IISE Transactions on Healthcare Systems Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uhse21

Likelihood ratio-based CUSUM charts for real-time monitoring the quality of service in a network of queues

Yanqing Kuang, Devashish Das, Mustafa Sir & Kalyan Pasupathy

To cite this article: Yanqing Kuang, Devashish Das, Mustafa Sir & Kalyan Pasupathy (2023): Likelihood ratio-based CUSUM charts for real-time monitoring the quality of service in a network of queues, IISE Transactions on Healthcare Systems Engineering, DOI: 10.1080/24725579.2023.2181470

To link to this article: https://doi.org/10.1080/24725579.2023.2181470

Published online: 19 Apr 2023.

IISE TRANSACTIONS ON HEALTHCARE SYSTEMS ENGINEERING
https://doi.org/10.1080/24725579.2023.2181470

Likelihood ratio-based CUSUM charts for real-time monitoring the quality of service in a network of queues

Yanqing Kuang[a], Devashish Das[b], Mustafa Sir[b#], and Kalyan Pasupathy[c]

[a]University of South Florida, Tampa, FL, USA; [b]Amazon.com Inc, Seattle, WA, USA; [c]University of Illinois Chicago, Chicago, IL, USA

ABSTRACT

Queuing networks (QNs) are widely used stochastic models for service systems include healthcare systems, transportation systems, and computer networks. While existing literature has extensively focused on modeling and optimizing resource allocation in QNs, very little research has been done on developing systematic statistical monitoring methods for QNs. This paper proposes cumulative sum (CUSUM) control charts that monitor the queuing information collected in real-time from the QN. We compare the proposed methods with existing statistical monitoring methods to demonstrate their ability to quickly detect a change in the service rate of one or more queues at the nodes in the QN. Simulation results show that the proposed CUSUM charts are more effective than existing statistical monitoring methods. The motivation for this research comes from the need to monitor the performance of a hospital emergency department (ED) with the goal of monitoring delays experienced by patients visiting the ED. A case study using the data from the ED of a large academic medical center shows that proposed methods are a promising tool for monitoring the timeliness of care provided to patients visiting the ED.

KEYWORDS

Queuing networks; statistical process control; CUSUM charts; likelihood ratio tests

1. Introduction

A QN is the representation of a service system consisting of a network of servers. Each node of a QN consists of a set of servers processing or serving arriving entities, such as customers in a service system or packets in a computer network (Reiser & Kobayashi, 1975). In recent years, QNs have been widely used in modeling many service systems, such as manufacturing systems (Askin & Hanumantha, 2018), computer networks (Menascé & Bardhan, 2019), transportation systems (Hoshino et al., 2007; Roy et al., 2016), and healthcare systems (Armony et al., 2015; Li et al., 2017; Mehandiratta, 2011). Especially in healthcare, QN models have been found valuable in modeling the flow of patients in the hospital emergency departments (ED).

There is a rich body of research on resource allocation and decision-making in ED using QN models. For example, Cochran and Roche (2009) develop an open QN model to increase the capacity of an ED for patient care by considering various types of arrival patterns and volumes in patients. Vass and Szabo (2015) apply QN models to determine the optimal allocation of trained personnel and specialized equipment in ED. Huang et al. (2015) and Xie et al. (2016) present QN models that consider triage patients as a multi-class queuing system to control the priority of patients' treatments. They model the ED as a traditional queuing system, such as M/M/1 and M/M/k queues, where the service capacity is bounded and constant. However, healthcare systems such as ED have time-varying staff policy and availability for medical resources, which make traditional M/M/1 and M/M/k queuing systems a poor fit for ED. Therefore, Shi et al. (2021) adopt a processor-sharing queue, where the service rates are functions of the number of patients over time, to study how to effectively integrate a new diagnostic test into the clinical environment in ED. They demonstrate the processor-sharing queue with a state-dependent service rate function is more flexible to accommodate the complexities commonly seen in the ED environment comparing to the traditional queue setting.

Resource allocation using performance modeling tools like QNs addresses the problem of optimizing scare emergency medical resources. But such problems are typically part of long-term operational decision making in the ED. For example, changing the staffing schedules too frequently could be opposed by ED healthcare providers (Sir et al., 2017). Therefore, it is important to augment such performance modeling methods with real-time performance monitoring methods, which will ensure the adherence to a high quality of care and detect deterioration in the hypothesized optimal flow of patients. The signals from the performance monitoring methods can be used to activate reactive strategies, like swing shifts, where physicians shifts are extendable when the patient census is higher than expected to ensure adequate levels of service (Forsyth et al. 2018;

CONTACT Devashish Das ddas.wisc@gmail.com Amazon.com Inc., Seattle, WA, USA
#This work was done by the authors prior to joining Amazon.com Inc. and is not related to their roles at Amazon.com Inc.

Harrison et al. 2020; Hertzum 2021). Statistical process control (SPC) charts are increasingly being used in healthcare to monitor and measure the process variation and identify changes that indicate deterioration in quality (Woodall et al., 2012). It is important to note that the Center for Medicare and Medicaid Services requires that the hospitals report performance measures of the EDs, such as average length of stay of patients visiting the ED. Deterioration of these indicators can quickly bring down the quality of care. Research shows that lower service rates in ED can result in longer queue length and waiting time, which might increase the risk of adverse outcomes for patients (Johnson & Winkelman, 2011; Wen et al., 2020). Therefore, it is imperative to develop statistical performance monitoring methods for evaluating the quality of healthcare delivery in the ED.

Statistical process control (SPC) methods have been studied in the context of monitoring the quality of service in the ED (Mohammed, 2004). Salient examples using Shewhart-type control charts include the application of p-chart to monitor the variability of the number of patients leaving the ED (Kaminsky et al., 1997), $\bar{x}$-chart to monitor the door-to-reperfusion time for patients who have acute ST myocardial infarction (Callahan & Griffen, 2003), and run charts are developed to monitor the patient mortality rate (Rogers et al., 2008) and daily demand in order to identify the start and end of the winter surge of pediatric patients in ED (Pagel et al., 2018). Unlike the Shewhart-type charts depended on only the current observation, the charts based on CUSUM and EWMA schemes accumulate information from past observations. For example, Moran and Solomon (2013) implemented an EWMA chart to detect significant changes in the average number of deaths in the intensive care units of hospitals in Australia and New Zealand. Chen and Zhou (2015) developed advanced CUSUM charts for monitoring the performance of typical queuing systems with single queuing node. These methods focus on monitoring of specific quality indicators, such as the queue length of an individual queue, using univariate control charts. Service systems like the ED have a networked structure, so we cannot ignore the multidimensionality and granularity of the data obtained from electronic health records that can capture the delay experienced by patients at various stages of the care delivery process. Therefore, a multivariate statistical monitoring scheme based on advanced stochastic models like QNs is crucial and needs to be developed.

The most appropriate and widely used multivariate control charts are multivariate EWMA (MEWMA) and multivariate CUSUM (MCUSUM) charts. Their good performance in monitoring the changes of process means, especially for small changes, has been validated by many papers (Lee et al., 2015; Zou & Tsung, 2008). However, these methods assume the process data follow a time-homogeneous multivariate normal distribution. It needs to be clarified that, many large sample approximations of queue performance metrics, such as diffusion approximation, also follow multivariate normal distribution (Reiman, 1982). In practice, the normality assumption is usually difficult to justify for a real time queuing performance metric obtained from a nonstationary QN, so that the statistical properties of MEWMA and MCUSUM charts could be affected. In addition, we observed most of the existing papers focus on monitoring the queue length or waiting time in a service system modeled as a queue (Chen et al., 2011; Chen & Zhou, 2015; Qi et al., 2017; Shore, 2006; Zhao & Gilbert, 2015), limited attention has been paid of detecting the changes of the system parameters like service rate, which is the key factor that reflect the service ability of a service system like ED.

To overcome the limitation caused by the multivariate normal distribution assumption and fill the gap of monitoring the system parameters of a service system, this paper proposes new CUSUM charts based on the likelihood ratio statistics to monitor the service rates for a QN with time-inhomogeneous state-dependent queues. The likelihood ratio statistics pose no constraint to the underlying process distribution and have demonstrated to be generally more powerful than other alternative methods (Zhang, 2002). The proposed methods are evaluated based on the delay in detecting the change in service rate of one or more nodes of the QN. Our simulation results show that the proposed charts are more effective compared with conventional MCUSUM and MEWMA charts. Also, a real case study focusing on monitoring the daily patient flow of an emergency department demonstrates the efficacy of the proposed methods in real application.

The remainder of the paper is organized as follows. Section 2 introduces the QN model, and Section 3 introduces the statistical monitoring scheme for the QN. Section 4 derives the CUSUM charts based on different likelihood ratio statistics to monitor the service rate of QN. Their numerical performances are investigated in Section 5. In Section 6, we demonstrate the application of the proposed methods using a real-data example from the ED of a large academic medical center. Finally, the conclusions of this research and future research directions are described in Section 7.

2. Queuing network model

Consider an open network with I nodes where the service rate of the nodes depends on the number of customers at each node. In this network, external arrivals to node i, for $i \in \{1, 2, ..., I\}$, occurs as a Poisson process with rate $\lambda_i(t)$, where t denotes the time of a day. Let $B_i(t)$ denote the number of customers at node i at time t. We assume that the service rate of queue at node i follows an exponential distribution with rate $\mu_i(t) = f_i(B_i(t), \theta_i)$, where θ_i is the vector of parameters that define f_i. The arrival process and service time for each node are assumed to be mutually independent. The data from a queue network consisting of queues indexed by $i = 1, 2, ...I$ will have following events:

1. $\tau_i^1, \tau_i^2, ...\tau_i^{A_i(t)}$: The external arrivals to a queue node between times $[0, t]$ are independent of everything happening inside the network.
2. $\delta_i^1, \delta_i^2, ...\delta_i^{D_i(t)}$: The departures from a queue node between times $[0, t]$ are independent of everything except the number of customers at $\delta_i^1, \delta_i^2, ...\delta_i^{D_i(t)}$.

3. $e_i^1, e_i^2, \ldots e_i^{D_i(t)}$: The index of the queue that a customer leaving node i joins. $e_i^{n_i} = 0$ indicates that the n_ith customer leaving node i is either deterministic or dependent on the transition probabilities p_{ij}, where p_{ij} is the probability of a customer leaving node i to join node j and $p_{i0} = 1 - \sum_{j=1, j\neq i}^{I} p_{ij}$.

Since,

$$\mathbb{P}(t \leq \delta_i^{n_i} < t + dt || \text{ all events that have occurred on or before } t) = \mu_i(t)dt$$

for $n_i \in 1, 2, \ldots N_i(t)$ and

$$\mathbb{P}(t \leq \tau_i^{a_i} < t + dt || \text{ all events that have occurred on or before } t) = \lambda_i(t)dt$$

for $a_i \in 1, 2, \ldots A_i(t)$. The log likelihood of this data is given as

$$l(t, \Theta) = \sum_{i=1}^{I}\sum_{n_i=1}^{D_i(t)} \log \mu_i(\delta_i^{n_i}) - \int_0^t \mu_i(s)ds + \sum_{i=1}^{I}\sum_{a_i=1}^{A_i(t)} \log \lambda_i(\tau_i^{a_i}) - \int_0^t \lambda_i(s)ds + \sum_{i=1}^{I}\sum_{n_i=1}^{D_i(t)} \frac{\mathbb{I}(e_j^{n_i} = j)}{D_i(t)} \log p_{ij}.$$

Here $\Theta = [\theta_i]$, vector obtained from concatenating the service rate parameters from node i for $i = 1, \ldots I$. Let t_n denote the nth event in the ordered list of all arrivals $(\tau_i^{n_i})$ and departures $(\delta_i^{n_i})$, and let $l_n(\Theta) = l(t_n, \Theta)$. Thus,

Then, the log-likelihood function for the observed sample path in $(0, t_n]$ becomes

$$l_n(\Theta) = \sum_{i=1}^{I}\sum_{n_i=1}^{D_i(t_n)} \log \mu_i(\delta_i^{n_i}) - \int_0^{t_n} \mu_i(s)ds + \sum_{i=1}^{I}\sum_{a_i=1}^{A_i(t_n)} \log \lambda_i(\tau_i^{a_i}) - \int_0^{t_n} \lambda_i(s)ds + \sum_{i=1}^{I}\sum_{n_i=1}^{D_i(t_n)} \frac{\mathbb{I}(e_j^{n_i} = j)}{D_i(t_n)} \log p_{ij}. \quad (1)$$

In the following sections, instead of the likelihood ratio function, our proposed CUSUM charts are designed based on the log-likelihood function (1), which is both computationally easy and well-suited for the introduction of penalization methods.

3. Statistical monitoring scheme for QN

The likelihood function in (1) can be used to monitor any change in the QN. However, in practice, detecting deterioration in the performance of a QN caused by one or more queues at the nodes of the QN slowing down is more relevant than other types of changes. This detection problem is also the primary focus of the related research reviewed in Section 1. Therefore, we focus on building a statistical monitoring scheme to detect the change in the service rate of a QN. Let $\Theta_0 = \{\theta_1^0, \theta_2^0, \ldots, \theta_I^0\}$ represent parameters that define the service rate of each node in the QN when the system is in control, referred to as the in-control parameter, and let the parameter $\Theta = \{\theta_1, \theta_2, \ldots, \theta_I\}$ denote the true parameters corresponding to the service rate of each node. Hence, if the system is in control, then $\Theta_0 = \Theta$. The statistical monitoring scheme in this paper focuses on monitoring the QN only when an event such as arrival, departure, or movement of an entity from one node to another occurs. It is a framework also considered in prior research on monitoring single server queues (Chen et al., 2011). Therefore, the monitoring statistic is only updated at t_n, which represents the time when n^{th} event occurs.

Therefore, for each t_n, a test statistic h_n is defined to test the following hypothesis

$$H_0 : \Theta = \Theta_0$$
$$\text{vs.}$$
$$H_1 : \Theta \neq \Theta_0.$$

A decision rule is defined to test this hypothesis in a CUSUM scheme as follows:

$$h_n > g$$

where g is the threshold value. Once the CUSUM statistic h_n exceeds the control limit g, an alarm is triggered. A generated alarm means that the observed process is classified as out of control. Then the time t_n where such an out-of-control signal first happens is used to define the run length n. Here, the control limit g is determined such that the average run length (ARL) under the in-control scenario, denoted by ARL_0, meets the specified value (Knoth, 2006). The CUSUM statistic is defined in the next section.

4. Proposed CUSUM charts

The CUSUM chart, introduced by Page (1954), is one of the most popular sequential change-detection methods used in statistical quality control. It is based on not only current observations but also past observations. It has been demonstrated that the conventional CUSUM chart and its modifications are very effective in detecting a large class of change in model parameters (Chen & Zhou, 2015; Faisal et al., 2018; Sparks, 2000). Therefore, we develop CUSUM charts for monitoring the deterioration in the service rate of queues in a QN. In this section, CUSUM charts that are based on the likelihood ratio statistics are proposed to monitor the service rate of QNs. The first is a simple CUSUM (SCUSUM) chart, which is best suited when the practitioners can specify the potential out-of-control parameters. However, the performance of the SCUSUM chart might deteriorate if the real out-of-control parameters were far from the hypothesized out-of-control parameters. So, the general likelihood ratio and penalized likelihood ratio, which is computed by maximizing the likelihood ratio and penalized likelihood ratio respectively, are developed to construct the second type of CUSUM charts. They are called the

generalized CUSUM (G-CUSUM) chart and penalized CUSUM (P-CUSUM) chart.

4.1. The SCUSUM chart

For deriving the SCUSUM chart based on the likelihood ratio, a specified out of control parameter is needed. Let

$$\begin{aligned}\Theta_1 &= \{\theta_1, \theta_2, ..., \theta_I\} \\ &= \{(1+\Delta_1)\theta_1^0, (1+\Delta_2)\theta_2^0, ..., (1+\Delta_I)\theta_I^0\}\end{aligned}$$

represent the specified out of control service rates. Where Δ_i denotes the hypothesized degree of a shift away from in control parameter in μ_i^0. The sign of Δ_i should be consistent with the actual change direction of the service rate for each node $i = \{1, 2, ..., I\}$. For instance, if we are interested in detecting the decrease of the service rate for node i, then δ_i should be set as a negative value such as -10%.

Then, based on Eq. (1), we denote the log-likelihood ratio after the n^{th} event under the observed complete sample path $\{X(t_n)\}$ as ξ_n, which is

$$\begin{aligned}\xi_n &= l_n(\Theta_1) - l_n(\Theta_0) \\ &= \left[\sum_{i=1}^{I}\sum_{n_i=1}^{D_i(t_n)} \log\frac{f_i(B_i(\delta_i^{n_i}), \Theta_1)}{f_i(B_i(\delta_i^{n_i}), \Theta_0)} - \int_0^{t_n} \big(f_i(B_i(s), \Theta_1) - f_i(B_i(s), \Theta_0)\big)ds\right]\end{aligned} \tag{2}$$

Thus, the SCUSUM statistic h_n^s is defined as

$$h_n^s = \max\{0, h_{n-1}^s + \xi_n - \xi_{n-1}\} \qquad \text{where} \quad h_0^s = 0.$$

4.2. The G-CUSUM and P-CUSUM charts

The SCUSUM chart requires a set of specified design parameters, like Δ_i that indicate the type of change that the users want to detect. However, in practice, it is difficult for users to know the potential change in advance. For example, the user may need to specify the specific nodes of a QN that have the potential to slow down. For such cases, the SCUSUM chart may perform poorly when the actual change is different from the assumed change. As a solution to this problem, the specified service rate Θ_1 can be replaced by the maximum likelihood estimate (MLE) of the service rate for the server in each node by maximizing the log-likelihood ratio. The resulting CUSUM scheme results in a generalized-likelihood-ratio-based G-CUSUM chart.

Let τ_n^g denote the generalized log-likelihood ratio after the n^{th} event, which is the maximum of the log-likelihood ratio in (2), that is

$$\begin{aligned}\tau_n^g &= \max_{\Theta}\left[\sum_{i=1}^{I}\sum_{n_i=1}^{D_i(t)} \log\frac{f_i(B_i(\delta_i^{n_i}), \Theta)}{f_i(B_i(\delta_i^{n_i}), \Theta_0)} - \int_0^{t_n} \big(f_i(B_i(s), \Theta) - f_i(B_i(s), \Theta_0)\big)ds\right], \\ &= \sum_{i=1}^{I} \xi_{n,i}^g\end{aligned} \tag{3}$$

where

$$\xi_{n,i}^g = \max_{\Theta} \log\frac{f_i(B_i(\delta_i^{n_i}), \Theta)}{f_i(B_i(\delta_i^{n_i}), \Theta_0)} - \int_0^{t_n} \big(f_i(B_i(s), \Theta) - f_i(B_i(s), \Theta_0)\big)ds.$$

This maximization problem is simplified when f_i is a linear function in θ_i. Assume that f_i is functional linear model where:

$$f_i(B_i(t), \theta_i)) = \phi_i(B_i(t))^T\theta_i$$

where $\phi_i(B_i(t))^T$ is a vector-valued function of $B_i(t)$, which includes polynomial functions. Further, let $\Phi_{i,n} = \int_0^{t_n} \phi_i(B_i(s))ds$. Then,

$$\xi_{n,i}^g = \min_{\theta_i} \Phi_{i,n}^T(\theta_i - \theta_{i,0}) - \log\frac{\phi_i(B_i(\delta_i^{n_i}))^T\theta_i}{\phi_i(B_i(\delta_i^{n_i}))^T\theta_{i,0}}$$

which is a convex minimization problem and can be easily solved using gradient descent methods. Then, the G-CUSUM statistic h_n^g is then given as

$$h_n^g = \max\left\{0, h_{n-1}^g + \sum_{i=1}^{I}(\xi_{n,i}^g - \xi_{n-1,i}^g)\right\} \qquad \text{where} \quad h_0^g = 0.$$

The MLE-based likelihood ratio test can lead to poor change detection power when the dimensionality of Θ is large. To overcome this problem, a Lasso penalty term can be added for estimating the MLE of the service rate in each node, which reduces the dimensionality of changed parameters by inducing sparsity to the MLE in generalized likelihood ratio test (Zou et al., 2012; Zou & Qiu, 2009). In large QNs, it is reasonable to assume that only a few service rates will deviate from the in-control values. The Lasso method induces sparsity in the estimated Θ, and therefore increases the probability to select the θ_i that change.

Adding a penalty term to MLE-based statistic is equivalent to maximize the penalized log-likelihood ratio, which is as follows

$$\begin{aligned}\xi_n^\psi &= \max_{\Theta}\left[\sum_{i=1}^{I}\sum_{n_i=1}^{D_i(t)} \log\frac{f_i(B_i(\delta_i^{n_i}), \Theta)}{f_i(B_i(\delta_i^{n_i}), \Theta_0)} - \int_0^{t_n} \big(f_i(B_i(s), \Theta) - f_i(B_i(s), \Theta_0)\big)ds\right] \\ &\quad - \psi||\Theta - \Theta_0||_1 \\ &= \sum_{i=1}^{I} \xi_{n,i}^\psi,\end{aligned} \tag{4}$$

where I is the number of nodes in the QN, $i \in \{1, 2, ..., I\}$, and

$$\xi_{n,i}^\psi = \min_{\theta_i} \Phi_{i,n}^T(\theta_i - \theta_{i,0}) - \log\frac{\phi_i(B_i(\delta_i^{n_i}))^T\theta_i}{\phi_i(B_i(\delta_i^{n_i}))^T\theta_{i,0}} + \psi||\theta_i - \theta_{i,0}||_1.$$

Similarly, the test statistic for the P-CUSUM with a penalty ψ is defined as

$$h_n^{\psi} = \max\left\{0, h_{n-1}^{\psi} + \sum_{i=1}^{I}(\xi_{n,i}^{\psi} - \xi_{n-1,i}^{\psi})\right\} \quad \text{where} \quad h_0^{\psi} = 0.$$

It is worth noting that the derivations of ξ_n, ξ_n^g, ξ_n^{ψ} in Eqs (2–4) show that our proposed log-likelihood ratio based CUSUM statistic only require departure timestamps and number of customers in for each queue in a QN. In the implementation of the CUSUM chart, the optimization step converges in a few iterations and did not pose numerical issues.

5. Numerical study

In this section, the results of a simulation study to analyze the performance of the proposed CUSUM schemes in Section 4 are discussed. In the simulation experiments a QN with ten nodes will be examined, which is shown in Figure 1. Each node of the QN consists of a single server. It is important to note that the likelihood-ratio-based CUSUM schemes are agnostic to the number of servers in the queue. Entities arrive at the first node and depart the system from the last node. All the servers are independent of each other and their service times are exponentially distributed with the rate $\mu_i, i \in \{1, 2, ..., 10\}$. The in-control values of service rate for each node are all set to be 1.1, that is $\Theta_0 = [1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1]$. The external arrivals to the first node are according to a Poisson process with rate $\lambda = 1$. This QN is equivalent to ten connected M/M/1 queues.

Monte Carlo simulations are used to analyze the ARL performance of our proposed CUSUM methods: SCUSUM, G-CUSUM and P-CUSUM charts. The designed out-of-control parameters of the service rates for SCUSUM chart is set as $\Theta_1 = 0.9\Theta_0 = 0.9 * [1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1]$, which corresponds to a 10% decrease in service rates of all nodes. For the G-CUSUM, instead of a hypothesized change in service rate for each node, the MLE of the service rate for each node should be computed to obtain the test statistics. The control limits for all methods are all set such that the $ARL_0 = 100$.

We compare the performance of the proposed CUSUM scheme with two general multivariate SPC schemes: the multivariate CUSUM (MCUSUM) scheme (Pignatiello Jr and Runger, 1990) and the multivariate exponentially weighted moving average (MEWMA) chart (Lowry et al., 1992) to monitor the queue length for every t_n, the time when n^{th} event occurs. This is consistent with previous literature on monitoring queue length of single server queues (Chen & Zhou, 2015; Shore, 2006). Let $Q_n = [q_n^1, q_n^2, ..., q_n^{10}]^T$ denote the queue length for each of the ten nodes at t_n. The MEWMA and MCUSUM charts described here are meant to detect the change in service rate of the service nodes in the QN based on Q_n. The MEWMA test statistic for Q_n is defined as

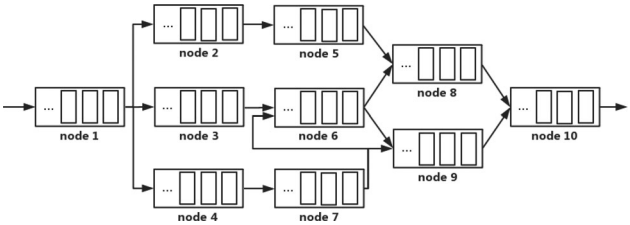

Figure 1. Structure of QN.

$$T_n^2 = \frac{2-\gamma}{\gamma} Z_n^T \Sigma^{-1} Z_n$$

where $\gamma \in (0, 1]$ is a weighting parameter and Z_n is a vector calculated in a recursive form

$$Z_n = \gamma(Q_n - Q^0) + (1-\gamma)Z_{n-1}$$

where $Z_0 = 0$, and Q^0 and Σ are the mean and covariance of queue lengths under in-control scenario, which are estimated from 10,000 simulations of Q_n under the in-control setting. The recommended values of γ is between 0.05 and 0.2 (Zou & Tsung, 2011). In the reported results $\gamma = 0.2$ was found to be the best for detecting small changes.

The MCUSUM statistic is defined as:

$$MC_n = \max\left\{0, \sqrt{D_n^T \Sigma^{-1} D_n} - \tilde{k}\omega_n\right\},$$

where

$$D_n = \sum_{i=n-\omega_n+1}^{n} (Q_i - Q^0)$$

and

$$\omega_n = \begin{cases} \omega_{n-1} + 1 & \text{if } MC_{n-1} > 0 \\ 1 & \text{otherwise} \end{cases}.$$

Also, following the recommendations in (Pignatiello Jr and Runger, 1990) and the $\Theta_1 = 0.9\Theta_0$ values, $\tilde{k} = 0.12$ was selected.

5.1. ARL comparisons for detecting the change of all nodes

Figure 2 presents the ARL comparisons for detecting the decrease in service rates of all nodes ranging from Θ_0 to $0.4\Theta_0$. The comparison shows that our proposed CUSUM charts significantly outperform the MEWMA and MCUSUM charts. Among them, the G-CUSUM is comparable to the SCUSUM in terms of the ARL performance and exhibits better sensitivity than the P-CUSUM chart in detecting the small changes of all service nodes. The Lasso penalty would force some of estimated service rates equal to their in-control values, which is not consistent with the fact that all service rates have been changed. Hence, P-CUSUM chart is less effective when all the nodes of the QN have slowed down.

On the other hand, when the actual service rates are much slower than the designed out-of-control parameter Θ_1, the performance of G-CUSUM chart deteriorates and SCUSUM chart is still sensitive. It is due to the fact that simple CUSUM charts are usually effective when actual change direction is similar to the hypothesized out-of-control change (Chen et al., 2011). On the other hand, when service rates decrease, the number of departure events decrease and the estimation error in G-CUSUM test statistic

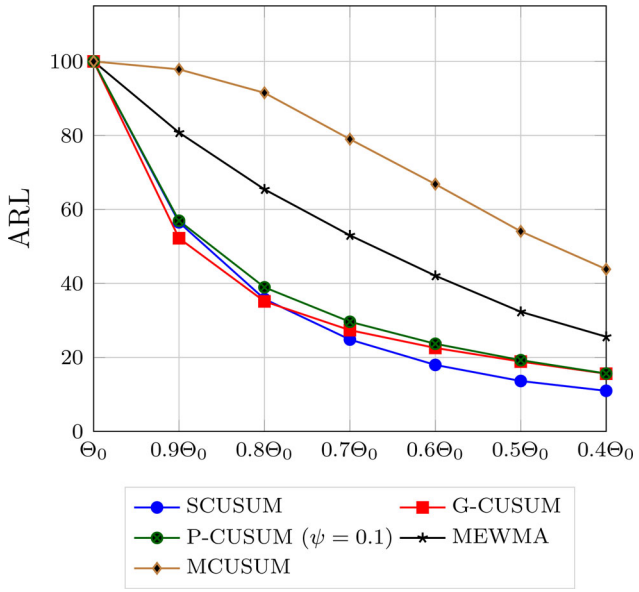

Figure 2. ARL comparisons in detecting the decrease of the service rates of all nodes.

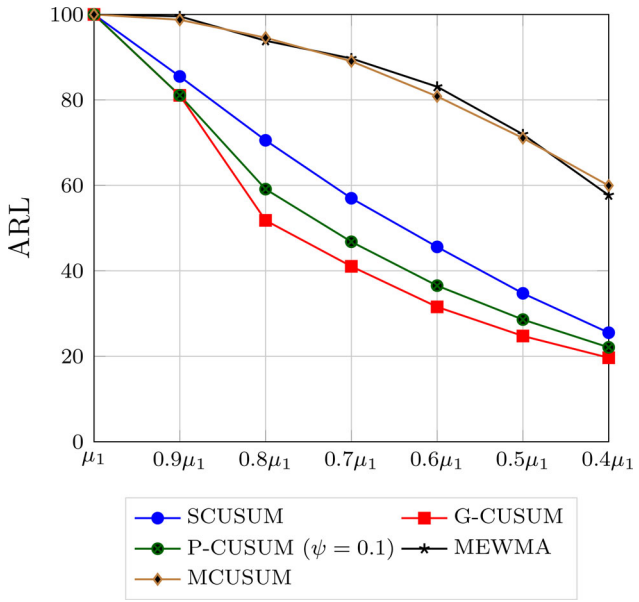

Figure 3. ARL comparisons in detecting the decrease of μ_1.

increases, which can explain the slight decrease in performance for smaller values of Θ observed in Figure 2.

5.2. ARL comparisons for detecting the change of single node

For detecting change of the single node in the QN, ARL comparisons are discussed for the first node, middle node and last node. Figures 3 and 4 show the ARL comparisons in detecting the decrease of the service rate of the first node and last node, respectively. Firstly, we demonstrate that the proposed CUSUM charts perform much better when compared to MEWMA and MCUSUM charts. Because the change in queue length of first node or last node significantly dominates that of other nodes when we only decrease the service rate of the first or last node,

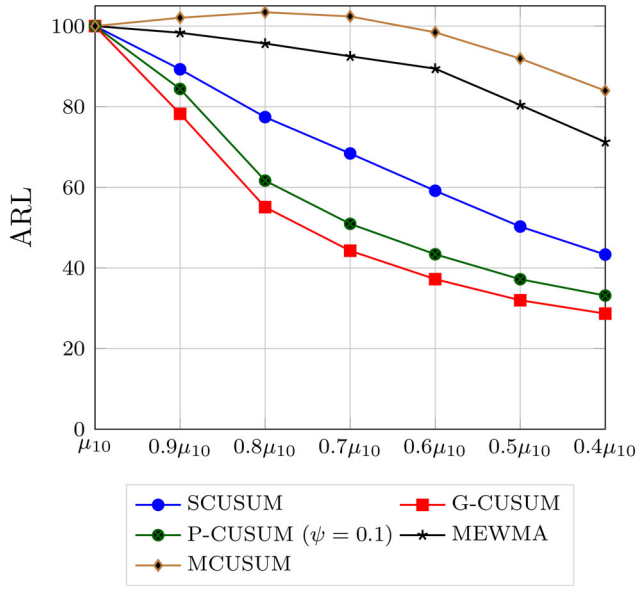

Figure 4. ARL comparisons in detecting the decrease of μ_{10}.

it makes MEWMA and MCUSUM less sensitive to the slowing in the service rate of other nodes. Among all the CUSUM charts, the G-CUSUM and P-CUSUM charts are more sensitive than SCUSUM chart in detecting any amount of decrease of the service rate for the first node and last node. However, this is not surprising. The designed out-of-control parameters of SCUSUM charts assume all the service rates have reduced, therefore its relatively poor ability to detect the change of service rate of a single node in comparison with G-CUSUM and P-CUSUM charts. Furthermore, it reveals that the G-CUSUM chart is more sensitive than the P-CUSUM chart. But the first and last node of the network are different than the other nodes. Change in service rate of either changes the performance of the whole QN. So, the need for detecting a sparse change, which is the goal in P-CUSUM chart, is not realized. Indeed the bias resulting from penalizing the likelihood could also impact the performance of P-CUSUM chart.

Figure 5 shows ARL comparisons for various monitoring schemes in detecting the decrease of the service rate of the fifth node, which located in the middle of the network. Again, Figure 5 shows that MEWMA and MCUSUM perform poorly. Also, G-CUSUM and P-CUSUM charts have better performance in detecting the change in service rate of a single node than SCUSUM. In addition, it is observed that the P-CUSUM chart is more effective in detecting a small decrease of the service rate in fifth node and the G-CUSUM chart exhibits better sensitivity in detecting the moderate and large decreases in service rate of a single node. The latter observation is consistent with the findings in Figures 3 and 4. Therefore, if the objective is to detect a small change in the service rate of a single node using a small sized sample, adding a penalty term is recommended.

5.3. Identity the exact out-of-control node using penalized CUSUM chart

Traditional multivariate SPC scheme like MEWMA and MCUSUM control chart statistics are computed based on

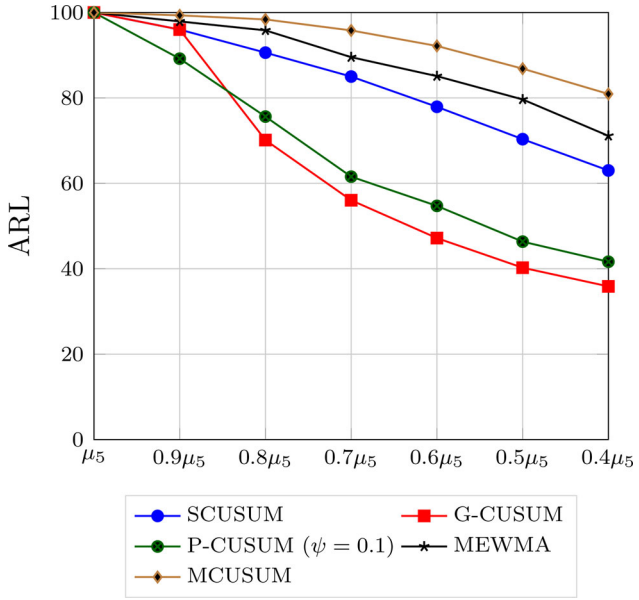

Figure 5. ARL comparisons in detecting the decrease of μ_5.

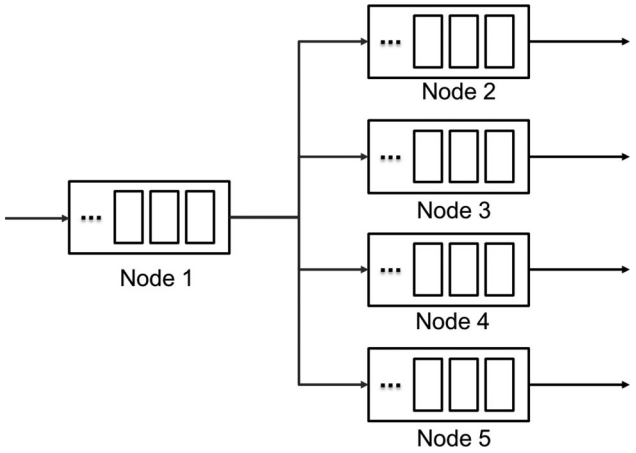

Figure 6. Parallel queuing Network.

the covariance matrix in data, they can be used to detect the potential change for multivariate process but not to identify which variate has changed, but the latter is more important for quality control practitioners. Thus, our proposed penalized CUSUM charts can overcome the limitation to identify the exact out of control node when only single node in a queuing network has changed. The designed penalized CUSUM charts can return us a set of estimated departure rates for each node that has potentially changed, then the node with the minimal estimated departure rate is signaled as out of control. In order to verify the efficiency of penalized CUSUM chart, a parallel network with 5 nodes is examined, which is shown in Figure 6.

Table 1 is the accuracy of P-CUSUM for identifying node 2 as out-of-control if only node 2 has decreased. It shows that the accuracy is increasing when the degree of change for node 2 is increasing. The accuracy is defined as the probability of identifying node 2 as out-of-control within all the five nodes during a process.

Table 1. The accuracy of P-CUSUM for identifying node 2 as out-of-control if only node 2 has decreased.

Actual change of μ_2	Accuracy of P-CUSUM
$0.9\mu_2$	30%
$0.8\mu_2$	40%
$0.7\mu_2$	52%
$0.6\mu_2$	65%
$0.5\mu_2$	76%
$0.4\mu_2$	85%

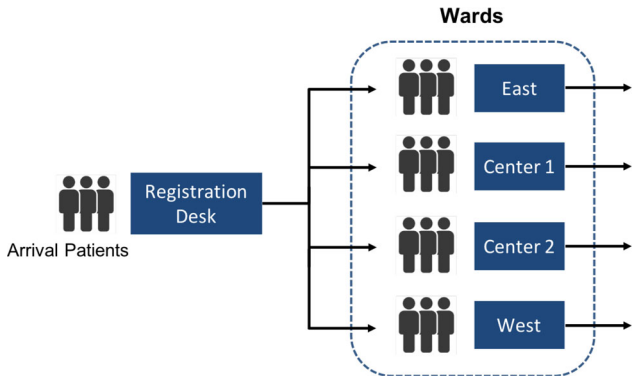

Figure 7. Patient visit flow of the emergency department (ED) of a large academic medical center.

6. Case study: monitoring the flow of patients in an ED

In this section, the proposed monitoring schemes are evaluated to monitor the flow of patients in the emergency department of Mayo Clinic. The patient flow in the ED is modeled as a QN with five nodes, which are registration desk and four clusters of beds with a team of healthcare providers in each cluster. They are called East, Center 1, Center 2 and West nodes. These correspond to the distinct pods in the ED from where the data is collected and illustrated in Figure 7. Patients visiting the ED wait until they are assigned to different wards. The patients leave the ED (admitted to the hospital wards or discharged) after been served in the wards. We are interested in monitoring the service rate of these five nodes, shown in Figure 7.

Here we assume each service node in our partner ED is a processor-sharing queue with a state-dependent service rate function. Processor sharing queue is a model in which the available service capacity is shared by the number of customers presented in queue. The ED is a complex service environment with many shared resources (nurses, doctors, equipment, hallways, laboratory, etc) and multitasking situations, which are conceptually similar to queuing models with shared processors. Thus, the processor-sharing queue is more flexible to accommodate these complexities commonly seen in the ED service environment compared to traditional queuing systems. Other papers also have considered it for similar reasons (Armony et al., 2015; Shi et al., 2021; Whitt & Zhang, 2017). Figure 8 illustrates an example for the empirical distribution of the patient occupancy levels in one of the service nodes at our partner ED of Mayo Clinic. The occupancy level at a given time represents the total number of patients in the node. We find that assuming a processor-sharing queue with a state-dependent service rate for our

partner ED can best replicate the empirical occupancy distribution curve compared to the conventional M/M/1 queue, which clearly deviates from the empirical distribution.

We select the first 183 days in Year 2016 as training data to estimate the in-control departure rate using model fitting methods. We adopted a linear form with $\mu_i(t) = \theta_i f_i(n_i(t))$ to define the service rate of node i in ED, where $n_i(t)$ denotes the number of patients in node i at time t and θ_i is a parameter corresponding to f_i, and f_i represents a transform or function of $n_i(t)$ such as the logarithm, square root, square and cube of $n_i(t)$. Then, 10-fold cross validation (CV) method is applied on the in control data to select the best model. Table 2 is the CV errors for different models. Among different models, we can see that Model 3 (a linear function with respect to the square root of number of patients) produces the minimum errors, which is then chosen as the best model to explain the relationship between total service rate and the number of patients for all nodes. As a result,

$$\mu_1(t) = 203.1\sqrt{n_1(t)}, \mu_2(t) = 49.93\sqrt{n_2(t)}, \mu_3(t) = 47.43\sqrt{n_3(t)},$$
$$\mu_4(t) = 51.2\sqrt{n_4(t)}, \mu_5(t) = 51.61\sqrt{n_5(t)},$$

We use the last 183 days in Year 2016 as the test data set. The control limits are set as the 90% percentile of the test statistics for the training data. The proposed CUSUM charts are used to detect the decrease in the service rates and then the MEWMA and MCUSUM charts are used to compare with proposed CUSUM charts. The number of days classified as in control and out of control are presented in the confusion matrices in Tables 3–6. Table 3 shows that there are total 51 days in the test dataset are labeled as out of control by SCUSUM chart, while 36 days among the testing set are signed as out of control by MEWMA chart, in which 21 days are identified as out of control by both SCUSUM and MEWMA charts, and 30 days are identified as out of control by SCUSUM chart only. Similar results for the comparisons for P-CUSUM and MEWMA charts are given in Table 4, SCUSUM and MCUSUM charts are given in Table 5, and P-CUSUM and MCUSUM charts are given in Table 6.

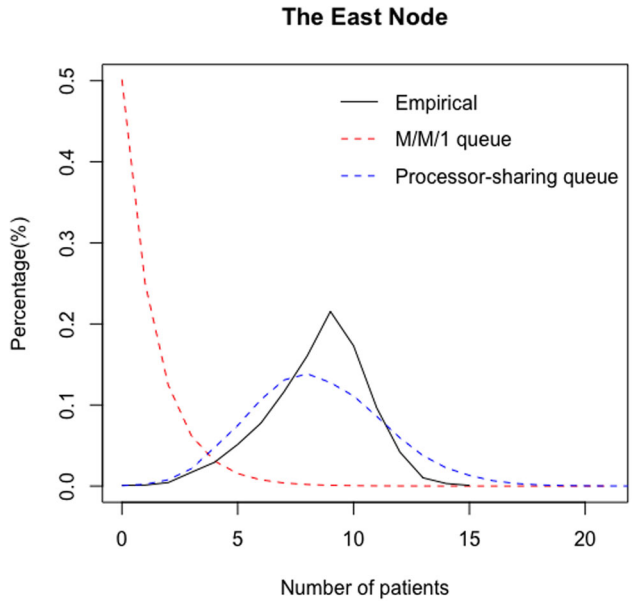

Figure 8. Histogram of patient occupancy of the east node in ED of Mayo Clinic. The x-axis is the state namely the number of patients in the east node, and the y-axis corresponds to the frequency of the state.

In the test dataset, there are 15 days signaled out-of-control by both SCUSUM and P-CUSUM but not the MEWMA and MCUSUM charts. To get further insight into the reason for this, we study October 20, 2016 in further detail. Our methods found there is an overall decrease in service rate for all the nodes on Oct 20. However, the east node has decreased dramatically compared to other nodes. Then we compare the actual departure rate on Oct 20 with the in-control departure rate for the east node in Figure 9. This figure shows the departure rate has clearly dropped at every time of the day on Oct 20. The departure rate is fitted using a nonparametric functional method based on kernel density fitting process proposed by Wu et al. (2013).

Table 7 shows actual average queue length, and their in-control values for each node on October 20, 2016. We can observe that, except for the registration node, the average queue length of all the other nodes on October 20 just slightly deviated from the in-control average value, hence, the MEWMA and MCUSUM test statistics were not able to distinguish them as out of control. However, as shown in Figure 9, the service rate for the east node had clearly decreased on October 20, and we are able to detect it using the proposed methods.

The analysis of the real data leads to an important conclusion that monitoring the service rate in the ED is needed. It also shows that traditional performance measures of queuing system such as queue length often are unable to reflect the service ability in ED. The result can assist operations managers to improve the timeliness of care in the ED. The

Table 3. Confusion matrix for SCUSUM and MEWMA charts.

		SCUSUM		
	$N = 183$	Out of control	In control	Total
MEWMA	Out of control	21	15	36
	In control	30	117	147
	Total	51	132	

Table 2. CV errors for different models.

	Model 1 $\mu_i(t) = \theta_i n_i(t)$	Model 2 $\mu_i(t) = \theta_i \log(n_i(t))$	Model 3 $\mu_i(t) = \theta_i\sqrt{n_i(t)}$	Model 4 $\mu_i(t) = \theta_i n_i^2(t)$	Model 5 $\mu_i(t) = \theta_i n_i^3(t)$
Node 1	755.9	746.7	**702.6**	796.6	811.8
Node 2	299.2	299.2	**299.2**	301.1	304.4
Node 3	293.2	293.2	**293.2**	295.1	298.6
Node 4	290.6	289.9	**289.7**	294.6	298.8
Node 5	336.3	336.2	**336.2**	339.6	345.4

Table 4. Confusion matrix for P-CUSUM and MEWMA charts.

		P-CUSUM		
	$N = 183$	Out of control	In control	Total
MEWMA	Out of control	20	16	36
	In-control	26	121	147
	Total	46	137	

Table 5. Confusion matrix for SCUSUM and MCUSUM charts.

		SCUSUM		
	$N = 183$	Out of control	In control	Total
MCUSUM	Out of control	33	26	59
	In control	18	106	124
	Total	51	132	

Table 6. Confusion matrix for P-CUSUM and MCUSUM charts.

		P-CUSUM		
	$N = 183$	Out of control	In control	Total
MCUSUM	Out of control	29	30	59
	In control	17	107	124
	Total	46	137	

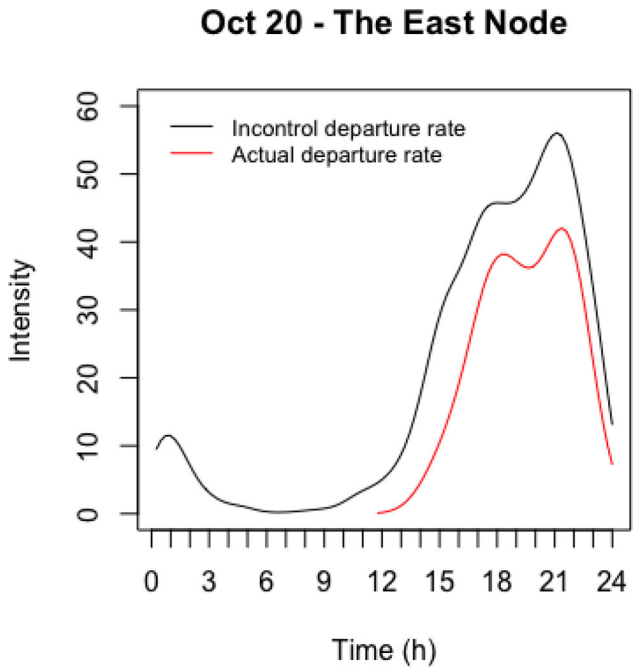

Figure 9. Departure intensity comparison on Oct 20, which was signaled out-of-control both SCUSUM and P-CUSUM but not the MEWMA and MCUSUM charts.

proposed methods can be potentially used in two ways: (1) Identify repeated patterns of inefficiency and reevaluate strategic decisions. If specific hours of a day repeatedly experience slow in service rate, the ED can reevaluate staffing schedules and allocate more clinician hours to those hours; (2) React in real time by increasing staffing levels. EDs may utilize swing shifts to increase staffing levels temporarily to meet patient demand. These decisions are currently rule based, and there is a lack of research in development of systematic methods to activate such responses. Having a real-time monitoring method can help in timely activation of such decisions on reactive responses and avoid unnecessary reactions. Alternative methods like using Markov Decision

Table 7. The average queue length comparisons on October 20.

		Registration	East	Center 1	Center 2	West
	October 20	7.8	3.17	8.32	9.55	9.64
Ave. Queue Length	In-control	4.31	4.25	8.67	8.02	9.68

Processes models for evaluating the cost of such decisions can be prohibitive to build for all EDs as they require construction of models and developing efficient algorithms. The proposed methods use easily available timestamp data and relatively simple calculation of test statistics.

7. Conclusion

In this paper, we propose new CUSUM control charts based on count data to monitor the service rate of a QN with state-dependent queues. The proposed CUSUM charts are compared with the MEWMA and MCUSUM charts using the ARL criteria to detect out-of-control scenarios. A major contribution of this research is the development of an easy to implement and efficient likelihood-ratio-based CUSUM charts, G-CUSUM and P-CUSUM charts for monitoring QNs, which could overcome the limitation of the normality assumption and do not need to know the potential change in service rate of the queuing nodes in a QN, and thus have important practical applications. Numerical studies based on a simulated QN demonstrated that the proposed CUSUM charts can outperform traditional approaches on a variety of out-of-control scenario detection tests. Further, a case study focusing on monitoring the daily patient flow of an ED demonstrates the efficacy of the proposed methods in a real application.

There are several extensions of the methods developed in this paper. The current monitoring scheme is based on the likelihood ratio statistic, which requires the sample path can be observed completely. However, there are some challenges associated with obtaining the likelihood ratio statistic when only limited and partial samples can be observed. Generalizations and extensions of this method to study problems such as changes in optimal routing policies and dependence on factors external to an ED that can cause delays in the ED are part of our ongoing research. This would lead to development of real-time monitoring methods that evaluate the cost of making such decisions. Another important extension of this paper could involve the study of approximation methods in establishing theoretical understanding of statistical monitoring of QNs. Specifically, the application of diffusion approximation methods can help establish theoretical performance guarantees of CUSUM methods developed here. In addition, other than the real case application in ED, monitoring the patient flows in other units, such as the intensive care units is important for further methodological development and application of the proposed methods in quality control of healthcare systems.

Consent and approval

This study has been exempt from the requirement for approval by an institutional review board. We have only used secondary data.

Disclosure statement

The authors report no conflict of interest.

Funding

No specific funding was received for this work.

ORCID

Yanqing Kuang http://orcid.org/0000-0001-9229-6678
Kalyan Pasupathy http://orcid.org/0000-0002-4760-2805

References

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., & Yom-Tov, G. B. (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, *5*(1), 146–194. https://doi.org/10.1287/14-SSY153

Askin, R. G., & Hanumantha, G. J. (2018). Queueing network models for analysis of nonstationary manufacturing systems. *International Journal of Production Research*, *56*(1-2), 22–42. https://doi.org/10.1080/00207543.2017.1398432

Callahan, C. D., & Griffen, D. L. (2003). Advanced statistics: Applying statistical process control techniques to emergency medicine: A primer for providers. *Academic Emergency Medicine*, *10*(8), 883–890. https://doi.org/10.1197/aemj.10.8.883

Chen, N., & Zhou, S. (2015). CUSUM statistical monitoring of m/m/1 queues and extensions. *Technometrics*, *57*(2), 245–256. https://doi.org/10.1080/00401706.2014.923787

Chen, N., Yuan, Y., & Zhou, S. (2011). Performance analysis of queue length monitoring of m/g/1 systems. *Naval Research Logistics (NRL)*, *58*(8), 782–794. https://doi.org/10.1002/nav.20483

Cochran, J. K., & Roche, K. T. (2009). A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research*, *36*(5), 1497–1512. https://doi.org/10.1016/j.cor.2008.02.004

Faisal, M., Zafar, R. F., Abbas, N., Riaz, M., & Mahmood, T. (2018). A modified CUSUM control chart for monitoring industrial processes. *Quality and Reliability Engineering International*, *34*(6), 1045–1058. https://doi.org/10.1002/qre.2307

Forsyth, K. L., Hawthorne, H. J., Cammon, W. D., Linden, A. R., & Blocker, R. C. (2018). Perceived workload and an automated workload alert system: A comparison in the emergency department. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 62, pp. 573–577). SAGE Publications Sage CA. https://doi.org/10.1177/1541931218621131

Harrison, E. M., Walbeek, T. J., Maggio, D. G., Herring, A. A., & Gorman, M. R. (2020). Circadian profile of an emergency medicine department: Scheduling practices and their effects on sleep and performance. *The Journal of Emergency Medicine*, *58*(1), 130–140. https://doi.org/10.1016/j.jemermed.2019.10.007

Hoshino, S., Ota, J., Shinozaki, A., & Hashimoto, H. (2007). Optimal design methodology for an agv transportation system by using the queuing network theory. In R. Alami, R. Chatila, & H. Asama (Eds.), *Distributed autonomous robotic systems* (Vol. 6, pp. 411–420). Springer.

Huang, J., Carmeli, B., & Mandelbaum, A. (2015). Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, *63*(4), 892–908. https://doi.org/10.1287/opre.2015.1389

Johnson, K. D., & Winkelman, C. (2011). The effect of emergency department crowding on patient outcomes: A literature review. *Advanced Emergency Nursing Journal*, *33*(1), 39–54. https://doi.org/10.1097/TME.0b013e318207e86a

Kaminsky, F. C., Maleyeff, J., Providence, S., Purinton, E., & Waryasz, M. (1997). Using SPC to analyze quality indicators in a healthcare organization. *Journal of Healthcare Risk Management: The Journal of the American Society for Healthcare Risk Management*, *17*(4), 14–22. https://doi.org/10.1002/jhrm.5600170404

Knoth, S. (2006). The art of evaluating monitoring schemes – How to measure the performance of control charts?. In H.-J. Lenz, & P.-T. Wilrich (Eds.), *Frontiers in statistical quality control* (Vol. 8, pp. 74–99). Springer.

Lee, M. L., Goldsman, D., & Kim, S.-H. (2015). Robust distribution-free multivariate CUSUM charts for spatiotemporal biosurveillance in the presence of spatial correlation. *IIE Transactions on Healthcare Systems Engineering*, *5*(2), 74–88. https://doi.org/10.1080/19488300.2015.1017674

Li, N., Kong, N., Li, Q., & Jiang, Z. (2017). Evaluation of reverse referral partnership in a tiered hospital system–a queuing-based approach. *International Journal of Production Research*, *55*(19), 5647–5663. https://doi.org/10.1080/00207543.2017.1327731

Lowry, C. A., Woodall, W. H., Champ, C. W., & Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, *34*(1), 46–53. https://doi.org/10.2307/1269551

Hertzum, M. (2021). How demanding is healthcare work? a meta-analytic review of TLX scores. *Context Sensitive Health Informatics: The Role of Informatics in Global Pandemics*, *286*, 55.

Mehandiratta, R. (2011). Applications of queuing theory in health care. *International Journal of Computing and Business Research*, *2*(2), 2229–6166.

Menascé, D. A., & Bardhan, S. (2019). TDQN: Trace-driven analytic queuing network modeling of computer systems. *Journal of Systems and Software*, *147*, 162–171. https://doi.org/10.1016/j.jss.2018.10.036

Mohammed, M. A. (2004). Using statistical process control to improve the quality of health care. *Quality and Safety in Health Care*, *13*(4), 243–245. https://doi.org/10.1136/qshc.2004.011650

Moran, J. L., & Solomon, P. J. (2013). Statistical process control of mortality series in the Australian and New Zealand intensive care society (ANZICS) adult patient database: Implications of the data generating process. *BMC Medical Research Methodology*, *13*(1), 66. https://doi.org/10.1186/1471-2288-13-66

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1-2), 100–115. https://doi.org/10.1093/biomet/41.1-2.100

Pagel, C., Ramnarayan, P., Ray, S., & Peters, M. J. (2018). Development and implementation of a real time statistical control method to identify the start and end of the winter surge in demand for paediatric intensive care. *European Journal of Operational Research*, *264*(3), 847–858. https://doi.org/10.1016/j.ejor.2016.08.023

Pignatiello, J. J., Jr., & Runger, G. C. (1990). Comparisons of multivariate CUSUM charts. *Journal of Quality Technology*, *22*(3), 173–186. https://doi.org/10.1080/00224065.1990.11979237

Qi, D., Li, Z., Zi, X., & Wang, Z. (2017). Weighted likelihood ratio chart for statistical monitoring of queueing systems. *Quality Technology & Quantitative Management*, *14*(1), 19–30. https://doi.org/10.1080/16843703.2016.1189184

Reiman, M. I. (1982). The heavy traffic diffusion approximation for sojourn times in Jackson networks. In R. L. Disney, & T. J. Ott (Eds.), *Applied probability–computer science: The interface* (pp. 409–421). Springer.

Reiser, M., & Kobayashi, H. (1975). Queuing networks with multiple closed chains: Theory and computational algorithms. *IBM Journal of Research and Development*, *19*(3), 283–294. https://doi.org/10.1147/rd.193.0283

Rogers, H., Gilligan, S., & Walters, M. (2008). Quality improvements in hospital flow may lead to a reduction in mortality. *Clinical Governance: An International Journal*, *13*(1), 26–34. https://doi.org/10.1108/14777270810850607

Roy, D., Gupta, A., & De Koster, R. B. (2016). A non-linear traffic flow-based queuing model to estimate container terminal throughput with AGVS. *International Journal of Production Research*, *54*(2), 472–493. https://doi.org/10.1080/00207543.2015.1056321

Shi, P., Helm, J. E., Heese, H. S., & Mitchell, A. M. (2021). An operational framework for the adoption and integration of new diagnostic tests. *Production and Operations Management*, *30*(2), 330–354. https://doi.org/10.1111/poms.13263

Shore, H. (2006). Control charts for the queue length in a g/g/s system. *IIE Transactions*, *38*(12), 1117–1130. https://doi.org/10.1080/07408170600737336

Sir, M. Y., Nestler, D., Hellmich, T., Das, D., Laughlin, M. J., Jr, Dohlman, M. C., & Pasupathy, K. (2017). Optimization of multidisciplinary staffing improves patient experiences at the mayo clinic. *Interfaces*, *47*(5), 425–441. https://doi.org/10.1287/inte.2017.0912

Sparks, R. S. (2000). CUSUM charts for signalling varying location shifts. *Journal of Quality Technology*, *32*(2), 157–171. https://doi.org/10.1080/00224065.2000.11979987

Vass, H., & Szabo, Z. K. (2015). Application of queuing model to patient flow in emergency department. case study. *Procedia Economics and Finance*, *32*, 479–487. https://doi.org/10.1016/S2212-5671(15)01421-5

Wen, J., Geng, N., & Xie, X. (2020). Real-time scheduling of semi-urgent patients under waiting time targets. *International Journal of Production Research*, *58*(4), 1127–1143. https://doi.org/10.1080/00207543.2019.1612965

Whitt, W., & Zhang, X. (2017). A data-driven model of an emergency department. *Operations Research for Health Care*, *12*, 1–15. https://doi.org/10.1016/j.orhc.2016.11.001

Woodall, W. H., Adams, B. M., & Benneyan, J. C. (2012). The use of control charts in healthcare. *Statistical Methods in Healthcare* (pp. 251–267).

Wu, S., Müller, H. G., & Zhang, Z. (2013). Functional data analysis for point processes with rare events. *Statistica Sinica*, *23*(1), 1–23. https://doi.org/10.5705/ss.2010.162

Xie, J., Cao, P., Huang, B., & Ong, M. E. H. (2016). Determining the conditions for reverse triage in emergency medical services using queuing theory. *International Journal of Production Research*, *54*(11), 3347–3364. https://doi.org/10.1080/00207543.2015.1109718

Zhang, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *64*(2), 281–294. https://doi.org/10.1111/1467-9868.00337

Zhao, X., & Gilbert, K. (2015). A statistical control chart for monitoring customer waiting time. *International Journal of Data Analysis Techniques and Strategies*, *7*(3), 301–321. https://doi.org/10.1504/IJDATS.2015.071366

Zou, C., & Qiu, P. (2009). Multivariate statistical process control using lasso. *Journal of the American Statistical Association*, *104*(488), 1586–1596. https://doi.org/10.1198/jasa.2009.tm08128

Zou, C., & Tsung, F. (2008). Directional MEWMA schemes for multistage process monitoring and diagnosis. *Journal of Quality Technology*, *40*(4), 407–427. https://doi.org/10.1080/00224065.2008.11917746

Zou, C., & Tsung, F. (2011). A multivariate sign EWMA control chart. *Technometrics*, *53*(1), 84–97. https://doi.org/10.1198/TECH.2010.09095

Zou, C., Ning, X., & Tsung, F. (2012). Lasso-based multivariate linear profile monitoring. *Annals of Operations Research*, *192*(1), 3–19. https://doi.org/10.1007/s10479-010-0797-8