom_line(data = gcs, aes(x = t, y = sig_prob)) +
  geom_point(data = gcs, aes(x = t, y = sig_prob, col = as.factor(block_identifier) )) +
  scale_color_manual(guide=FALSE, values = palette) +
  scale_y_continuous(name = "Signal Probability", limit = c(0,1))+
  theme_bw()
```



The complete run can be plotted with:

```
nblock <- max(gcs$block_identifier)

p <- ggplot(gcs)

for ( i in 1: nblock){
  dblock <- gcs[gcs$block_identifier == i,]
  col <- ifelse(i %% 2 == 0,col2,col1)
  dblock_before <- dblock[1,]
  dblock_before$t <- dblock_before$t - .5
```

```
  dblock_after <- dblock[nrow(dblock),]
  dblock_after$t <- dblock_after$t + .5
  dblock_n <- rbind(dblock, dblock_before, dblock_after)

  p <- p +
    geom_ribbon(data = dblock_n, aes(x = t, ymin = min, ymax = max), fill = col, alpha = 0.2) +
    geom_ribbon(data = dblock_n, aes(x = t, ymin = q05, ymax = q95), fill = col, alpha = 0.2) +
    geom_ribbon(data = dblock_n, aes(x = t, ymin = q25, ymax = q75), fill = col, alpha = 0.2)

}

p <- p +
  geom_line(data = gcs, aes(x = t, y = median)) +
  geom_point(data = gcs, aes( x = t, y = median, fill = as.factor(block_identifier)), size=2, pch = 21)+
  geom_hline(aes(yintercept = cusum_limit), linetype = 2) +
  theme_bw() +
  scale_y_continuous(name = "RACUSUM Distribution") +
  scale_x_continuous(name = "Sequence of Observations") +
  scale_fill_manual(values = palette, guide = FALSE) +
  labs(subtitle = "RA-GSCUSUM")
p
```



2 R Code to evaluate hospital performance using CUSUM, GSCUSUM and conventional EQA

This code shows the evaluation of hospital performance using the `cusum` R package to construct CUSUM and GSCUSUM charts and the `IQTIGpvci` R package for conventional EQA performance evaluation.

The input data frame (`qdat`) consists of the following variables:

- `qi`: ID of performance indicator. Possible values are 11724 (risk-adjusted), 54030 and 51838 (non-risk-adjusted)
- `id`: Hospital ID
- `year`: Reporting year, used to select Phase I and Phase II data
- `y`: Patient outcomes
- `yhat`: Patient individual risk (`NA` in non-risk-adjusted performance indicators)
- `date`: Date of documentation (sequence of observations in CUSUM; block indices in GSCUSUM)
- `rho`: Target value of performance indicator

```
# select quality data for Phase I and Phase II ####
data_1 <- qdat[qdat$qi == qi & qdat$id == id & qdat$year == 2016,]
data_2 <- qdat[qdat$qi == qi & qdat$id == id & qdat$year == 2017,]
n_sim <- 100000
rho <- unique(data_1$rho[data_1$qi == qi]

 if (qi == 11724){
    # Risk-adjusted performance indicator ######################
    # Control Limit ####
    # estimate on Phase I data
    cl <- cusum::racusum_limit_sim(patient_risks = data_1$yhat,
                                odds_multiplier = 2,
                                alpha = 0.05,
                                n_simulation = n_sim,
                                seed = 7112015)
    # CUSUM ####
    cs <- cusum::racusum(patient_risks = data_2$yhat,
                      patient_outcomes = data_2$y,
                      limit = cl,
                      reset = FALSE)
    cs_signal <- ifelse(max(cs$signal) >0, 1,0)
    # GSCUSUM ####
    data_2$block <- data_2 %>% dplyr::group_indices(date) # group by date
    probability_ae <- data_2$yhat
    odds_multiplier <- 2
    ws <- log((1)/ (1 - probability_ae + odds_multiplier * probability_ae))
```

```
  wf <- log((odds_multiplier) / ((1 - probability_ae + odds_multiplier *
      probability_ae) * 1))
  gscs <- cusum::ragscusum(input_ra_outcomes = matrix(c(data_2$y,wf, ws, data_2
      $block), ncol = 4),
                           limit = cl,
                           quantiles = c(0,0.5,1),
                           max_num_shuffles = n_sim,
                           seed = 10082018)
  gscs <- as.data.frame(gscs)
  names(gscs) <- c("sig_prob", "avg", "min", "q50","max")
  gscs_signal <- ifelse(max(gscs$sig_prob) >.5, 1,0)
  # IQTIG ####
  ci <- IQTIGpvci::compute_oe_ci(o = sum(data_2$y), e = sum(data_2$yhat))
  arithm_dev <- ifelse(mean(data_2$y) >= data_2$rho, 1,0)
  stat_dev <- ifelse(ci$lower >= rho), 1,0)
} else if (qi == 54030 | qi == 51838){
   # Non-risk-adjusted performance indicator ######################
   # Control Limit ####
  cl <- cusum::cusum_limit_sim(failure_probability = rho,
                               n_patients = nrow(data_1),
                               odds_multiplier = 2,
                               alpha = 0.05,
                               n_simulation = n_sim,
                               seed = 7112015)
  # CUSUM ####
  cs <- cusum::cusum(failure_probability = rho,
                     patient_outcomes = data_2$y,
                     limit = cl,
                     reset = FALSE)
  cs_signal <- ifelse(max(cs$signal) >0,1,0)
  # GSCUSUM ####
  data_2$block <- data_2 %>% group_indices(date)
  gscs <- cusum::gscusum(input_outcomes = matrix(
                           c(data_2$y, data_2$block),
                           ncol = 2),
                         failure_probability = rho,
                         odds_multiplier = 2,
                         limit = cl,
                         quantiles = c(0,0.5,1),
                         max_num_shuffles = n_sim,
                         seed = 10082018)
  gscs <- as.data.frame(gscs)
  names(gscs) <- c("sig_prob", "avg", "min","q50","max")
  gscs_signal <- ifelse(max(gscs$sig_prob) > .5,1,0)
  # IQTIG ####
  ci <- IQTIGpvci::compute_rate_ci(o = sum(data_2$y), n = nrow(data_2))
  arithm_dev <- ifelse(mean(data_2$y) >= rho,1,0)
  stat_dev <- ifelse(ci$lower >= rho, 1,0)
```

Affidavit

I hereby declare, that the submitted thesis entitled:

Statistical Process Monitoring to
Improve Quality Assurance of Inpatient Care

is my own work. I have only used the sources indicated and have not made unauthorized use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

London, 5 October 2020	Lena Hubig
Place, date	Signature

Confirmation of Congruency

I hereby declare, that the electronic version of the submitted thesis entitled:

Statistical Process Monitoring to
Improve Quality Assurance of Inpatient Care

is congruent with the printed version both in content and format.

London, 5 October 2020	Lena Hubig
Place, date	Signature

Acknowledgements

I would like to thank all those who have contributed to the success of this project, in particular:

- Ulrich Mansmann for supervising this work, and for supporting and guiding me throughout my project.
- Rolf Holle for co-advising my thesis with helpful criticism and tireless optimism.
- Nicholas Lack for his ongoing support during my Master and Ph.D., and especially for introducing me to the subject of Statistical Process Control.
- The BAQ for taking me on as an intern and giving me access to their data.
- My colleagues at the IBE for creating a welcoming atmosphere to work in.
- The coordinators at MMRS - Epidemiology and Public Health for always being helpful and organising the retreats, and all Ph.D. students for being such a great group of people.
- My family for their love and encouragement, and especially my parents, Barbara and Jürgen, who made sure I had the chance to become everything I ever wanted.
- Claudius for countless useful comments and discussions, and moreover for being the best partner one can possibly imagine.