ISSN: 2688-8645

Improve Response-Time Performance of Rapid Response Process in Acute Care: A System-Theoretic Method

Ran Li*

School of Shanghai University, Shanghai, China

Abstract

In order to improve the response time performance (RTP) of rapid response process in acute care delivery, an analytical model and system-theoretic method are introduced to study the response time performance. The analytical model is used to systematically describe the rapid response process, the decomposition method and the assumption that each step is approximate to gamma distribution are used to evaluate the response time performance (RTP). The results show that compared with previous studies, this method can evaluate the response time performance more quickly and accurately, and then improve the emergency efficiency of medical staffs and the safety of patients in acute care delivery. In addition, we study the system characteristics such as monotonicity and bottleneck analysis, and verify the feasibility of gamma approximate distribution. According to the data collected in University of Kentucky Chandler Hospital, further verify the accuracy of the model. Besides, compared with the previous model in the effectiveness and computational efficiency of RTP, which shows that the model in this paper is more effective in RTP evaluation and has more faster computing efficiency. After that, make improvement analysis to determine the biggest bottleneck of the rapid response process and improve to improve the decision-making efficiency. What-if analysis, consider the impact of other parameters on response time performance (RTP). Finally, the future research direction is discussed. potential future work to extend the study is discussed.

Keywords

Acute Care; Rapid Response; Response-Time Performance; Bottleneck; Gamma Approximate Distribution.

1. Introduction

In order to intervene and deal with patients before cardiopulmonary arrest or other life-threatening events, and avoid further deterioration of clinical conditions, Rapid Response Teams (RRT) have been set up in many foreign hospitals to reduce the incidence of cardiopulmonary arrest outside ICU and hospital mortality, and improve the satisfaction of patients, family members and employees [1, 2].

Numerous studies have investigated the effectiveness of RRT and found that the establishment of RRT can provide a systematic rapid response process for patients with deteriorating conditions, so as to significantly reduce the hospital mortality [3, 4]. Such studies mainly focus on the average response time and coefficient of variation (CV), which characterize the effectiveness of rapid response process [3]. Or a more direct measure is "response time performance", that is the probability of making a final decision within the expected period of time [4]. For example, Xie et al. (2012) initially studied the rapid response process in acute care, modeled as a complex network with split, merge and parallel structure and developed an analytical method to evaluate the decision time (from detection of patient deteriorating to a doctor's decision for treatment) and its variability [3]. In 2014, a closed formula to evaluate the

ISSN: 2688-8645

RTP by assuming exponential response time, and investigate the system-theoretic properties [4]. In addition, Chen et al. (2018) developed an analytical model to analyze the number, availability and waiting time distribution of care providers and to evaluate and improve the total waiting time of patients throughout the rapid response system [5]. Zeng et al. (2019) introduced a split, merge and parallel network model with complex structures to study the rapid response process and introduced an iterative method to evaluate the average decision-making time of multiple patients who decline at the same time [6].

From the perspective of methodology, the application trend of operational research methods in healthcare systems has become more and more obvious [7], Markov model and empirical model have been widely used [8]-[10]. Although discrete event computer simulation has been widely applied to provide detailed analysis, many of them are based on case studies and may require long model development and simulation time [11].

The current literature lacks a more rapid, accurate and feasible quantitative tool to improve the rapid response system from the perspective of average decision time, its variability and response time performance (RTP). In addition, a systematic review of the literature found that delayed activation of rapid response team seriously affected hospital mortality, cardiac arrest and intensive care transfer rate [12]. Literature [13] describes that the rapid response system can be redesigned to improve the interaction between RRT members and between RRT members and patients. Therefore, how to decrease response time or increase response time performance of rapid response system, adjust and improve the existing rapid response process model is very necessary for hospitals and patients.

In order to achieve this, firstly, this paper introduces a complex network including parallel, split and merge structures to simulate the rapid response process. The analysis model can provide fast and accurate estimation without relying on the detailed description of the process. More importantly, this quick approach is suitable for sensitivity analysis and design considering numerous scenarios. Then, the complex network is decomposed into multiple serial processes [10, 14], an analysis method is introduced to evaluate the performance of the serial subprocess and the whole process. Through the stochastic modeling of the rapid response process, the probability of completing the rapid response process in the expected or given time interval can be calculated, that is, the response time performance (RTP), as well as the average value and coefficient of variation (CV) of the response time. Then its structural characteristics are discussed and the bottleneck is analyzed. A case study of the University of Kentucky Chandler Hospital is introduced to illustrate the applicability of the method and to investigate suggestions for continuous improvement. Finally, what-if analysis is conducted to discuss the impact of other parameters on response time performance, and the potential future work to expand the study is discussed.

The main contribution of this paper is to introduce a complex network with parallel, split and merge structures to simulate the rapid response process. The network is decomposed into multiple serial processes, and each step is assumed to approximate gamma distribution to evaluate the performance of the serial subprocess and the whole process. Compared with previous literature studies, the analysis method based on the assumption of approximate gamma distribution in each step has faster and accurate calculation results, and provides a new quantitative analysis model for rapid response process.

The remainder of this paper is organized as follows. Section 2 describes the system and formulates the problem. The performance evaluation of response-time performance is introduced in Section 3. The improvement analysis is discussed in Section 4. Section 5 presents a validation study at the University of Kentucky Chandler Hospital and what-if analysis are carried out. Finally, conclusions and future works are summarized in Section 6.

ISSN: 2688-8645

2. System Description

A typical rapid response process in acute care system is showed in Figure 1. Patients are continuously monitored. When the signal drops, the nurse acting as primary care provider should respond quickly and seek help according to the patient 's condition, or call a higher level of care provider response, and finally make the final decision with the activated care provider response composed of a certain number and combination of doctors.

In order to evaluate the rapid response performance of such a process, the patient, care providers, and decision flow are defined by the following assumptions.

Patient:

(i) In each patient ward, there is one patient who may exhibit clinical deterioration.

Care providers:

(ii) The care provider response set is defined as X= {nurse, RRT, Intern, Resident, Fellow, Attending, RRT & Intern, RRT & Resident, RRT & Fellow, RRT & Attending}.

(iii) The response process of care providers to diagnose and treat the declining patient is modeled with a random process with mean τ_i and CV cv_i, where i is the step number as shown in Fig. 1. and l represents the workflows of nurse, intern, RRT, resident, RRT & intern, RRT & resident, RRT & fellow, fellow. Assume that all successive steps are independent. Note that this includes the time waiting for the provider.

Decision flow:

(iv) The care provider's choice of calling for a higher-level help is denoted by probability α_{ij}, and i and j are the indices of care providers, where $i \in X$, ($i \neq$Attending, or RRT & Attending), and i is the current care provider, j is the upper-level provider, as illustrated in Fig. 1.

(v) The response time for type i care provider is x_i, with a probability density function of $f_i(x)$ and cumulative distribution function of $F_i(x)$. We assume that the response time distributions of the same type care providers are identical and independent.

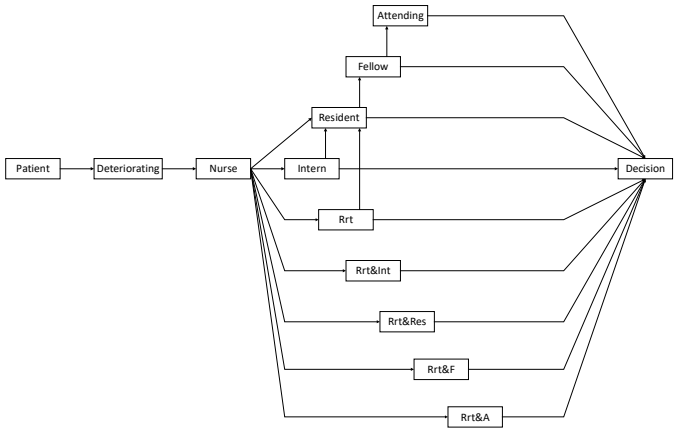


Figure 1. Rapid response process for declining patient in acute care.

3. Evaluation of Response-Time Performance

3.1. Process Decomposition

As shown in Figure. 1 and assumptions (i)-(v), the typical rapid response process in acute care can be described as a complex network with multiple splits, merges and parallel lanes and decision times, and each step has an unknown distribution type. All these aspects make the research of this kind of network more complicated, so that it is almost impossible to directly analyze the performance of response process.

ISSN: 2688-8645

However, considering that each patient can only take a specific procedure or route, we can decompose the deteriorating decision-making process into a set of procedures, and each procedure has a certain probability. Then we can decompose such a complex network into a series of processes. Therefore, in this framework, each serial process represents the route that a call to the helper might follow. As shown in Fig. 2, the complex network is decomposed into 21 path processes. Record as int1, int2, int3, int4, rrt1, rrt2, rrt3, rrt4, res1, res2, res3, rrt&int1, rrt&int2, rrt&int3, rrt&int4, rrt&res1, rrt&res2, rrt&res3, rrt&f1, rrt&f2, rrt&a. And subscripts n, int, rrt, res, f, a, rrt&int, rrt&res, rrt&f, rrt&a represent Nurse, Intern, RRT, Resident, Fellow, Attending, RRT & Intern, RRT & Resident, RRT & Fellow, and RRT & Attending, respectively.

Under assumptions (i)-(v), T_l is defined the time to finish each process, l =int1, int2, int3, int4, rrt1, rrt2, rrt3, rrt4, res1, res2, res3, rrt&int1, rrt&int2, rrt&int3, rrt&int4, rrt&res1, rrt&res2, rrt&res3, rrt&f1, rrt&f2, rrt&a. Then, the time to finish the overall response time, T, is characterized by

$$\begin{aligned} T = \max\{&T_{int1}, T_{int2}, T_{int3}, T_{int4}, T_{rrt1}, T_{rrt2}, T_{rrt3}, \\ &T_{rrt4}, T_{res1}, T_{res2}, T_{res3}, T_{\text{rrt\&int1}}, T_{\text{rrt\&int2,}}\ T_{\text{rrt\&int3,}} \\ &T_{\text{rrt\&int4}}, T_{\text{rrt\&res1}}, T_{\text{rrt\&res2}}, T_{\text{rrt\&res3,}} T_{\text{rrt\&f1}}, T_{\text{rrt\&f2}}, \\ &T_{\text{rrt\&a}}\} \end{aligned} \tag{1}$$

In an appropriately defined state space, the system with assumptions (i)-(v) forms a stationary random process. Let T_d be the response time. Introduce the notion of response-time performance, RTP, which is referred to as the probability that the response time is less than a desired or given time period T_{given}. Note that we assume that all decisions are appropriate. The quality of decision-making will be studied in the future. So RTP is a function of all system variables. Then

$$RTP(T_{given}) = \Pr(T \le T_{given}) = f\ (M, V, A, T_{given}) \tag{2}$$

where

$$\begin{aligned} M = [&\tau_{int}, \tau_{rrt}, \tau_{res}, \tau_{rrt\&int}, \tau_{rrt\&res}, \tau_{rrt\&f}, \tau_{rrt\&a}, \tau_f, \tau_a] \\ V = [&cv_{int}, cv_{rrt}, cv_{res}, cv_{rrt\&int}, cv_{rrt\&res}, \\ &cv_{rrt\&f}, cv_{rrt\&a}, cv_f, cv_a] \\ A = [&\alpha_{n,int}, \alpha_{n,rrt}, \alpha_{n,res}, \alpha_{n,rrt\&int}, \alpha_{n,rrt\&res}, \\ &\alpha_{n,rrt\&f}, \alpha_{n,rrt\&a}, \alpha_{int,res}, \alpha_{rrt\&int,res}, \\ &\alpha_{rrt,res}, \alpha_{res,f}, \alpha_{rrt\&res,f}, \alpha_{rrt\&f,a}, \alpha_{f,a}] \end{aligned}$$

Then the probability of each serial process can be expressed as

$$\begin{aligned} P = [&p_1^{int}, \dots\dots, p_4^{int}, p_1^{rrt}, \dots\dots, p_4^{rrt}, p_1^{res}, \dots\dots, p_3^{res} \\ &p_1^{rrt\&int}, \dots\dots p_4^{rrt\&int}, p_1^{rrt\&res}, \dots\dots p_3^{rrt\&res}, \\ &p_1^{rrt\&f}, p_2^{rrt\&f}, p^{rrt\&a}] \end{aligned} \tag{3}$$

Where the subscript corresponds to the care provider defined in each workflow, and

$$p_1^{int} = \alpha_{n,int}(1 - \alpha_{int,res})\dots$$
$$p_4^{int} = \alpha_{n,int}\alpha_{int,res}\alpha_{res,f}\alpha_{f,a}$$

ISSN: 2688-8645

$$
\begin{aligned}
p_1^{rrt} &= \alpha_{n,rrt}(1-\alpha_{rrt,res})\ldots \\
p_4^{rrt} &= \alpha_{n,rrt}\alpha_{rrt,res}\alpha_{res,f}\alpha_{f,a} \\
p_1^{res} &= \alpha_{n,res}(1-\alpha_{res,f})\ldots \\
p_3^{res} &= \alpha_{n,res}\alpha_{res,f}\alpha_{f,a} \\
p_1^{rrt\&int} &= \alpha_{n,rrt\&int}\left(1-\alpha_{rrt\&int,res}\right)\ldots \\
p_4^{rrt\&int} &= \alpha_{n,rrt\&int}\alpha_{rrt\&int,res}\alpha_{res,f}\alpha_{f,a} \\
p_1^{rrt\&res} &= \alpha_{n,rrt\&res}\left(1-\alpha_{rrt\&res,f}\right)\ldots \\
p_3^{rrt\&res} &= \alpha_{n,rrt\&res}\alpha_{rrt,\&res,f}\alpha_{f,a} \\
p_1^{rrt\&f} &= \alpha_{n,rrt\&f}\left(1-\alpha_{rrt\&f,a}\right) \\
p_2^{rrt\&f} &= \alpha_{n,rrt\&f}\ \alpha_{rrt\&f,a} \\
p_1^{rrt\&a} &= \alpha_{n,rrt\&a}
\end{aligned} \tag{4}
$$

In addition, n^ldenote the number of decomposed serial procedures in the workflow l described above, then

$$
\begin{gathered}
n^n = 7,\ n^{int} = 2,\ n^{rrt} = 2,\ n^{res} = 2, \\
n^{rrt\&int} = 2,\ n^{rrt\&res} = 2, \\
n^{rrt\&f} = 2,\ n^f = 2
\end{gathered}
$$

In addition

$$
\begin{gathered}
\sum_{k=1}^{n^l} p_k^l = 1 \\
l = n, int, rrt, res, rrt\&\ int, rrt\ \&\ res, rrt\ \&\ f, f
\end{gathered} \tag{5}
$$

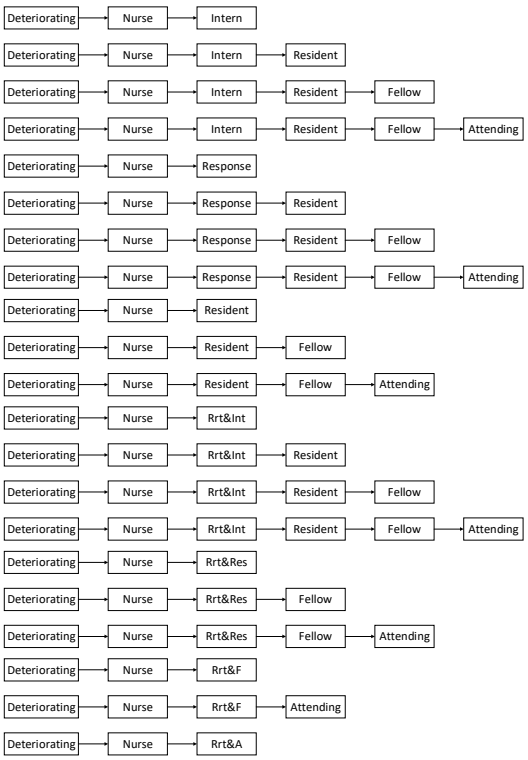


Figure 2. Decomposed serial procedures

ISSN: 2688-8645

Due to the independence assumption of parallel lanes, $RTP(T_{given})$ can be evaluated as

$$RTP(T_{given}) = \sum_l p_i^l RTP^l(T_{given}) \tag{6}$$

where $RTP^l(T_{given})$ indicates the probability to finish workflow l within time interval T_{given}. l =int1, int2, int3, int4, rrt1, rrt2, rrt3, rrt4, res1, res2, res3, rrt&int1, rrt&int2, rrt&int3, rrt&int4, rrt&res1, rrt&res2, rrt&res3, rrt&f1, rrt&f2, rrt&a.

Owing to the exact solutions of T_d and CV_d have been given in previous studies, we will not discuss them more here. Then, the problem to be studied can be described as follows. Under assumptions(i)-(v), develop a method to calculate RTP and investigate improvement strategies.

3.2. Performance Evalution

Even though the complex rapid response process can be decomposed into a series of subprocesses, each step is a random variable with arbitrary distribution, so it is difficult to analyze the decision-making time of each step. Each subprocess is the sum of multiple random variables with arbitrary distribution, so analyze the decision-making time of each process is difficult, and the overall process is the weighted aggregation of each subprocess, Therefore, it is more difficult to calculate the response time performance (RTP) of the whole rapid response process.

However, many existing literatures have proved that in many healthcare systems and manufacturing systems [9], when the CVs of the process is very small (less than 1), the overall performance may mainly depend on the first two moments of the process rather than complete distribution. Therefore, we assume the gamma distribution of the first two moments to approximate each service or step of the decision process. Note that this applies only when the variability of the process is small (i.e., CV < 1).

Specifically, denote the mean and CV of the j th service or decision time in serial procedure k of work flow l as $\tau_{k,(j)}^l$and $cv_{k,(j)}^l$, respectively. In addition, introduce n_k^las the number of steps in such a procedure (i.e., procedure k of workflow l). Then, we have

$$T_k^l = \sum_{j=1}^{n_k^l} \tau_{k,(j)}^l$$

$$CV_k^l = \frac{\sqrt{\sum_{j=1}^{n_k^l} \left(cv_{k,(j)}^l \cdot \tau_{k,(j)}^l\right)^2}}{T_k^l}$$

$$l = n, int, rrt, res, rrt\&\, int, rrt\ \&\ res, rrt\ \&\ f, f \tag{7}$$

By assuming that each step follows a gamma distribution, it is equivalent to evaluate the cumulative distribution function (cdf) of the sum of independently distributed gamma variables. If n^l is the number of serial procedures of workflow l, then the response-time performance can be calculated as follows.

Proposition 1: Under assumptions (i)–(v), if the response time at each step follows a gamma distribution, the response-time performance can be calculated as

$$RTP(T_{given}) = \frac{\sum_{k=1}^{n^l} p_k^l G_k^l(T_{given})}{\sum_{k=1}^{n^l} p_k^l} \tag{8}$$

ISSN: 2688-8645

$$l = n, int, rrt, res, rrt\& int, rrt \& res, rrt \& f, f$$

where $G_k^l(T_{given})$represents the cdf of serial procedure k in workflow l, i.e., the probability the time spent in this procedure is less than T_{given}, which can be evaluated as

$$G_k^l(T_{given}) = \prod_{j=1}^{n_k^l}\left(\frac{\beta_{k,min}^l}{\beta_{k,(j)}^l}\right)^{\eta_{k,(j)}^l} \cdot \sum_{m=0}^{\infty} \frac{\delta_{k,m}^l \gamma\left(\rho_k^l + m, \frac{T_{given}}{\beta_{k,min}^l}\right)}{\Gamma(\rho_k^l + m)} \tag{9}$$

and

$$\eta_{k,(j)}^l = \frac{1}{(cv_{k,(j)}^l)^2} \qquad \beta_{k,(j)}^l = (cv_{k,(j)}^l)^2 \cdot \tau_{k,(j)}^l$$

$$\beta_{k,min}^l = \min\left(\beta_{k,(j)}^l\right), j = 1, \ldots\ldots, n_k^l$$

$$\rho_k^l = \sum_{j=1}^{n_k^l} \eta_{k,(j)}^l, \quad \delta_{k,m}^l = 1$$

$$v_{k,m}^l = \frac{1}{m}\sum_{j=1}^{n_j^l} \eta_{k,(j)}^l \left(1 - \frac{\beta_{k,min}^l}{\beta_{k,(j)}^l}\right), \quad m = 1,2 \ldots$$

$$\delta_{k,m}^l = \frac{1}{m+1}\sum_{j=1}^{m+1} j\, v_{k,j}\delta_{k,m+1-j}, \quad m = 1,2 \ldots$$

$$\gamma(a,x) = \int_0^x y^{a-1} e^{-y} dy$$

$$\Gamma(a) = \lim_{x\to\infty} \gamma(a,x) \tag{10}$$

Proof:

It has been shown in [9] that if $\{X_i$, i = 1,..., n} are independently distributed gamma random variables with mean τ_i and standard deviation σ_i, then

$$G\left(T_{given}\right) = \prod_{i=1}^{n}\left(\frac{\beta_{min}}{\beta_i}\right)^{\eta_i} \cdot \sum_{k=0}^{\infty} \frac{\delta_k \gamma\left(\rho + k, \frac{T_{given}}{\beta_{min}}\right)}{\Gamma(\rho + k)}$$

where β_{min}, β_i, η_i, δ_k, ρ, $\gamma(.)$ and $\Gamma(.)$ are defined in (13). Using this result, the number of steps in serial process i can be defined by n_i^l, Then, RTP_l of procedure i of workflow l for time interval T_{given} can be evaluated by $G_i^l(T_{given})$. Then, RTP_l can be calculated using the weighted sum of the CDF of all the pathways in workflow l, i.e.,

$$RTP^l = \frac{\sum_I p_i^l G_i^l(T_{given})}{\sum_i p_i^l}$$

we obtain the overall response-time performance of the whole process.

4. Improvement Analysis

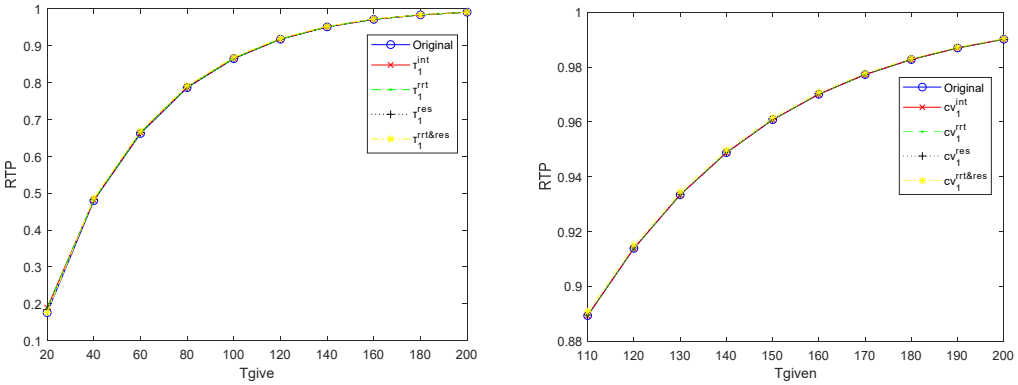
Use the performance evaluation method described above to analyze the performance of the rapid response system, so as to improve the system performance of the rapid response process.

ISSN: 2688-8645

It is known that the most effective way to improve system performance is to identify and alleviate bottlenecks. Bottlenecks are generally considered to be the processes that have the greatest impact on system performance. In the medical service delivery system, the bottleneck process is called response, and its improvement will lead to the greatest improvement of performance measurement. For example, the paper of Chen (2018) shows that RPH workflow (i. e., waiting for the doctor's discharge order) may be the bottleneck in the discharge process [14]. Therefore, firstly, the monotonicity of the system is studied to determine the potential improvement direction. Secondly, bottlenecks are studied to determine the response time and variability that most hinder service decision-making.

4.1. Monotonicity

Proposition 2: Under assumptions (i)–(v), the response-time performance is monotonically decreasing with respect to the mean response time of each step.



(a) (b)

Figure 3. RTP monotonicity

Proof:

For the mean of decision time τ_i^l, it is obvious that

$$\frac{\partial G_j^l(T_{given})}{\partial \tau_i^l} < 0$$

when route j of workflow l includes step τ_i^l; otherwise, $G_j^l(T_{given})$ does not change with τ_i^l.

In addition, for workflow $k \neq l$, k=n, int, rrt, res, rrt& int, rrt & res, rrt & f, f, we have

$$\frac{\partial G_j^k(T_{given})}{\partial \tau_i^l} = 0$$

That is to say, the response-time performance of a workflow decreases with respect to the mean of waiting time of the step from the same workflow and does not change with the one from a different workflow

$$\frac{\partial RTP^l}{\partial \tau_i^l} < 0$$

ISSN: 2688-8645

$$\frac{\partial RTP^k}{\partial \tau_i^l} = 0, for\ k \neq l$$

Thus, from (6), we obtain

$$\frac{\partial RTP}{\partial \tau_i^l} = \frac{\partial \int_0^\infty \left(1 - \sum_{k=n,int,rrt,res,rrt\&\ int,rrt\ \&\ res,rrt\ \&\ f,f} p_i^k RTP^k\right)}{\partial \tau_i^l} < 0$$

In order to verify this characteristic, a large number of simulation and numerical experiments have been carried out. As shown in Fig.3(a), when the mean response time is reduced by 10%, the RTP is always increase. However, different values of the mean response time of each step have different increases. In Fig.3(a), The corresponding RTP curves due to the decrease of is $\tau_1^{int}, \tau_1^{rrt}, \tau_1^{res}, \tau_1^{rrt\&res}$ given. As you can see, a small decrease in the average response time of step RRT & resident (i.e., seeking rapid response team and resident for help) lead to a more increase in RTP compared with other steps, which only result in a very small increase in RTP.

Note that when T_{given} is very small, the response-time performance is not monotonically decreasing with respect to the CV of each step. That is because with the decrease of CV, for smaller T_{given}, the RTP decreases, while for larger T_{given}, the RTP increases. Therefore, when we study the monotonicity of RTP with respect to CV, we only consider the case where T_{given} is large. In fact, a more practical way is to provide the probability of completion within the average decision time and estimate the larger T_{given} value. For those ranges, CV shows monotonicity behavior.

Inference 1: Under assumptions (i)–(v), the response-time performance is monotonically decreasing with respect to the CV of each step.

When T_{given} is large, as shown in Fig.2(b), The corresponding RTP curves due to the decrease of is $cv_1^{int}, cv_1^{rrt}, cv_1^{res}, cv_1^{rrt\&res}$ given. when the CV of response time is reduced by 20%, the RTP is increased. Similarly, a slight decrease in the coefficient of variation of step RRT & resident will lead to a more increase in RTP.

The monotonicity is consistent with our intuitions well. This indicates that RTP can be increased by reducing mean time or variability of response time.

4.2. Bottleneck Analysis

The monotonicity study shows that decreasing the τ and cv of each step can lead to the increase of RTP, but the increase of RTP may be significantly different when different steps are decreased. Therefore, which response time and its variability should be reduced to maximize enhance RTP? Or which step is the strongest way to impedes response time performance? This step is called the bottleneck step. Formally, for continuous function of RTP, the bottleneck step can be defined as follows.

Definition 1: Step r_i^l is the variability RTP bottleneck ($BN - RTP_\tau$) if

$$\left|\frac{\partial RTP}{\partial \tau_i^l}\right| > \left|\frac{\partial RTP}{\partial \tau_j^k}\right|, \quad \forall\{j,k\} \neq \{i,l\}$$

Definition 2: Step r_i^l is the variability RTP bottleneck ($BN - RTP_{cv}$) if

ISSN: 2688-8645

$$\left|\frac{\partial RTP}{\partial cv_i^l}\right| > \left|\frac{\partial RTP}{\partial cv_j^k}\right|, \quad \forall\{j,k\} \neq \{i,l\}$$

However, the partial derivatives are difficult to evaluate analytically. To identify the bottleneck steps, let $RTP\ (\tau_i^l - \delta_\tau \tau_i^l)$ and $RTP\ (cv_i^l - \delta_{cv} cv_i^l)$ be the RTP when the mean and cv of response time at step r_i^l are reduced by proportion δ_τ and δ_{cv}, respectively. In addition, $\delta_\tau \ll 1$ and $\delta_{cv} \ll 1$. Then, the following approach is used to identify the bottlenecks.

(1) Mean time τ_i^l is the $\text{BN} - \text{RTP}_\tau$ if

$$RTP\left(\tau_i^l - \delta_\tau \tau_i^l\right) > RTP\left(\tau_j^l - \delta_\tau \tau_j^l\right), \qquad \forall\{j,k\} \neq \{i,l\}$$

(2) $CV\ cv_i^l$ is the $\text{BN} - \text{RTP}_{\text{cv}}$ if

$$RTP\left(cv_i^l - \delta_{cv} cv_i^l\right) > RTP\left(cv_j^l - \delta_{cv} cv_j^l\right), \forall\{j,k\} \neq \{i,l\}$$

Note that, calculation of such indicators can be carried out using the data collected on the hospital floor without calculations of performances and their partial derivatives.

5. Improvement Analysis

Based on the previous description, we modeled and analyzed the rapid response process of acute care delivery at the University of Kentucky Chandler Medical Center.

5.1. Model Development

The rapid response process in UKCH acute care delivery is similar to the process introduced in Section 2 and 3. We found the relevant data from the previous literature [3]. It was included the necessary information such as call-for-help time, provider response information, diagnosis time, treatment and decisions. These data enable us to calculate the routing probabilities and response time. Based on this information, an analysis model is established. However, because some processes occur only a few times, we aggregate these processes (e.g., services by resident, fellow and attending doctors) into a single service to get a simplified process model, as shown in the Fig.4., the complex network is decomposed into 4 path processes. Then the probability of each serial process can be expressed as

$$P = [p_1^{rrt\&int}, p_2^{rrt\&int}, p_1^{rrt\&res}, p_1^{rrt\&a}]$$

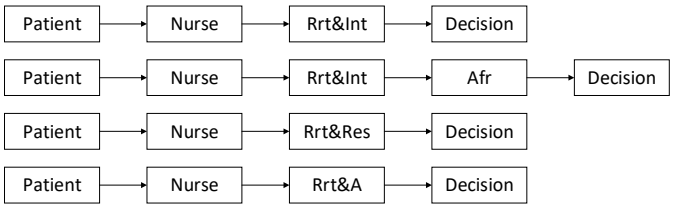


Figure 4. Decomposed serial procedures.

Using the decomposed serial procedures in Figure 4. and the information and data collected. From (3) and (4), we calculate the probability of each serial process p_i^l. $RTP\left(T_{given}\right)$ can be calculated from (8) and (9). Next, we can make further analysis.

ISSN: 2688-8645

5.2. Model Validation

Proposition 1 introduces an approximate method to calculate the response time performance. To verify the model, two issues need to be considered. First, does response time performance mainly depend on the average value and CV of each step rather than the complete distribution? If the answer to the first question is yes, the second question is related to the accuracy of the approximation, that is, whether proposition 1 can provide an accurate estimate of response time performance?

First, we investigate the feasibility of gamma approximation. Two simulation models are compared. One assumes that every step is represented by a gamma distribution, while the other assumes that each step follows a randomly selected distribution (either gamma, lognormal, Weibull, or a mixture of them). However, the same first two moments are assumed in both scenarios at each step.

Using the data collected in UKCH, we assume the average value and CV of each step in the initial model, then randomly generate the average service time of each step based on 50% to 150% of the average value, and randomly select the CVs between 0.5 and 1.0. The simulation settings are as follows: collect the data of 10000 patients and repeat the simulation for 20 times to ensure that the confidence interval is small enough to be less than 1% of the RTP value. Then, the response time characteristics of 20 data sets are simulated and compared. The results are shown in Table 1, where the average, minimum, and maximum differences between the two models for given T_{given} values are presented.

Table 1. RTP Differences in Distribution types.

T_g	Ave.diff	Min.diff	Max.diff
10	0.0023	0.0003	0.0048
20	0.0049	0.0006	0.0108
30	0.0043	0.0002	0.0118
40	0.0040	0.0002	0.0119
50	0.0039	0.0003	0.0105
60	0.0039	0.0004	0.0098
70	0.0036	0.0003	0.0093
80	0.0031	0.0002	0.0083
90	0.0028	0.0002	0.0079
100	0.0024	0.0002	0.0067
110	0.0022	0.0003	0.0063
120	0.0020	0.0001	0.0052
130	0.0016	0.0001	0.0045
140	0.0013	0.0001	0.0034
150	0.0011	0.0000	0.0030
160	0.0009	0.0001	0.0025
170	0.0008	0.0001	0.0020
180	0.0006	0.0000	0.0017
190	0.0005	0.0000	0.0015
200	0.0005	0.0000	0.0013

It can be seen from the table 1 that under different distribution types, the average, minimum and maximum differences of RTP are very small, which indicates that the response time performance is actually independent of the distribution type, but mainly depends on the

ISSN: 2688-8645

average value and CV of response time. This proves that gamma distribution can be used to characterize the response time of each step. Figure 5 shows four examples by comparing a model with gamma distribution at each step with a randomly selected lognormal, Weibull, or gamma distribution at each step (referred to as "mixed" in the figure). The blue line represents the RTP obtained through proposition 1, and the red line represents the RTP obtained through the mixed simulation of lognormal distribution, Weibull distribution and gamma distribution. Similarly, the difference of RTP is very small, which further verifies the feasibility of the model.

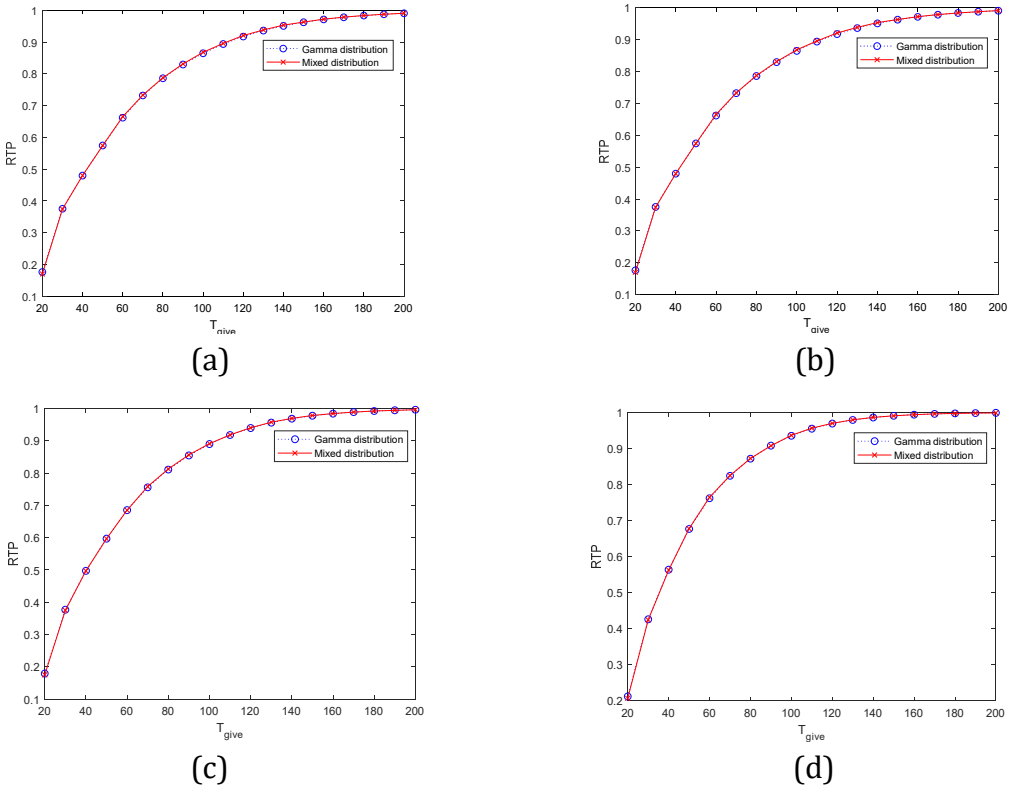


Figure 5. Comparison examples in distribution type.

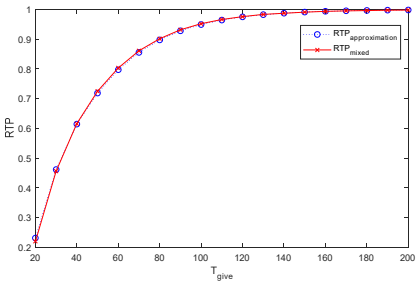


Figure 6. Comparison example using UKCH data

Data collected from UKCH are also used for comparison to simulate and compare the UKCH rapid response process model, and further verify the accuracy of response time performance. First, we calculate that the mean response time is 39.091 mins, the variance is evaluated as 834.782 mins. This leads to CV_d = 0.7391. These results match with the collected data well. Therefore, the model can be used for the subsequent analysis. Secondly, the response time performance calculated from proposition 1 is compared with the simulation assuming a

ISSN: 2688-8645

randomly selected distribution at each step. The results are shown in Figure 5. In all cases, the estimates have acceptable accuracy.

5.3. Comparative Analysis

In order to further verify the improvement effect of the model on the rapid response system, this paper compares the improved model with the model in [4] on the effectiveness and efficiency of RTP evaluation

The conclusion of the two models is that the difference between RTP is small in all comparisons for different distribution types. The average value of the difference in this model is 0.23%, the minimum value of the difference is 0.019%, and the maximum value of the difference is 0.62%. The model in literature [4], the average value is 0.59%, and the maximum value is 1.37%. The comparison shows that the proposed model has more accuracy in RTP evaluation compared with the previous model in different distribution types. At the same time, the paper [4] proposes a linear interpolation method to approximate RTP in the non-exponential case, and studies the accuracy of the linear interpolation estimation through simulation experiments. Under the lognormal and gamma assumptions of response time, the difference between RTP calculated by approximate formula and simulation is also small. The average values of difference with lognormal and gamma distribution are 0.7% and 0.8%, respectively. Likewise, the comparison again shows that the model in this paper has better accuracy for RTP estimation.

5.4. Improvement Analysis

Using the method described earlier, based on the hospital data, we evaluated the hospital's RTP improvement by assuming δ_τ= 0.1 and δ_{cv}= 0.2. The results are shown in Table 2. As one can see, the decrease of $\tau_{rrt\&res}$ and $cv_{rrt\&res}$ leads to the maximum improvement of RTP. Therefore, the RRT & resident step is the bottleneck. Reducing the probability of seek RRT & resident for help will greatly increase RTP and improve the rapid response process.

Table 2. RTP Improvements with Respect Mean or CV Reduction

(a)

T_{give}	$\tau_{rrt\&int}$	$\tau_{rrt\&res}$	$\tau_{rrt\&a}$	τ_{afr}
100	0.000512	0.010680	0.003925	0.001400
150	0.000044	0.001955	0.001695	0.000175
200	0.000003	0.000279	0.000636	0.000016

(b)

T_{give}	$cv_{rrt\&int}$	$cv_{rrt\&res}$	$cv_{rrt\&a}$	cv_{afr}
100	0.000093	0.014421	0.004446	0.002002
150	0.000009	0.003058	0.002773	0.000287
200	0.000007	0.000415	0.001113	0.000025

5.5. Comparative Analysis

In addition to bottleneck analysis to identify bottlenecks and improve bottleneck steps, we can also consider how other parameters affect the response time performance (RTP). In particular, we conducted what-if analysis on the probability of nurses calling the rapid response team and doctors.

(1) Nurse call RRT & intern ($\alpha_{n,rrt\&int}$): The probability of nurses calling RRT & intern is still relatively high in reality, because once the patient's condition deteriorates, nurses will first consider calling the first level nursing staff. As shown in table 3(a), when the probability $\alpha_{n,rrt\&int}$ decreases from 30% to 15%, the value of RTP has obvious changes.

ISSN: 2688-8645

(2) Nurse call RRT & resident ($\alpha_{n,rrt\&res}$): Similarly, when the probability $\alpha_{n,rrt\&res}$ decreases from 55% to 40%, the value of RTP decreases significantly, as shown in table 3(b).

(3) Nurse call RRT & attending ($\alpha_{n,rrt\&a}$): As shown in table 3(c), when the probability $\alpha_{n,rrt\&a}$ decreases from 15% to 9%, the value of RTP decreases. However, compared with the previous two probabilities (i.e., $\alpha_{n,rrt\&int}$, $\alpha_{n,rrt\&res}$), the reduction of RTP is not so significant. This is because nurses have a lower probability of seeking RTP & attaching when the patient's condition deteriorates. Therefore, changing the probability has no greater impact on RTP than the other two.

Table 3(a): What-if Analysis with Respect to $\alpha_{n,rrt\&int}$

T_{give}	$\alpha_{n,rrt\&int}$	0.03	0.25	0.20	0.15
100	RTP	0.9490	0.8997	0.8504	0.8011
	Percentage	-	-5.2%	-10.4%	-15.6%
125	RTP	0.9785	0.9287	0.8789	0.8292
	Percentage	-	-5.1%	-10.2%	-15.3%
150	RTP	0.9907	0.9408	0.8908	0.8409
	Percentage	-	-5.0%	-10.1%	-15.1%

(b): What-if Analysis with Respect to $\alpha_{n,rrt\&res}$

T_{give}	$\alpha_{n,rrt\&res}$	0.55	0.50	0.45	0.40
100	RTP	0.9490	0.9017	0.8544	0.8071
	Percentage	-	-5.0%	-10.0%	-15.0%
125	RTP	0.9785	0.9295	0.8805	0.8315
	Percentage	-	-5.0%	-10.0%	-15.0%
150	RTP	0.9907	0.9410	0.8914	0.8417
	Percentage	-	-5.0%	-10.0%	-15.0%

(c): What-if Analysis with Respect to $\alpha_{n,rrt\&a}$

T_{give}	$\alpha_{n,rrt\&a}$	0.15	0.13	0.11	0.09
100	RTP	0.9490	0.9312	0.9135	0.8957
	Percentage	-	-1.9%	-3.7%	-5.6%
125	RTP	0.9785	0.9598	0.9410	0.9223
	Percentage	-	-1.9%	-3.8%	-5.8%
150	RTP	0.9907	0.9714	0.9521	0.9327
	Percentage	-	-2.0%	-3.9%	-5.9%

6. Conclusion and Future Work

This paper introduces an analytical model to study the rapid response process and response time performance (RTP) in acute care delivery. The analytical model is used to systematically describe the rapid reaction process. The decomposition method and the assumption that each step is approximate to gamma distribution are used to evaluate the response time performance. The results show that compared with previous studies, this method can evaluate the response time performance more quickly and accurately, and improve patient safety in Acute Care. In addition, we also study the monotonicity and bottleneck analysis, verify the feasibility of the model assuming gamma approximate distribution. According to the data collected in University of Kentucky Chandler Hospital, we further verified the accuracy of the model, compared with

the previous model in the effectiveness and computational efficiency of RTP, and conduct improvement analysis. Finally, what-if analyses are carried out to consider how other parameters affect the response time performance.

This model provides a quantitative tool for hospital management to study and improve the performance of rapid response system in acute care delivery, so as to improve the quality of care and patient safety. In addition, the method can also be used in the engineering field of production, product development and other parallel processes.

There are also limitations in this study, which is based on the reaction process of a single patient and assumes that the provider is always available. In practice, there are multiple patients on the floor, and more than one patient may deteriorate at the same time, while the number of providers on the hospital floor is limited. Therefore, in order to expand this research, we can make a more in-depth analysis of the rapid response process. In future work, we can use this model to study multiple patients on the hospital floor who are deteriorating at the same time, and a limited number of providers need to treat multiple patients who are deteriorating at the same time. In this case, because providers are limited and may not reach multiple deteriorating patients at the same time, care delivery may be delayed, which increases the decide time and treat time. This will greatly affect the safety of patients and the quality of care. Therefore, it is meaningful to study it.

References

[1] K. K. Hall, A. Lim and B. Gale, The Use of Rapid Response Teams to Reduce Failure to Rescue Events: A Systematic Review, Journal of Patient Safety, vol. 16(2020).

[2] C. Li, R. Teuma Custo and J. Trapani, The impact of rapid response systems on mortality and cardiac arrests – A literature review, Intensive and Critical Care Nursing, 2020.

[3] X. Xie, J. Li, C. H. Swartz and P. DePriest, Modeling and Analysis of Rapid Response Process to Improve Patient Safety in Acute Care, in IEEE Transactions on Automation Science and Engineering, vol. 9(2012) No. 2, p. 215-225. T.D`. Zhang, A.J. Shih and E. Levin: Annals of the CIRP, Vol. 43 (1994) No.3, p.305-307.

[4] X. Xie, J. Li, C. H. Swartz and P. DePriest, Improving Response-Time Performance in Acute Care Delivery: A Systems Approach, in IEEE Transactions on Automation Science and Engineering, vol. 11(2014), No. 4, p. 1240-1249.

[5] N. Chen, M. Wang, X. Xie, L. Zheng and C. H. Swartz, Modeling and Analysis of the Waiting Time of Rapid Response Process in Acute Care, in IEEE Robotics and Automation Letters, vol. 3(2018), No. 1, p. 336-343.

[6] Z. Zeng, Z. Fan, X. Xie, et al. A two-level iteration approach for modeling and analysis of rapid response process with multiple deteriorating patients, Flex Serv Manuf J, vol. 32(2020), p. 35–71.

[7] J. L. Wiler, R. T. Griffey and T. Olsen, Review of Modeling Approaches for Emergency Department Patient Flow and Crowding Research, Academic Emergency Medicine, vol. 18(12)(2011), p. 1371–1379.

[8] X. Shao, J. Li and D. A. Wiegmann, A Markov chain approach to study flow disruptions on surgery in emergency care, 2013 IEEE International Conference on Automation Science and Engineering (CASE), pp. 990-995, 2013.

[9] X. Xie, J. Li, C. H. Swartz, Y. Dong and P. DePriest, Modeling and Analysis of Ward Patient Rescue Process on the Hospital Floor, in IEEE Transactions on Automation Science and Engineering, vol. 13(2016), No. 2, p. 514-528.

[10] H. K. Lee et al., A System-Theoretic Method for Modeling, Analysis, and Improvement of Lung Cancer Diagnosis-to-Surgery Process, in IEEE Transactions on Automation Science and Engineering, vol. 15(2018), No. 2, p. 531-544.

ISSN: 2688-8645

[11] M. P. Fanti, A. M. Mangini, M. Dotoli and W. Ukovich, A Three-Level Strategy for the Design and Performance Evaluation of Hospital Departments, in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 43(2013), No. 4, p. 742-756.

[12] M.K. Xu, K.G. Dobson, L. Thabane, et al., Evaluating the effect of delayed activation of rapid response teams on patient outcomes: a systematic review protocol, Syst Rev, vol. 7(2018), p. 42.

[13] R. Chalwin, L. Giles, A. Salter, et al., Re-designing a rapid response system: effect on staff experiences and perceptions of rapid response team calls, BMC Health Serv Res, vol. 20(2020), p. 480.

[14] N. Chen et al., Improving Discharge Process at the University of Wisconsin Hospital: A System-Theoretic Method, in IEEE Transactions on Automation Science and Engineering, vol. 16(2019), No. 4, p. 1732-1749.