Flexible Services and Manufacturing Journal (2020) 32:35–71
https://doi.org/10.1007/s10696-019-09347-6

A two-level iteration approach for modeling and analysis of rapid response process with multiple deteriorating patients

Zexian Zeng[1] · Zhenghao Fan[2] · Xiaolei Xie[2] · Colleen H. Swartz[3] · Paul DePriest[4] · Jingshan Li[5]

Published online: 15 April 2019

Abstract
In acute care, a patient's clinical deterioration is often a precursor to serious and often fatal outcomes. To reduce the severity and frequency of negative outcomes, care providers need to response rapidly by providing quick evaluation, triage, and treatment to patients with declining conditions. However, a provider's availability to respond can be constrained when multiple patients are deteriorating at the same time. To study the multiple patients rapid response process, we introduce a network model with complex structures, such as split, merge, and parallel. Iterative methods are presented to evaluate the mean decision time (i.e., the average time from the detection of a patient's declining to a physician's treatment decision being made). It is shown that such methods lead to convergent results and high accuracy in performance evaluation. Such a model provides a quantitative tool for healthcare professionals to design and improve rapid response systems.

Keywords Rapid response · Decision time · Mean waiting time · Multiple patients · Patient deterioration · Iterations

1 Introduction

After the publication of the US Institute of Medicine's report "To Err is Human" (Kohn et al. 2000), there has been a national initiative in the US to improve patient safety (Watcher 2004; Leape and Berwick 2005; Berwick et al. 2006; Brindley 2010). In addition to regular care services, rapid response teams (RRTs), also referred to as medical emergency teams (METs), or critical care outreach (CCO), have been implemented in many hospitals to provide quick evaluation, triage, and treatment to patients with clinical signs of deterioration on the hospital floor

✉ Xiaolei Xie
xxie@tsinghua.edu.cn

Extended author information available on the last page of the article

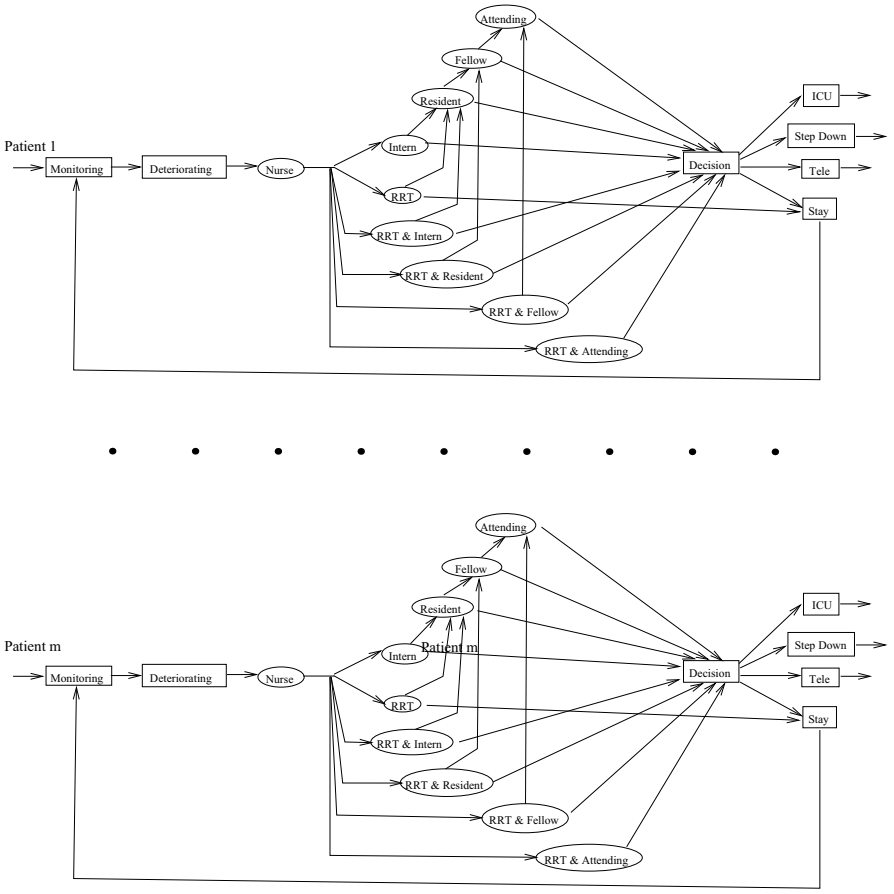

Fig. 1 Rapid response process with multiple patients

(DeVita et al. 2006, 2011). However, recent studies have indicated that there exist inconsistent results regarding the effectiveness of implementing RRTs. Therefore, an in-depth study of the efficacy of RRTs is necessary.

To study this issue, in this paper, we consider a rapid response system with multiple patients in a teaching hospital environment. In such systems, multiple patients could deteriorate simultaneously. Each requires timely diagnosis and treatment from a limited number of providers, who need to make prompt decisions through a hierarchical referral procedure. When a patient's deterioration is identified by a monitoring nurse, he/she can inform one of the providers or the RRT, as shown in Fig. 1. In other words, the nurse can notify either the intern, or the RRT, or both RRT and one provider (intern, resident, fellow, or attending). The RRT can either keep the patient "stay" or can notify the resident and rely on his/her judgment. If a provider is asked, either a diagnosis and treatment decision can be made, or assistance from an upper level provider can be sought. For instance, an intern (or RRT, or RRT & intern) may seek help

from the resident. Similarly, the resident (or RRT & resident) may either make a decision or ask help from the fellow. The fellow (or RRT & fellow) again can make a decision or request the attending's help. The attending (or RRT & attending) must make a final decision, in one of the four options: elevating the patient to be admitted to "ICU", monitoring for progressive care (referred to as "step down"), moving to a telemetry bed ("tele"), or keeping the patient for observation (i.e., "stay").

As one can see, the rapid response process requires integrated and collaborative operations of multiple care providers from different divisions or departments. Early identification, better recognition, as well as prompt response and treatment, play key roles. Therefore, a systematic study of the whole rapid response system (RRS), rather than an individual response or a provider, is necessary and important (DeVita et al. 2006). Among various performance measures, the mean decision time, i.e., the average time from decline to an appropriate medical decision is important since clinical studies have shown that patient safety is strongly correlated to the decision time (Hillman et al. 2001). Therefore, evaluation of the mean decision time is of critical importance for RRS, which is the focus in this study.

Although extensive clinical studies have been devoted to the rapid response process, the investigation from a systems engineering point of view is still limited. To bridge this gap, both discrete event simulation and analytical methods are viable. They are complement to each other and have different advantages and limitations, which can provide results and insights from different perspectives. In this paper, we focus on developing an analytical method. The contribution of this paper is to introduce an iteration method to evaluate the average decision time in a multi-patient rapid response process, where the extra waiting times, due to unavailability of the providers are taken into account and are updated through recursive procedures. To our best knowledge, no such study is available in the literature. Using such a method, the estimation of the mean decision time can be obtained, which is critical to patient safety. Such a method provides a quantitative tool for operation management of the rapid response process with multiple patients. In addition, such a model could also enable us to identify the response time that is most critical to the overall decision time through sensitivity analysis. Then efforts can be organized to decrease this response time, such as increasing number of resident doctors, thus reducing the number of patients each covers, so that the overall decision time can be improved.

The remainder of the paper is structured as follows: Sect. 2 briefly reviews the related literature. Section 3 introduces the rapid response process in multiple patients environment and formulates the problem. By considering limited resource, a two-level iteration method to estimate the mean decision time is presented in Sect. 4. Finally, Sect. 5 presents conclusions and summarizes future work. All proofs are provided in the "Appendix".

2 Related literature

The historical report "To Err is Human," published by the US Institute of Medicine, has estimated that the number of potentially preventable hospital deaths in the US is up to 100,000 per year (Kohn et al. 2000). Since its publication, there

has been a worldwide concern regarding patient safety; numerous efforts have been made to improve processes (see, for example, reviews by Watcher 2004; Leape and Berwick 2005; Berwick et al. 2006). Moreover, the data in the report has triggered a trend in the US to implement RRTs (or METs, CCO) in hospitals (DeVita et al. 2011). Numerous studies have been carried out to investigate the effectiveness of RRTs and RRSs. In some studies, it is found that the implementation of RRT has provided a systematic response procedure to patients with deterioration episodes, leading to substantial reduction in mortality in some hospitals (e.g., Priestley et al. 2004; DeVita et al. 2006; Dacey et al. 2007). However, in other hospitals, such positive improvements were not observed and there is no consistent clinical conclusion regarding the effectiveness of RRTs (see Massey et al. 2010; Hillman et al. 2005; Winters et al. 2007; Ranji et al. 2007; Chan et al. 2010). As most of the available studies are observational or trial-based, a systematic study using mathematical models could provide a new perspective and generate guidance to adjust and optimize the existing rapid response system (Downey et al. 2008).

Clinical studies have suggested that the majority of the patients show signs of deterioration before ICU admission (Hillman et al. 2001; Downey et al. 2008; Xie et al. 2012, 2014), and the time of quick response and treatment to patient decline is critical to reduce safety risk and ICU burden (McGloin et al. 1999; McArthur-Rouse 2001). Thus, a quantitative study of the response and decision time in RRS becomes important. However, in the current literature, only the single patient scenario has been studied (see papers by Xie et al. 2012, 2014). In Xie et al. (2012), the mean decision time and its variability are analyzed via a response network model with split and merge. The most impeding response, i.e., the bottleneck response with respect to improvement in individual response time can be identified. To further investigate the system behavior, the response time performance (RTP), i.e., the probability that an appropriate decision can be made within a desired time duration, is proposed by Xie et al. (2014). A closed formula is presented under exponential assumption of response time, and an empirical modification law for the general case is introduced. The bottleneck response from the RTP perspective is also analyzed.

However, these studies are based on response process for a single patient and assumes providers are always available. In practice, there are multiple patients on the floor, and more than one patients may deteriorate simultaneously while the number of providers on the hospital floor is limited. Thus, there is a chance that the limited providers may need to treat multiple patients who are deteriorating simultaneously. In this case, care delivery may be delayed due to the unavailability of providers. This will significantly impact patient safety and quality of care. Unfortunately, such a scenario has not been investigated yet.

From the methodology point of view, among substantial efforts contributing to healthcare systems research, simulation has been used as a prevailing tool (see reviews by Jacobson et al. 2006; Gunal and Pidd 2010; Wiler et al. 2011; Zhong et al. 2015). Although simulation is a viable approach, this paper provides an alternative and complement method based on network analysis. Analytical model can provide a fast and accurate estimation and is not dependent on the detailed description of the process. More importantly, such a quick approach enables us to study numerous scenarios related to sensitivity analysis and design considerations to

find better solutions. For example, queueing models and Markov chain approach are often used [see monographs by Brandeau et al. 2004; Hall 2006 and papers by Schaefer et al. 2005; Green 2006; Fomundam and Herrmann 2007; Lakshmi and Iyer 2013; Garg et al. 2010; Mayhew and Smith 2008; Wang et al. 2012, 2013, 2014]. However, in some cases, some specific assumptions (e.g., poisson arrivals, exponential service time) may limit their applications. For example, a Markov chain model of ward patient rescue process is presented by Xie et al. (2016). Although RRT is involved in the model, the main focus is on estimating the steady state probabilities of patient status based on exponential assumption of intervention time.

To summarize, developing a novel analytical model to study rapid response system with multiple declining patients and limited provider availability is necessary, which is pursued using an analytical approach in this study.

3 System assumptions and problem formulation

Consider a rapid response system under a multi-layer referral mechanism, shown in Fig. 1. The variables used to characterize the rapid response process throughout the paper are summarized in Table 1.

Remark 1 In the US medical system (Whitlock 2017), particularly in teaching hospitals, "interns" refer to the doctors who have completed their first year of post-medical school training; the residency follows the intern year. Fellows are the physicians who have completed their residency and have elected to complete further training in a specialty. Finally, attending physicians are those who have completed their training and practise independently in their chosen specialty.

The following assumptions define the patients, the providers, and their interactions.

(1) There are m patients in the system. Each patient is continuously monitored. When a decline in vital signs, such as heart rate, blood pressure, or respiratory rate, is detected, the primary nurse will respond to notify the RRT or the intern, or inform both the RRT and a provider (intern, resident, fellow, and attending) for help.
(2) Once the nurse call is received, the provider should arrive immediately and carry out appropriate diagnosis and treatment. A decision will be made according to the patient's condition. The decision includes sending help requests to a higher level provider (as shown in Fig. 1), or one of the following four options: ICU, step down, tele, or stay. The RRT can only make a "stay" decision. If the attending is called for help, his/her decision is final.
(3) The probability of provider i's possible action j (making a final decision or asking for higher level help) is defined as $\alpha_{i,j}$, where $i \in \{nur, int, rrt, res, fel, int\&rrt, res\&rrt, fel\&rrt\}$, and $j \in \{rrt, int, res, fel, atn, int\&rrt, res\&rrt, fel\&rrt, atn\&rrt\}$.

Table 1 Variables

Providers	
rrt	Rapid response team
nur	Nurse
int	Intern doctor
res	Resident doctor
fel	Fellow doctor
atn	Attending doctor
Time	
τ_i	Mean response, diagnosis, and treatment time of provider i
$\tau_{k,r}$	Mean decision time including patient k's waiting time for provider r
T_{normal}	Average time period a patient is not in declining status
T_{in}	Mean decision time in Level-1 iteration, including additional waiting time
T_{final}	Mean decision time after 2-level iterations
T_d	Mean decision time
Probability	
$\alpha_{i,j}$	Probability that provider i will ask for help from provider j
p_i	Probability that response i has been carried out
$p_{k,r}$	Probability that provider r is treating patient k with another request
ρ_k	Percentage of time the patient is in a deteriorating status
λ_k	Probability patient k is declining with other patients

(4) Each patient may exhibit his/her own deteriorating characteristics. In other words, the patients' deteriorations are independent. Thus, it is possible that multiple patients decline simultaneously. However, each provider can only take care of one patient at a time.
(5) The response time (including diagnosis and possible treatment time) of provider i is modeled by a general random distribution with mean τ_i. However, if multiple patients are declining and need the specific provider at the same time, the provider will work with the first requesting patient, and other patients will wait until the current response is finished.

Remark 2 The above assumptions imply that the providers follow a first-come-first-serve policy to respond to patient deteriorations (assuming that all clinical declines have the same priority), which is typical in most acute care environments.

Remark 3 In practice, when the higher level provider is busy, the staff who initiated the request may wait or seek help from other providers, which depends on patient status, clinical protocols, and physicians' preferences, etc. The latter one is indeed considered from routing probability perspective, where the scenarios of routing to

the second provider are already included in the data. Thus, in the current model, we assume the former case (i.e., waiting) only.

Such a rapid response process can be viewed as a complex network model consisting of split, merge, and parallel features. Thus, we define a resource set of RRT, intern, resident, fellow and attending doctors, and their joint groups as $X = \{rrt, int, res, fel, atn, rrt\&int, rrt\&res, rrt\&fel, rrt\&atn\}$. Let t_d denote the decision time, i.e., from the time a decline is detected to the time a final decision is made. In addition, introduce T_d as the mean decision time. Clearly, T_d is not a simple summation of all the response times since it includes the possible unknown waiting time due to interactions between all the providers and patients. Thus, T_d is a function of all processes involved, including patients' decline, responses from all the providers, and decisions. As one can see, T_d cannot be estimated directly due to the complexity of the system. Developing a method to estimate such a time is needed.

Therefore, the problem to be addressed in this paper is formulated as: *Under assumptions (1)–(5), develop an analytical method to evaluate the mean decision time in the multiple patients rapid response system.*

As one can see, the rapid response process is complex involving multiple care providers (nurse, intern, resident, RRT, fellow, attending, and a combination of them) and various routings for response. In addition, if several patients decline simultaneously, providers will be unavailable, making the process even more complicated. Thus, direct analysis is constrained by the curse of dimensionality (e.g., using Markov chain or state-based methods), and the non-exponential nature of service time will again increase the level of difficulty. Therefore, a hierarchical structure and a two-level iteration method are proposed for this study.

Remark 4 In practice, improving operation management is of critical importance in healthcare delivery. Using the analysis method introduced in this paper, one can adjust the system parameters to predict the system performance and compare them to find an appropriate direction or strategy for operation improvement. For instance, one can adjust a provider's mean response time to find out whose response is more critical and then reduce the response time. One can also evaluate the impacts of different team configurations and compare them, such as a more experienced nurse with a quick response time working with a junior resident needing more time to response, or a recently graduated nurse working with a senior resident.

The details of the iterative method to solve the problem are presented in Sect. 4.

4 Performance analysis method

We first review the case of single patient. Then, using a three-patient example, the idea of the iteration approach is introduced. Finally, the general case is discussed.

4.1 Single patient case

When there is only one patient involved in the rapid response system, a formula to evaluate the mean decision time T_d is introduced in Xie et al. (2012).

$$T_d = \sum_{i \in X} p_i \tau_i,$$

where τ_i, $i \in X$, is the average response time of provider i, and p_i is the probability that response i, $i \in X$, has been carried out. Then p_i can be calculated as follows:

$$p_i = \begin{cases} 1, & \text{if } i = nur, \\ \alpha_{nur,i}, & \text{if } i \in \{int, rrt, rrt\&int, rrt\&res, \\ & \qquad rrt\&fel, rrt\&atn\}, \\ \alpha_{nur,res} + \sum_{j=int,rrt,rrt\&int} \alpha_{j,res} p_j, & \text{if } i = res, \\ \alpha_{res,fel} p_{res} + \alpha_{res\&rrt,fel} p_{res\&rrt}, & \text{if } i = fel, \\ \alpha_{fel,atn} p_f + \alpha_{rrt\&fel,atn} p_{rrt\&fel}, & \text{if } i = atn. \end{cases}$$

When there is only one patient, the providers are always available. If there are multiple patients and more than one patient is deteriorating simultaneously, a provider can only take care of one patient at a time so that the other patients may need to wait for additional time. To study such cases, we start with a three-patient example.

4.2 A three-patient example

When multiple patients are declining simultaneously, they may need to share the limited resource (i.e., providers). For example, as shown in Fig. 2 (where "N" and "D" represent normal and declining states, respectively), patient 2 starts declining and immediately requests help from the RRT. During the time period of RRT diagnosis and treatment to patient 2, patient 1 starts deteriorating and also asks for help from the RRT. However, patient 1 needs to wait until the RRT finishes the treatment for patient 2 and requests higher level provider's intervention. The wide dark bar represents the waiting time due to RRT sharing. Similar scenarios can be observed for all other care providers, where the patients need to share the same resource.

However, the extra waiting time due to such sharing is not easy to analyze. First, when a patient in a hospital ward may decline at any time, resource sharing can only occur when multiple patients are deteriorating during the same time period. Second, even if multiple patients are declining, they may request different providers; which resource being shared and the length of extra waiting time are still not clear. The time depends on the probability that a provider is called and his/her response time. Thus, a closed form formula to estimate waiting time is extremely difficult to obtain.

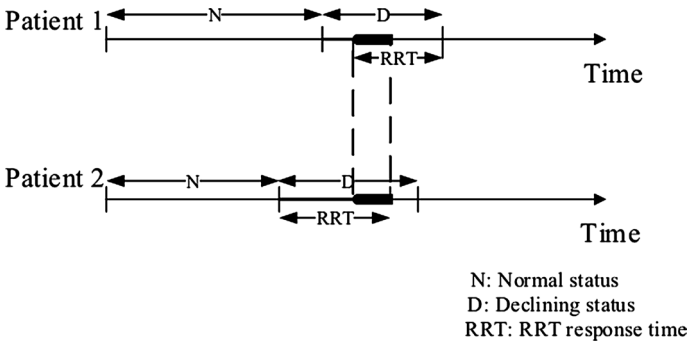

Fig. 2 RRT is shared by two patients

To solve this problem, an iteration approach is introduced. As there are two factors that initiate the waiting time: multiple patients decline simultaneously, and they all request the same provider, we introduce a two-level iteration method. First, the waiting time due to requests to the same provider is addressed. Since such analysis rely on the probability that the provider is treating other patients, which is unknown, we introduce iterations, referred to as Level 1 iteration. Secondly, using the information from Level 1, the waiting time due to multiple patients' simultaneously declining is studied. Again since this depends on another unknown probability, the probability that multiple patients are declining, we introduce iterations again, referred to as Level 2 iteration. An illustration of both Level-1 and Level-2 iterations is shown in Fig. 3.

As one can see, in Level-1 iteration, we consider each patient k iteratively. Using parameters p_i, τ_i, and T_d in single patient case, the possibility that the same provider can be requested by k patients simultaneously is investigated and the response time (including additional waiting time) for patient i is quantified. Using this result for the next patient, we calculate the similar information. Then the same process to is carried out the third patient. Afterwards, using the updated information, the procedure restarts with the first patient. Upon Level-1 iteration convergence, the mean decision time T_{in}, including T_d and the waiting time, is obtained.

Using T_{in}, and the time that the patient is in non-deteriorating condition, T_{normal}, Level-2 iteration is carried out. We calculate the probability that multiple patients are deteriorating and the mean decision time (including the waiting time) for each patient. The result is then supplied to the next patient and evaluate its probability and decision time. When all patients' information are updated, we start the next iteration. Upon convergence, the final decision time T_{final} is obtained. Below, through a three-patient example, the details of the two-level iterations are explained.

4.2.1 Level-1 iteration

In Level-1 iteration, the single patient model is used to evaluate the mean decision time, T_d, and to calculate the probability a provider is requested for help, p_i, $i \in X$, under the assumption that all providers are available.

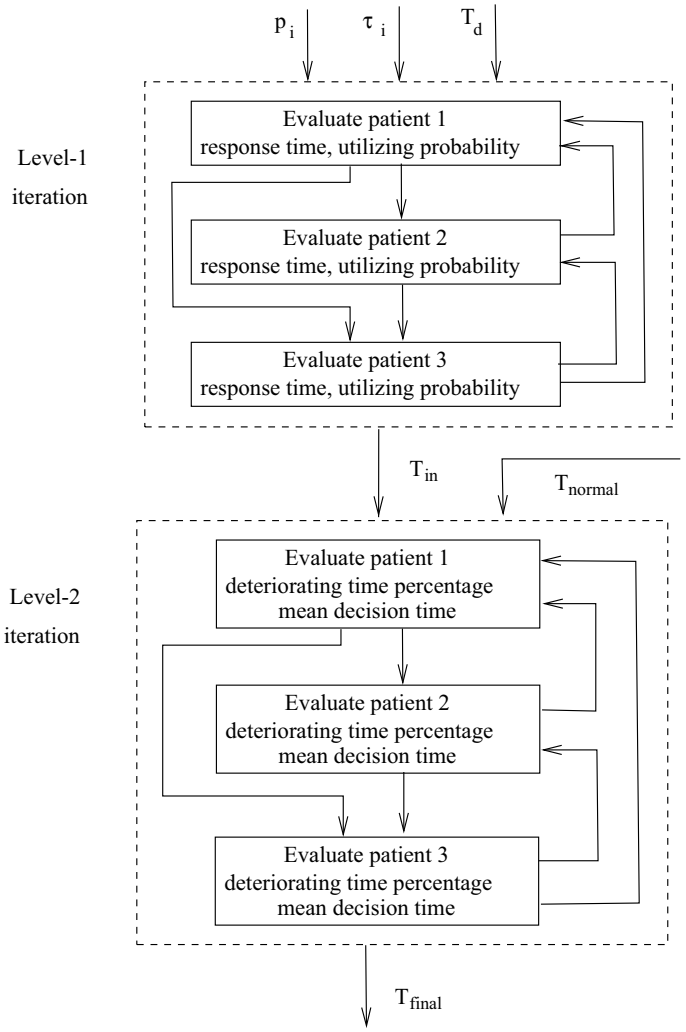

Fig. 3 Illustration of the two-level iteration procedure

Consider patient k, $k = 1, 2, 3$, and provider r, $r \in X$. Denote $\tau_{k,r}$ as the mean decision time that includes patient k's waiting time for provider r. Let $p_{k,r}$ be the probability that provider r is treating patient k and there is another request for provider r.

First, consider patient 1 requesting help from an intern. For this patient, he/she needs to wait if he/she requests help from an intern but the intern is treating the second or the third patient. We denote such probabilities as $p_{2,int}$ and $p_{3,int}$, respectively. If these probabilities are known, then the average response time of the intern includes the time when only the intern is requested, $p_{int}\tau_{int}$, and the time when both intern and RRT are requested, $p_{rrt\&int}\tau_{rrt\&int}$. Therefore, the mean decision time, $\tau_{1,int}$, will include the actual time to make decision when provider is available, T_d, and the first patient's

waiting time for the intern, $p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}$, multiplied by the probability the second or third patient is being treated by the intern, $p_{2,int} + p_{3,int}$. Therefore, we obtain:

$$\tau_{1,int} = T_d + (p_{2,int} + p_{3,int})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}).$$

Using $\tau_{1,int}$, we can evaluate $p_{1,int}$, which is the probability that the first patient is working with the intern when the second or the third patient also requests help from the intern. Again such a request can occur in both single provider (*int* only) and joint providers (both *rrt*&*int*) scenarios. For the single provider case, $p_{int}\tau_{int}/\tau_{1,int}$ represents the percentage of time that the intern is working. Analogously, for the joint providers case, $p_{rrt\&int}\tau_{rrt\&int}/\tau_{1,int}$ represents the time percentage the intern is working (jointly with RRT). Multiplied by p_{int} or $p_{rrt\&int}$, respectively, we obtain the weighted probability that the intern is serving another patient. Therefore, considering both cases, we have,

$$p_{1,int} = \frac{p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}}{\tau_{1,int}}.$$

Analogously, if we know probabilities $p_{1,int}$ and $p_{3,int}$, we can evaluate the second patient's waiting time for the intern, $\tau_{2,int}$, as well as probability $p_{2,int}$. In other words, we have

$$\tau_{2,int} = T_d + (p_{1,int} + p_{3,int})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}),$$
$$p_{2,int} = \frac{p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}}{\tau_{2,int}}.$$

Using the same logic, from probabilities $p_{1,int}$ and $p_{2,int}$, the third patient's waiting time for the intern, $\tau_{3,int}$, and probability $p_{3,int}$, can be evaluated.

$$\tau_{3,int} = T_d + (p_{1,int} + p_{2,int})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}),$$
$$p_{3,int} = \frac{p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}}{\tau_{3,int}}.$$

Since probabilities $p_{i,int}$, $i = 1, 2, 3$, are unknown, we introduce iterations to continuously update $p_{i,int}$ and $\tau_{i,int}$, $i = 1, 2, 3$, until convergence.

For resident, RRT, fellow and attending, similar updates can be carried out. Note that for RRT, there will be multiple joint service scenarios (RRT & intern, RRT & resident, RRT & fellow, and RRT & attending). A detailed description of such an iteration procedure is presented in "Appendix 1".

When the procedure converges, the mean decision time, T_{in}, which includes additional waiting time, can be calculated. This finishes Level-1 iteration.

$$T_{in} = T_d + \Sigma_{r,r\in X}P_r\mathrm{T}_r.$$

4.2.2 Level-2 iteration

Level-1 iteration provides the results if the same provider is requested by multiple patients. We also need to know when multiple patients will be declining. Thus, Level-2 iteration is carried out. Denote ρ_k, $k = 1, 2, 3$, as the percentage of time the patient is in a deteriorating status. In addition, let λ_k, $k = 1, 2, 3$, update the mean decision time, including the case patient k is declining with other patients. In other words, when patient 1 is declining, the additional waiting will occur if patient 2 or patient 3 is also declining, and such a probability can be estimated as $\rho_1(\rho_2 + \rho_3)$ if ρ_i is known. Thus, λ_1 can be evaluated as follows:

$$\lambda_1 = T_{in}[1 + \rho_1(\rho_2 + \rho_3)].$$

Note that the probability of all three patients declining is typically very small so that this scenario is ignored.

Next, calculate ρ_1 as the time percentage that a patient is in deteriorating status during a normal-declining cycle, i.e.,

$$\rho_1 = \frac{\lambda_1}{\lambda_1 + T_{normal}},$$

where T_{normal} is the average time period a patient is not in declining status.

Moving to patients 2 and 3, we obtain ρ_2 and ρ_3 using the same logic.

$$\begin{aligned}
\lambda_2 &= T_{in}[1 + \rho_2(\rho_1 + \rho_3)], \\
\rho_2 &= \frac{\lambda_2}{\lambda_2 + T_{normal}}, \\
\lambda_3 &= T_{in}[1 + \rho_3(\rho_1 + \rho_2)], \\
\rho_3 &= \frac{\lambda_3}{\lambda_3 + T_{normal}}.
\end{aligned}$$

As ρ_i, $i = 1, 2, 3$, is unknown, we introduce another iteration. Assuming all ρ_i's starting from 0, we calculate λ_i's and re-evaluate ρ_i's. The process is repeated until the procedure converges. Finally, denote T_{final} as the final value of mean decision time, we obtain

$$T_{final} = \lambda_1 = \lambda_2 = \lambda_3.$$

A detailed description of Level-2 iteration is provided in "Appendix 1" as well.

4.3 General procedure

Considering that there are m patients in the system. Using the similar idea in three-patient example, the waiting time of patient k, $k = 1, \ldots, m$, for provider r, $r \in X$, needs to consider all the possibilities that the provider is treating patient i, $i = 1, \ldots, m$, $i \neq k$. "Appendix 1" provides a formal presentation of the iteration

method, referred to as Procedure 1. Such a procedure includes two algorithms: Level-1 iterations and Level-2 iterations. The convergence of Level-1 iteration can be rigorously proved if the number of patients in the network equals to 2. The Level-2 iteration can be mathematically proved to be convergent for any number of patients. These results are presented below.

Proposition 1 *Under assumptions* (1)–(5), *when* $m = 2$, *Level-1 iteration of Procedure* 1 *is convergent, i.e.,*

$$\lim_{j\to\infty} \tau_{i,r}^{(j)} = \tau_{i,r}, \qquad \lim_{j\to\infty} p_{i,r}^{(j)} = p_{i,r}, \qquad i = 1, 2, \quad r \in X. \tag{1}$$

Proof See the "Appendix". □

Proposition 2 *Under assumptions* (1)–(5), *Level-2 iteration of Procedure* 1 *is convergent, i.e.,*

$$\lim_{j\to\infty} \lambda_i^{(l)} = \lambda_i, \qquad \lim_{j\to\infty} \rho_i^{(l)} = \rho_i, \quad i = 1, \ldots, m. \tag{2}$$

Proof See the "Appendix". □

If more than two patients present in the network, it is extremely difficult to provide a mathematical proof of convergence for Level-1 iteration due to its nature of oscillating pattern. Thus, extensive numerical investigation of the convergence of such a procedure is conducted. Numerous examples are generated by randomly selecting parameters. In all the examples, the procedure converges and a unique solution is obtained. Therefore, we formulate the results as a numerical fact:

Numerical Fact 1 *Under assumptions* (1)–(5), *Level-1 iteration of Procedure* 1 *is convergent when more than two patients present in the network, i.e.,*

$$\lim_{j\to\infty} \tau_{i,r}^{(j)} = \tau_{i,r}, \qquad \lim_{j\to\infty} p_{i,r}^{(j)} = p_{i,r}, \qquad i = 1, 2, \ldots, m. \tag{3}$$

The convergence of $\tau_{i,res}$, $\tau_{i,rrt}$, $p_{i,res}$ and $p_{i,int}$ in Level-1 iteration is illustrated in Figs. 4, 5, 6, 7. For Level-2 iteration, Figs. 8 and 9 illustrate the convergence of λ_i and ρ_i, respectively. Other variables exhibit similar convergence properties.

Clearly, in all figures, the procedure only needs 3 iterations to converge. In fact, the convergence is always observed in 3-5 iterations in all the examples we tested. The accuracy of the procedure is investigated next.

4.4 Accuracy

The accuracy of Procedure 1 has been investigated numerically. Dozens examples were generated to compare with the simulation results. In each example, uniform distribution between 20 and 40 min is assumed for the response time of each provider. The routing probabilities are randomly generated between 0 and

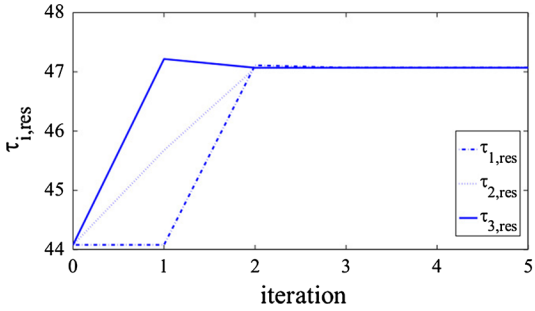

Fig. 4 Illustration of convergence of Level-1 iteration: $\tau_{i,res}$

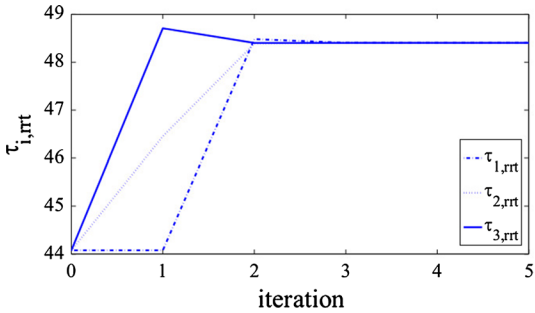

Fig. 5 Illustration of convergence of Level-1 iteration: $\tau_{i,rrt}$

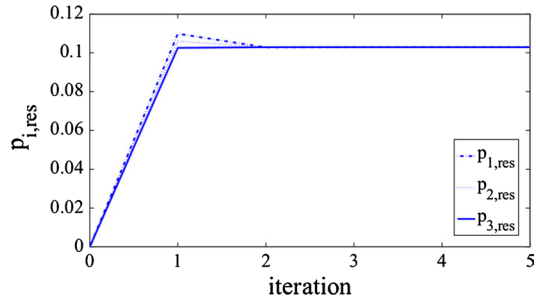

Fig. 6 Illustration of convergence of Level-1 iteration: $p_{i,res}$

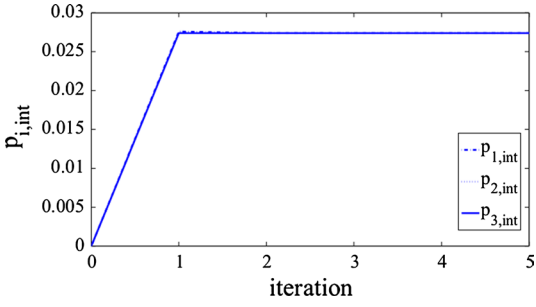

Fig. 7 Illustration of convergence of Level-1 iteration: $p_{i,int}$

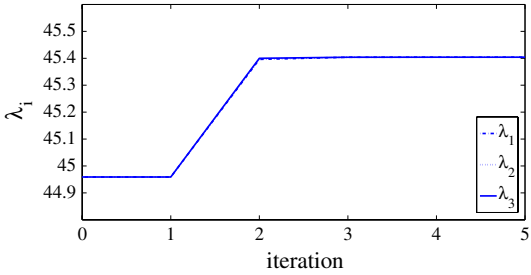

Fig. 8 Illustration of convergence of Level-2 iteration: λ_i

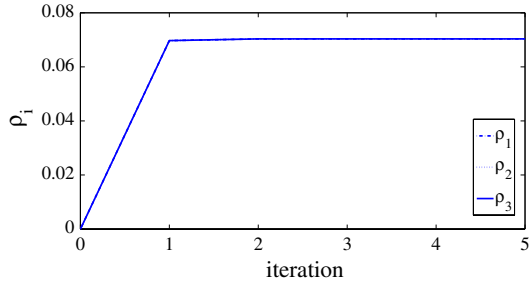

Fig. 9 Illustration of convergence of Level-2 iteration: ρ_i

1, again following uniform distribution. An exponential distribution is assumed for the patient time in the normal status. Since, in the wards, the patients are in normal status most of the time, the ratio between declining and normal status for patients is always assumed to be less than 20%. Finally, the simulations are executed using Plant Simulation 9. In the simulation model, each patient is presented as an entity. The service stations are introduced to characterize single processes. The waiting queues are described as buffers. Flow controllers are used to determine the destination of the patients based on probabilities. For each simulation experiment, 50,000 units of warm-up time are assumed. The next 5,000,000 units simulation time are carried out, and 10 replications are conducted, to ensure the confidence interval is less than 1% of the performance measure. Note that the computation time of simulation is in the order of minutes or longer, while the analytical model, programmed using Matlab R2016a, can be computed within a few seconds (see Table 2).

Remark 5 Note that the simulation speed can be increased by optimizing the setting and interface while the speed for analytical calculation can also be improved by using executable programs. The comparison in Table 2 only illustrates one aspect of the methods, the computation efficiency. There are many other aspects where simulations have an advantage, such as detailed outcomes and complexity modeling. Thus, both approaches are viable from different perspectives, and complement each other .

Table 2 Comparison of computation time

	2 patients	3 patients	4 patients	5 patients
Analytical model	2.18	2.66	12.46	15.51
Simulation	556.79	803.77	1106.99	1451.81
Ratio (simulation/analytical model)	255.41	302.17	88.84	93.60

Denote $T_{final}^{sim,i}$ and $T_{final}^{iter,i}$ as the mean decision times obtained by simulation and by Procedure 1 for example i, respectively. Then ϵ_i defines the relative difference between $T_{final}^{sim,i}$ and $T_{final}^{iter,i}$, i.e.,

$$\epsilon_i = \frac{|T_{final}^{sim,i} - T_{final}^{iter,i}|}{T_{final}^{sim,i}} \cdot 100\%.$$

The mean value of ϵ_i characterizes the average relative error and is denoted as $\bar{\epsilon}$. The results of accuracy are shown in Table 3, which provide the maximal, minimal and average accuracy as a function of number of patients and normal time, respectively.

As one can see, $\bar{\epsilon}$ is typically increasing when the normal time becomes shorter or the number of declining patients is higher. Particularly, when the normal time is not too short, i.e., more than 350 min (i.e., about 6 h), and the number of patients deteriorating is not high, e.g., less than 6, the accuracy is within 6%. Such errors may due to the heuristic updates in each iteration. When the normal time becomes short and the number of declining patients increases, such as 200 to 300 min normal time with number of patients up to 8 or 10, the accuracy decreases from 15% to 30% (even up to 50% for the worst case with 200 min normal time and 10 patients). In addition, the minimum and maximum of ϵ_i's are also included in the table. Similar trends are observed for the minimal and maximal differences. However, the cases with large discrepancies seldom happen, because these scenarios imply a substantial number of patients (e.g., 10 patients) could decline simultaneously and also quite frequently (deteriorating every 3 or 4 h), then these patients could already have been elevated to ICU or more providers have been called for help. Thus, the errors are small in most practical scenarios. We thus claim that the iteration procedure can result in an acceptable accuracy in estimating the mean decision time. In the scenarios of extreme cases, the simulation approach should be pursued to ensure the accuracy of the analysis.

From the above results, we conclude that the two-level shared resource iteration method can be used for performance evaluation of a multiple patients rapid response system.

Table 3 Accuracy of two-level iteration method, $\bar{\epsilon}$

	Normal time (min)	200	250	300	350	400	450	500
2 patients	Minimum (%)	0.09	0.06	0.04	0.02	0.02	0.04	0.06
	Average (%)	1.01	0.88	0.78	0.69	0.61	0.57	0.52
	Maximum (%)	2.30	1.96	1.65	1.39	1.27	1.35	1.40
3 patients	Minimum (%)	0.25	0.15	0.21	0.09	0.02	0.00	0.04
	Average (%)	2.36	2.03	1.76	1.51	1.33	1.18	1.04
	Maximum (%)	5.58	4.58	3.81	3.21	2.75	2.41	2.11
4 patients	Minimum (%)	0.23	0.80	0.79	0.71	0.36	0.39	0.07
	Average (%)	4.31	3.62	3.07	2.64	2.23	1.99	1.68
	Maximum (%)	10.02	7.84	6.41	5.42	4.51	3.95	3.51
5 patients	Minimum (%)	0.39	0.18	0.69	1.20	1.25	0.78	0.74
	Average (%)	7.28	5.70	4.79	3.95	3.50	2.91	2.61
	Maximum (%)	15.99	11.98	9.67	7.62	6.80	5.94	4.75
6 patients	Minimum (%)	3.27	2.40	1.60	0.53	0.11	0.58	1.36
	Average (%)	12.42	9.36	7.59	6.14	5.22	4.47	4.13
	Maximum (%)	24.49	17.69	14.14	11.33	9.71	8.53	6.84
8 patients	Minimum (%)	11.35	8.13	5.69	4.54	3.19	2.30	1.64
	Average (%)	29.59	19.10	14.83	12.06	10.10	8.73	7.66
	Maximum (%)	57.48	33.47	25.05	20.13	16.64	14.36	12.75
10 patients	Minimum (%)	24.78	15.99	11.05	8.98	6.84	5.32	4.49
	Average (%)	54.32	35.62	24.43	19.30	15.81	13.61	11.71
	Maximum (%)	74.21	63.67	40.16	30.96	25.00	21.20	18.25

	Normal time (min)	550	600	650	700	750	800	850	900
2 patients	Minimum (%)	0.09	0.07	0.01	0.04	0.12	0.02	0.01	0.06
	Average (%)	0.51	0.46	0.44	0.43	0.44	0.41	0.41	0.41
	Maximum (%)	1.51	1.54	1.59	1.63	1.69	1.71	1.74	1.77

Table 3 (continued)

	Normal time (min)	550	600	650	700	750	800	850	900
3 patients	Minimum (%)	0.10	0.21	0.17	0.05	0.11	0.19	0.10	0.03
	Average (%)	0.97	0.91	0.84	0.79	0.76	0.74	0.71	0.72
	Maximum (%)	2.24	2.38	2.43	2.57	2.63	2.69	2.72	2.84
4 patients	Minimum (%)	0.07	0.14	0.01	0.12	0.34	0.19	0.21	0.21
	Average (%)	1.46	1.35	1.24	1.21	1.17	0.97	0.95	0.92
	Maximum (%)	3.04	2.94	2.88	3.34	3.35	3.17	3.24	3.44
5 patients	Minimum (%)	0.05	0.12	0.13	0.02	0.25	0.41	0.05	0.07
	Average (%)	2.17	1.97	1.75	1.52	1.43	1.33	1.37	1.30
	Maximum (%)	4.38	3.71	3.14	3.23	3.01	3.16	3.99	3.79
6 patients	Minimum (%)	1.48	0.87	0.66	0.49	0.09	0.43	0.15	0.23
	Average (%)	3.80	3.06	2.77	2.54	2.17	2.17	1.96	1.57
	Maximum (%)	6.62	6.04	5.40	4.96	4.54	4.10	3.97	3.62
8 patients	Minimum (%)	1.29	0.91	0.26	0.12	1.22	1.02	1.40	1.53
	Average (%)	6.73	5.88	5.23	4.77	4.41	4.04	3.66	3.49
	Maximum (%)	11.42	10.15	8.36	8.40	7.37	8.05	7.27	5.66
10 patients	Minimum (%)	3.87	2.30	2.61	1.81	0.49	1.50	0.44	0.47
	Average (%)	10.53	9.45	8.41	7.43	6.82	6.27	6.25	5.16
	Maximum (%)	16.35	15.34	13.05	11.80	11.26	10.30	10.26	9.31

4.5 Distribution sensitivity analysis

In the accuracy study, an exponential distribution is assumed in simulations for the normal time, i.e., when a patient is not declining. In practice, such times may not exhibit exponential behavior. Investigating the case of non-exponential normal time is necessary. Therefore, gamma and lognormal distributions were used since one can easily alter their coefficients of variation (CV). In addition, in the rapid response process, a patient's risk of deterioration becomes higher as the time elapses, which leads to CV smaller than 1 (Li and Meerkov 2005). Therefore, we focus on four data points, $CV = 0.25, 0.5, 0.75$ and 1. First, additional accuracy studies using Lognormal and Gamma distributions with CV = 0.25, 0.5, and 0.75 are carried out. As shown in Table 4, the average accuracy is at the same level as exponential normal time.

Second, we hypothesize that the variability's impact on the mean decision time will be small. A dozen examples assuming Lognormal distributions were randomly generated and the largest possible relative error was recorded. Denote $T_{i,j}$ as the mean decision time obtained from simulation, where j represents the experiment number, and i indicates the CV value, where $i = 1, 2, 3, 4$ refer to $CV = 0.25, 0.5, 0.75, 1$, respectively. Then the difference between the largest and smallest mean decision time in experiment j for any given CV is defined as δ_j, which represents the maximal deviation under different variability.

$$\delta_j = \frac{\max_i T_{i,j} - \min_i T_{i,j}}{\min_i T_{i,j}} \cdot 100\%.$$

Table 4 Accuracy of two-level iteration method: non-exponential case

Normal time (min)	300	350	400	450	500	550	600
(*a*) *Lognormal distribution*							
2 patients (%)	0.68	0.61	0.58	0.53	0.50	0.48	0.44
3 patients (%)	1.43	1.28	1.15	1.05	0.97	0.90	0.84
4 patients (%)	2.54	2.22	1.94	1.70	1.51	1.39	1.25
5 patients (%)	3.93	3.38	2.92	2.56	2.22	1.92	1.73
6 patients (%)	6.26	5.27	4.56	4.12	3.66	3.12	2.82
8 patients (%)	12.83	10.48	8.86	7.63	6.71	5.84	5.24
10 patients (%)	22.15	17.25	14.29	12.17	10.68	9.62	9.02
(*b*) *Gamma distribution*							
2 patients (%)	0.72	0.64	0.58	0.55	0.50	0.49	0.46
3 patients (%)	1.55	1.36	1.23	1.10	0.99	0.93	0.87
4 patients (%)	2.71	2.53	2.05	1.80	1.57	1.38	1.23
5 patients (%)	4.28	3.58	3.12	2.67	2.39	2.04	1.82
6 patients (%)	6.68	5.58	4.80	4.21	3.72	3.33	2.93
8 patients (%)	13.44	10.99	9.28	8.02	7.03	6.15	5.44
10 patients (%)	22.65	17.77	14.81	12.64	11.03	9.77	8.76

Table 5 Accuracy of two-level iteration: lognormal distribution case, $\bar{\delta}$

Normal time (min)	300	350	400	450	500	550	600
2 patients (%)	0.09	0.07	0.07	0.06	0.04	0.05	0.05
3 patients (%)	0.27	0.16	0.10	0.09	0.07	0.06	0.06
4 patients (%)	0.51	0.35	0.26	0.21	0.18	0.19	0.21
5 patients (%)	0.85	0.55	0.25	0.31	0.28	0.21	0.26
6 patients (%)	1.38	0.81	0.49	0.39	0.28	0.33	0.28
8 patients (%)	3.22	1.82	1.00	0.57	0.46	0.43	0.42
10 patients (%)	6.62	3.78	2.09	1.21	0.92	0.53	0.50

Table 6 Accuracy of two-level iteration: gamma distribution case, $\bar{\delta}$

Normal time (min)	300	350	400	450	500	550	600
2 patients (%)	0.23	0.16	0.13	0.11	0.09	0.08	0.06
3 patients (%)	0.51	0.36	0.29	0.22	0.19	0.15	0.12
4 patients (%)	0.91	0.66	0.44	0.42	0.31	0.26	0.29
5 patients (%)	1.46	0.93	0.83	0.56	0.48	0.37	0.37
6 patients (%)	2.12	1.38	1.01	0.80	0.64	0.63	0.41
8 patients (%)	4.32	2.82	1.86	1.42	1.11	0.93	0.70
10 patients (%)	7.68	4.76	3.14	2.25	1.56	1.28	1.12

Using the average value of δ_j, denoted as $\bar{\delta}$, we study the impact of non-exponential normal time.

$$\bar{\delta} = \frac{\sum_{j=1}^{10} \frac{|\max_i T_{i,j} - \min_i T_{i,j}|}{\min_i T_{i,j}}}{10} \cdot 100\%.$$

The results are presented in Table 5, where normal time is between 300 to 600 min under scenarios of 2 to 10 patients. It can be seen that the differences ($\bar{\delta}$'s) are quite small, less than 0.5% for the cases of 2–4 patients and up to 3.78% for more-patient cases except the worst one (with 300 min normal time and 10 patients, making the difference as high as 6.62%).

Similarly, the results of a gamma distribution of normal status time are shown in Table 6. Again $\bar{\delta}$ is smaller than 1% for cases up to 5 patients. If more patients are deteriorating, the errors can increase to 4.76%, and the worst one (seldom happens) is 7.68% with 300 min normal time and 10 patients. These results are also satisfactory to support the hypothesis that the model reasonably represents reality. Therefore, we conclude that, even with different patients' declining distributions, the iteration method introduced in this paper can provide an acceptable estimation of mean decision time in the multiple patients system.

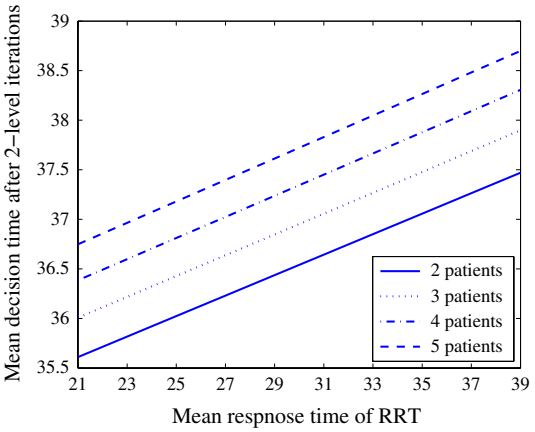

Fig. 10 Monotonicity of mean decision time with respect to response time

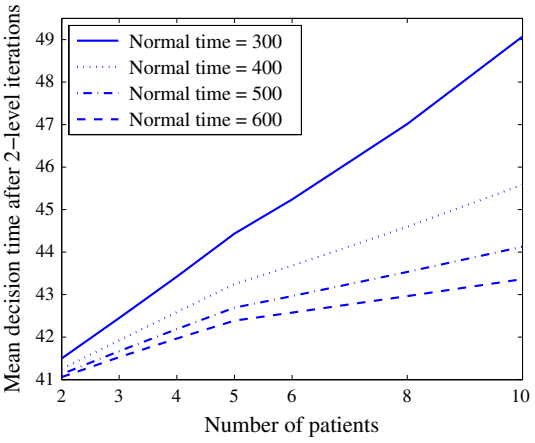

Fig. 11 Monotonicity of mean decision time with respect to number of patients

4.6 Monotonicity

Using the iteration method introduced above, we can efficiently investigate the monotonic properties of decision time with respect to its parameters, such as a provider's response time, number of patients, and the normal time. Based on extensive numerical experiments, we observe:

Numerical Fact 2 *Under assumptions* (1)–(5), *them mean decision time* T_d *is monotonically increasing with respect to each provider's response time* τ_i, *normal time* T_{normal}, *and number of patients* m.

An illustration of such monotonicity is shown in Figs. 10, 11, 12. As one can see, the monotonic properties can provide the direction of operation improvement to reduce mean decision time. Decreasing the RRT's response time, the number of

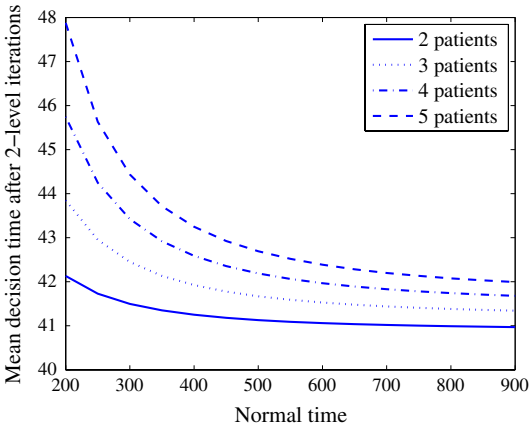

Fig. 12 Monotonicity of mean decision time with respect to normal time

patients (the providers are responsible for), and increasing normal time, all lead to reduction of mean decision time.

Using such properties, we can evaluate the impact of improvement efforts to identify the most critical factor that will lead to the largest improvement, which is referred to as bottleneck.

Definition 1 The provider response time τ_i is the bottleneck response time if

$$\left|\frac{\partial T_d}{\partial \tau_i}\right| > \left|\frac{\partial T_d}{\partial \tau_j}\right|, \quad \forall j \neq i.$$

Since even the evaluation of T_d is difficult, calculating the partial derivatives becomes all but impossible. Therefore, sensitivity analysis is carried out. Specifically, response time τ_i becomes the bottleneck if

$$T_d(\tau_i - \eta\tau_i) > T_d(\tau_j - \eta\tau_j), \quad \forall j \neq i,$$

where $0 < \eta \ll 1$.

After identifying the bottleneck, improvement efforts can be focused on how to reduce the bottleneck response time. For instance, assigning patients with specific diseases to residents who have more experience, reducing the frequency calling for residents, etc., could be investigated. As one can see, this will requires repeated calculation and comparison. Thus, the performance evaluation method introduced in this paper enables a quick analysis in such activities.

Remark 6 The above model provides a quantitative tool for hospital management to design continuous improvement activities. Note that due to shortage and/or multiple job functions of critical care providers and nurses (Buchman et al. 2017), and restricted rules on their duty hours (Meyers et al. 2017), adding more staff is typically difficult to achieve. However, the system performance can be improved by reorganize the workforce to find out the optimal or improved option of team configuration of staffs, such as pairing a more experienced nurse with a new resident doctor.

5 Conclusions

This paper introduces the study of rapid response system with multiple patients and limited provider availability. An iteration method is introduced to evaluate the mean decision time for multiple simultaneously declining patients. The convergence of the iteration procedure is justified both analytically and numerically. It is shown that the procedure converges within a few iterations and a reasonable accuracy is obtained in the test cases. Such a method presents an effective quantitative tool for performance evaluation of multiple patients rapid response system. The model and outcome of this study have been well received by healthcare professionals.

Clearly, the proposed method also exists limitations. In future work, we plan to address these limitations. Specifically, we will further explore to completely prove the convergence of the recursive procedure. Also, analysis of the systems with more complex structures should be conducted, for instance, multiple same type providers may work on the floor simultaneously, providers may seek help from multiple higher level resources at the same. In addition to average decision time, the variabilities in decision time, such as coefficients of variation and response-time performance (probability to make a decision within a given time interval) are also critical. Developing methods to evaluate the variability is strongly needed. Moreover, efforts can be devoted to evaluating and comparing the impacts of different team configurations to design appropriate staffing policy. Besides, patients are critical elements in healthcare delivery. Introducing a patient model to characterize the dynamic behavior of the patient and declining status is necessary, and such a model should be integrated with the response model. Furthermore, other methods, both simulations and analytical models, such as Petri Nets, Markov chains, should be investigated. Finally, more insights, patterns, and protocol implications should be derived from the analytical study, and all the developed methods and models will be validated and applied on the hospital floor. The successful development of these works can provide hospital management quantitative tools and decision support to improve patient safety and quality of care.

Acknowledgements This work is supported in part by National Science Foundation Grant No. CMMI-1536987 and by National Natural Science Foundation of China Grant No. 71501109.

Appendix 1: Iteration procedures

Three-patient example: Level-1 iteration procedure

Denote $\tau_{k,r}^{(j)}$, $k = 1, 2, 3$, $r \in X$, as the mean decision time that includes patient k's waiting time for provider r during the j-th iteration, $j = 1, 2, \ldots$. Let $p_{k,r}^{(j)}$ be the probability that provider r is treating patient k and there is another request for provider r during the j-th iteration. At the beginning of iteration, assume

$$\tau_{k,r}^{(0)} = T_d \quad \text{and} \quad p_{k,r}^{(0)} = 0, \qquad k = 1, 2, 3, \quad r \in X.$$

First, consider patient 1. During the first iteration, $\tau_{1,int}^{(1)}$ can be updated as:

$$\tau_{1,int}^{(1)} = T_d + (p_{2,int}^{(0)} + p_{3,int}^{(0)})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}).$$

The $p_{1,int}^{(1)}$ can be updated as:

$$p_{1,int}^{(1)} = \frac{p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}}{\tau_{1,int}^{(1)}}.$$

Next, Consider patient 2. Decision time $\tau_{2,int}^{(1)}$ and probability $p_{2,int}^{(1)}$ can be calculated.

$$\tau_{2,int}^{(1)} = T_d + (p_{1,int}^{(1)} + p_{3,int}^{(0)})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}),$$
$$p_{2,int}^{(1)} = \frac{p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}}{\tau_{2,int}^{(1)}}.$$

Lastly, consider patient 3, we have

$$\tau_{3,int}^{(1)} = T_d + (p_{1,int}^{(1)} + p_{2,int}^{(1)})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}),$$
$$p_{3,int}^{(1)} = \frac{p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}}{\tau_{3,int}^{(1)}}.$$

This completes the update of the intern.

Similar updating process for the resident can be carried out. First, we study patient 1:

$$\tau_{1,res}^{(1)} = T_d + (p_{2,res}^{(0)} + p_{3,res}^{(0)})(p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}),$$
$$p_{1,res}^{(1)} = \frac{p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res}}{\tau_{1,res}^{(1)}}.$$

Next, consider patient 2:

$$\tau_{2,res}^{(1)} = T_d + (p_{1,res}^{(1)} + p_{3,res}^{(0)})(p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}),$$
$$p_{2,res}^{(1)} = \frac{p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res}}{\tau_{2,res}^{(1)}}.$$

Then, patient 3 is included:

$$\tau_{3,res}^{(1)} = T_d + (p_{1,res}^{(1)} + p_{2,res}^{(1)})(p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}),$$
$$p_{3,res}^{(1)} = \frac{p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res}}{\tau_{3,res}^{(1)}}.$$

Similarly, all the rest of providers are updated. Particularly, for the RRT, considering patient 1, we obtain:

$$\begin{aligned}\tau^{(1)}_{1,rrt} &= T_d + (p^{(0)}_{2,rrt} + p^{(0)}_{3,rrt})(p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} + p_{rrt\&res}\tau_{rrt\&res} \\ &\quad + p_{rrt\&fel}\tau_{rrt\&fel} + p_{rrt\&atn}\tau_{rrt\&atn}), \\ p^{(1)}_{1,rrt} &= (p^2_{rrt}\tau_{rrt} + p^2_{rrt\&int}\tau^2_{rrt\&int} + p^2_{rrt\&res}\tau_{rrt\&res} + p^2_{rrt\&fel}\tau_{rrt\&fel} \\ &\quad + p^2_{rrt\&atn}\tau_{rrt\&atn})/\tau^{(1)}_{1,rrt}.\end{aligned}$$

Regarding patient 2, we have

$$\begin{aligned}\tau^{(1)}_{2,rrt} &= T_d + (p^{(1)}_{1,rrt} + p^{(0)}_{3,rrt})(p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} + p_{rrt\&res}\tau_{rrt\&res} \\ &\quad + p_{rrt\&fel}\tau_{rrt\&fel} + p_{rrt\&atn}\tau_{rrt\&atn}), \\ p^{(1)}_{2,rrt} &= (p^2_{rrt}\tau_{rrt} + p^2_{rrt\&int}\tau^2_{rrt\&int} + p^2_{rrt\&res}\tau_{rrt\&res} + p^2_{rrt\&fel}\tau_{rrt\&fel} \\ &\quad + p^2_{rrt\&atn}\tau_{rrt\&atn})/\tau^{(1)}_{2,rrt}.\end{aligned}$$

Furthermore, parameters of patient 3 are updated:

$$\begin{aligned}\tau^{(1)}_{3,rrt} &= T_d + (p^{(1)}_{1,rrt} + p^{(1)}_{2,rrt})(p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} + p_{rrt\&res}\tau_{rrt\&res} \\ &\quad + p_{rrt\&fel}\tau_{rrt\&fel} + p_{rrt\&atn}\tau_{rrt\&atn}), \\ p^{(1)}_{3,rrt} &= (p^2_{rrt}\tau_{rrt} + p^2_{rrt\&int}\tau^2_{rrt\&int} + p^2_{rrt\&res}\tau_{rrt\&res} + p^2_{rrt\&fel}\tau_{rrt\&fel} \\ &\quad + p^2_{rrt\&atn}\tau_{rrt\&atn})/\tau^{(1)}_{3,rrt}.\end{aligned}$$

Then for the fellow, patients 1 to 3 are considered:

$$\begin{aligned}\tau^{(1)}_{1,fel} &= T_d + (p^{(0)}_{2,fel} + p^{(0)}_{3,fel})(p_{fel}\tau_{fel} + p_{rrt\&fel}\tau_{rrt\&fel}), \\ p^{(1)}_{1,fel} &= \frac{p^2_{fel}\tau_{fel} + p^2_{rrt\&fel}\tau_{rrt\&fel}}{\tau^{(1)}_{1,fel}}, \\ \tau^{(1)}_{2,fel} &= T_d + (p^{(1)}_{1,fel} + p^{(0)}_{3,fel})(p_{fel}\tau_{fel} + p_{rrt\&fel}\tau_{rrt\&fel}), \\ p^{(1)}_{2,fel} &= \frac{p^2_{fel}\tau_{fel} + p^2_{rrt\&fel}\tau_{rrt\&fel}}{\tau^{(1)}_{2,fel}}, \\ \tau^{(1)}_{2,fel} &= T_d + (p^{(1)}_{1,fel} + p^{(1)}_{2,fel})(p_{fel}\tau_{fel} + p_{rrt\&fel}\tau_{rrt\&fel}), \\ p^{(1)}_{2,fel} &= \frac{p^2_{fel}\tau_{fel} + p^2_{rrt\&fel}\tau_{rrt\&fel}}{\tau^{(1)}_{3,fel}}.\end{aligned}$$

Finally, for the attending, we again address all three patients:

$$\tau_{1,atn}^{(1)} = T_d + (p_{2,atn}^{(0)} + p_{3,atn}^{(0)})(p_{atn}\tau_{atn} + p_{rrt\&atn}\tau_{rrt\&atn}),$$
$$p_{1,atn}^{(1)} = \frac{p_{atn}^2\tau_{atn} + p_{rrt\&atn}^2\tau_{rrt\&atn}}{\tau_{1,atn}^{(1)}},$$
$$\tau_{2,atn}^{(1)} = T_d + (p_{1,atn}^{(1)} + p_{3,atn}^{(0)})(p_{atn}\tau_{atn} + p_{rrt\&atn}\tau_{rrt\&atn}),$$
$$p_{2,atn}^{(1)} = \frac{p_{atn}^2\tau_{atn} + p_{rrt\&atn}^2\tau_{rrt\&atn}}{\tau_{2,atn}^{(1)}},$$
$$\tau_{3,atn}^{(1)} = T_d + (p_{1,atn}^{(1)} + p_{2,atn}^{(1)})(p_{atn}\tau_{atn} + p_{rrt\&atn}\tau_{rrt\&atn}),$$
$$p_{3,atn}^{(1)} = \frac{p_{atn}^2\tau_{atn} + p_{rrt\&atn}^2\tau_{rrt\&atn}}{\tau_{3,atn}^{(1)}}.$$

When the first iteration is finished, all the updated parameters will be used for the second iteration to calculate $\tau_{k,r}^{(2)}$, $k = 1, 2, 3$, $r \in X$, and $p_{k,r}^{(2)}$. The process is repeated until procedure converges. Let $\delta = 10^{-5}$. When

$$|\tau_{i,r}^{(j+1)} - \tau_{i,r}^{(j)}| \le \delta, \qquad |p_{i,r}^{(j+1)} - p_{i,r}^{(j)}| \le \delta, \qquad i = 1, 2, 3, \quad r \in X,$$

the procedure is convergent, i.e.,

$$\lim_{j\to\infty} \tau_{i,r}^{(j)} = \tau_{i,r}, \qquad \lim_{j\to\infty} p_{i,r}^{(j)} = p_{i,r}, \qquad i = 1, 2, 3.$$

In particular, all $\tau_{i,r}, i = 1, 2, 3$, are identical and all $p_{i,r}, i = 1, 2, 3$, are the same. Then the mean decision time (including waiting time) T_r and provider utilization P_r can be obtained:

$$\tau_{1,r} = \tau_{2,r} = \tau_{3,r} := \mathrm{T}_r, \qquad p_{1,r} = p_{2,r} = p(3, r) := P_r.$$

The mean decision time T_{in} includes the additional waiting time.

$$T_{in} = T_d + \Sigma_{r,r\in X} P_r \mathrm{T}_r.$$

Three-patient example: Level-2 iteration procedure

Denote $\rho_k^{(l)}$, $k = 1, 2, 3$, $l = 1, 2, \ldots$, as the percentage of time the patient is in a deteriorating status in iteration j, and $\lambda_k^{(l)}$, $k = 1, 2, 3$, $l = 1, 2, \ldots$, as the updated mean decision time in iteration j by including the time percentage patient k is declining. When the iteration starts, assume all

$$\rho_k^{(0)} = 0 \quad \text{and} \quad \lambda_k^{(0)} = T_{in}, \qquad k = 1, 2, 3.$$

Considering patient 1, $\lambda_1^{(1)}$ can be updated as:

$$\lambda_1^{(1)} = T_{in}\left[1 + \rho_1^{(0)}\left(\rho_2^{(0)} + \rho_3^{(0)}\right)\right].$$

The time percentage that the first patient is in deteriorating status can be calculated as

$$\rho_1^{(1)} = \frac{\lambda_1^{(1)}}{\lambda_1^{(1)} + T_{normal}}.$$

Next consider patients 2 and 3, where $\lambda_i^{(1)}$ and $\rho_i^{(1)}$, $i = 2, 3$, can be obtained:

$$\begin{aligned}
\lambda_2^{(1)} &= T_{in}\left[1 + \rho_2^{(0)}\left(\rho_1^{(1)} + \rho_3^{(0)}\right)\right],\\
\rho_2^{(1)} &= \frac{\lambda_2^{(1)}}{\lambda_2^{(1)} + T_{normal}},\\
\lambda_3^{(1)} &= T_{in}\left[1 + \rho_3^{(0)}\left(\rho_1^{(1)} + \rho_2^{(1)}\right)\right],\\
\rho_3^{(1)} &= \frac{\lambda_3^{(1)}}{\lambda_3^{(1)} + T_{normal}}.
\end{aligned}$$

This finishes the first iteration. Then $\rho_k^{(1)}$, $k = 1, 2, 3$, and $\lambda_k^{(1)}$ are used for the second iteration to evaluate $\rho_k^{(2)}$ and $\lambda_k^{(2)}$. The process is repeated until the procedure converges. When the following criteria is met:

$$|\lambda_i^{(j+1)} - \lambda_i^{(l)}| \le \delta, \qquad |\rho_i^{(j+1)} - \rho_i^{(l)}| \le \delta, \qquad i = 1, 2, 3,$$

the procedure is convergent. Again $\delta = 10^{-5}$. Upon converges, we have

$$\lim_{l\to\infty} \lambda_i^{(l)} = \lambda_i, \qquad \lim_{l\to\infty} \rho_i^{(l)} = \rho_i, \qquad i = 1, 2, 3.$$

The final mean decision time can be obtained:

$$\lambda_1 = \lambda_2 = \lambda_3 = T_{final}.$$

General iteration procedure

Procedure 1 (1) Level-1 iteration

Step 1.1 Initialization: Calculate p_i, $i \in X$, and T_d using the results in Xie et al. (2012). Set $j = 0$ and

$$\tau_{k,i}^{(j)} = p_{k,i}^{(j)} = 0.$$

Step 1.2 Update $\tau_{k,i}^{(j)}$ and $p_{k,i}^{(j)}$: For patient 1,

$$
\begin{aligned}
\tau_{1,int}^{(j+1)} &= T_d + \Sigma_{i=2}^{n} p_{i,int}^{(j)} (p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}),\\
p_{1,int}^{(j+1)} &= (p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}) \big/ \tau_{1,int}^{(j+1)},\\
\tau_{1,res}^{(j+1)} &= T_d + \Sigma_{i=2}^{n} p_{i,res}^{(j)} (p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}),\\
p_{1,res}^{(j+1)} &= (p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res}) \big/ \tau_{1,res}^{(j+1)},\\
\tau_{1,rrt}^{(j+1)} &= T_d + \Sigma_{i=2}^{n} p_{i,rrt}^{(j)} (p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} + p_{rrt\&res}\tau_{rrt\&res}\\
&\quad + p_{rrt\&fel}\tau_{rrt\&fel} + p_{rrt\&atn}\tau_{rrt\&atn}),\\
p_{1,rrt}^{(j+1)} &= (p_{rrt}^2\tau_{rrt} + p_{rrt\&int}^2\tau_{rrt\&int}^2 + p_{rrt\&res}^2\tau_{rrt\&res}\\
&\quad + p_{rrt\&fel}^2\tau_{rrt\&fel} + p_{rrt\&atn}^2\tau_{rrt\&atn}) \big/ \tau_{1,rrt}^{(j+1)},\\
\tau_{1,fel}^{(j+1)} &= T_d + \Sigma_{i=2}^{n} p_{i,fel}^{(j)} (p_{fel}\tau_{fel} + p_{rrt\&fel}\tau_{rrt\&fel}),\\
p_{1,fel}^{(j+1)} &= (p_{fel}^2\tau_{fel} + p_{rrt\&fel}^2\tau_{rrt\&fel}) \big/ \tau_{1,fel}^{(j+1)},\\
\tau_{1,atn}^{(j+1)} &= T_d + \Sigma_{i=2}^{n} p_{i,atn}^{(j)} (p_{atn}\tau_{atn} + p_{rrt\&atn}\tau_{rrt\&atn}),\\
p_{1,atn}^{(j+1)} &= (p_{atn}^2\tau_{atn} + p_{rrt\&atn}^2\tau_{rrt\&atn}) \big/ \tau_{1,atn}^{(j+1)}.
\end{aligned}
\tag{4}
$$

For patient $k = 2, \ldots, m-1$,

$$
\begin{aligned}
\tau_{k,int}^{(j+1)} &= T_d + \left(\Sigma_{i=1}^{k-1} p_{i,int}^{(j+1)} + \Sigma_{i=k+1}^{m} p_{i,int}^{(j)}\right) \cdot (p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}),\\
p_{k,int}^{(j+1)} &= (p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}) \big/ \tau_{k,int}^{(j+1)},\\
\tau_{k,res}^{(j+1)} &= T_d + \left(\Sigma_{i=1}^{k-1} p_{i,res}^{(j+1)} + \Sigma_{i=k+1}^{m} p_{i,res}^{(j)}\right) \cdot (p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}),\\
p_{k,res}^{(j+1)} &= (p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res}) \big/ \tau_{k,res}^{(j+1)},\\
\tau_{k,rrt}^{(j+1)} &= T_d + \left(\Sigma_{i=1}^{k-1} p_{i,rrt}^{(j+1)} + \Sigma_{i=k+1}^{m} p_{i,rrt}^{(j)}\right) \cdot (p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int}\\
&\quad + p_{rrt\&res}\tau_{rrt\&res} + p_{rrt\&fel}\tau_{rrt\&fel} + p_{rrt\&atn}\tau_{rrt\&atn}),\\
p_{k,rrt}^{(j+1)} &= (p_{rrt}^2\tau_{rrt} + p_{rrt\&int}^2\tau_{rrt\&int}^2 + p_{rrt\&res}^2\tau_{rrt\&res} + p_{rrt\&fel}^2\tau_{rrt\&fel}\\
&\quad + p_{rrt\&atn}^2\tau_{rrt\&atn}) / \tau_{k,rrt}^{(j+1)},\\
\tau_{k,fel}^{(j+1)} &= T_d + (\Sigma_{i=1}^{k-1} p_{i,fel}^{(j+1)} + \Sigma_{i=k+1}^{m} p_{i,fel}^{(j)}) \cdot (p_{fel}\tau_{fel} + p_{rrt\&fel}\tau_{rrt\&fel}),\\
p_{k,fel}^{(j+1)} &= (p_{fel}^2\tau_{fel} + p_{rrt\&fel}^2\tau_{rrt\&fel}) \big/ \tau_{k,fel}^{(j+1)},\\
\tau_{k,atn}^{(j+1)} &= T_d + \left(\Sigma_{i=1}^{k-1} p_{i,atn}^{(j+1)} + \Sigma_{i=k+1}^{m} p_{i,atn}^{(j)}\right) \cdot (p_{atn}\tau_{atn} + p_{rrt\&atn}\tau_{rrt\&atn}),\\
p_{k,atn}^{(j+1)} &= (p_{atn}^2\tau_{atn} + p_{rrt\&atn}^2\tau_{rrt\&atn}) \big/ \tau_{k,atn}^{(j+1)}.
\end{aligned}
\tag{5}
$$

For patient m,

$$
\begin{aligned}
\tau_{m,int}^{(j+1)} &= T_d + \Sigma_{i=1}^{m-1} p_{i,int}^{(j+1)}(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}),\\
p_{m,int}^{(j+1)} &= \frac{p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}}{\tau_{k,r}^{(j+1)}},\\
\tau_{m,res}^{(j+1)} &= T_d + \Sigma_{i=1}^{m-1} p_{i,res}^{(j+1)}(p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}),\\
p_{m,res}^{(j+1)} &= (p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res})\big/\tau_{k,res}^{(j+1)},\\
\tau_{n,rrt}^{(j+1)} &= T_d + \Sigma_{i=1}^{m-1} p_{i,rrt}^{(j+1)}(p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} + p_{rrt\&res}\tau_{rrt\&res}\\
&\quad + p_{rrt\&fel}\tau_{rrt\&fel} + p_{rrt\&atn}\tau_{rrt\&atn}),\\
p_{m,r}^{(j+1)} &= (p_{rrt}^2\tau_{rrt} + p_{rrt\&int}^2\tau_{rrt\&int}^2 + p_{rrt\&res}^2\tau_{rrt\&res} + p_{rrt\&fel}^2\tau_{rrt\&fel}\\
&\quad + p_{rrt\&atn}^2\tau_{rrt\&atn})\big/\tau_{m,rrt}^{(j+1)},\\
\tau_{m,fel}^{(j+1)} &= T_d + \Sigma_{i=1}^{m-1} p_{i,fel}^{(j+1)}(p_{fel}\tau_{fel} + p_{rrt\&fel}\tau_{rrt\&fel}),\\
p_{m,fel}^{(j+1)} &= (p_{fel}^2\tau_f + p_{rrt\&fel}^2\tau_{rrt\&fel})\big/\tau_{m,fel}^{(j+1)},\\
\tau_{m,atn}^{(j+1)} &= T_d + \Sigma_{i=1}^{m-1} p_{i,atn}^{(j+1)}(p_{atn}\tau_{atn} + p_{rrt\&atn}\tau_{rrt\&atn}),\\
p_{m,atn}^{(j+1)} &= (p_{atn}^2\tau_{atn} + p_{rrt\&atn}^2\tau_{rrt\&atn})\big/\tau_{m,atn}^{(j+1)}.
\end{aligned}
\tag{6}
$$

Step 1.3 Iteration: Set $j = j + 1$. If the terminating criteria is not met, go back to Step 1.2. Let $\delta = 10^{-5}$, the Level-1 iteration is finished if

$$
|\tau_{i,r}^{(j+1)} - \tau_{i,r}^{(j)}| \le \delta, \qquad |p_{i,r}^{(j+1)} - p_{i,r}^{(j)}| \le \delta, \qquad i = 1, 2, \ldots, m.
$$

Step 1.4 Termination: If the stopping conditions are met, set

$$
\begin{aligned}
\tau_{i,r}^{(j+1)} &= \mathrm{T}_r, \qquad p_{i,r}^{(j+1)} = P_r, \qquad i = 1, \ldots, m,\\
T_{in} &= T_d + \Sigma_{r,r\in X} P_r \mathrm{T}_r.
\end{aligned}
\tag{7}
$$

(2) Level-2 iteration

Step 2.1 Initialization: Set $l = 0$ and

$$
\rho_1^{(l)} = 0, \qquad \lambda_1^{(l)} = T_{in}.
$$

Step 2.2 Update $\rho_k^{(l)}$ and $\lambda_k^{(l)}$: For patient 1,

$$\begin{aligned}\lambda_1^{(l+1)} &= T_{in}(1+\rho_1^{(l)}\Sigma_{i=2}^{m}\rho_i^{(l)}),\\ \rho_1^{(l+1)} &= \frac{\lambda_1^{(l+1)}}{\lambda_1^{(l+1)}+T_{normal}}.\end{aligned} \tag{8}$$

For patient $k = 2, \ldots, m-1$,

$$\begin{aligned}\lambda_k^{(l+1)} &= T_{in}(1+\rho_k^{(l)}(\Sigma_{i=1}^{k-1}\rho_i^{(l+1)}+\Sigma_{i=k+1}^{m}\rho_i^{(l)})),\\ \rho_k^{(l+1)} &= \frac{\lambda_k^{(l+1)}}{\lambda_k^{(l+1)}+T_{normal}}.\end{aligned} \tag{9}$$

For patient m,

$$\begin{aligned}\lambda_m^{(l+1)} &= T_{in}(1+\rho_k^{(l)}\Sigma_{i=1}^{m-1}\rho_i^{(l+1)}),\\ \rho_m^{(l+1)} &= \frac{\lambda_k^{(l+1)}}{\lambda_m^{(l+1)}+T_{normal}}.\end{aligned} \tag{10}$$

Step 2.3 Iteration: Set $l = l + 1$. If the terminating criteria is not met, go back to Step 2.2.

$$|\lambda_i^{(l+1)} - \lambda_i^{(l)}| \le \delta, \qquad |\rho_i^{(l+1)} - \rho_i^{(l)}| \le \delta, \quad i = 1, \ldots, m.$$

Step 2.4 Termination: If the terminating condition is met, set

$$\lambda_i^{(l+1)} = \lambda_i, \quad \rho_i^{(l+1)} = \rho_i, \quad \lambda_i = T_{final}, \quad i = 1, \ldots, m.$$

Appendix 2: Proofs

To prove Proposition 1, Lemmas 1 and 2 are needed.

Lemma 1 *Under assumptions* (1)–(5), *when* $m = 2$, *if* $p_{2,r}^{(j)} > p_{2,r}^{(j-1)}$, $r \in X$, $j = 1, 2, \ldots$, *then* $\tau_{1,r}^{(j+1)} > \tau_{1,r}^{(j)}$, $p_{1,r}^{(j+1)} < p_{1,r}^{(j)}$, $\tau_{2,r}^{(j+1)} < \tau_{2,r}^{(j)}$, $p_{2,r}^{(j+1)} > p_{2,r}^{(j)}$.

Lemma 2 *Under assumptions* (1)–(5), *when* $m = 2$, *the sequences* $p_{1,r}^{(j)}$ *and* $\tau_{2,r}^{(j)}$ *are monotonically decreasing, while the sequences* $p_{2,r}^{(j)}$ *and* $\tau_{1,r}^{(j)}$ *are monotonically increasing.*

Proof of Lemma 1 From all the equations related to the update of $\tau_{i,r}^{(j)}$ and $p_{i,r}^{(j)}$, which are from (4) to (6), define $C_{1,r}$ and $C_{2,r}$ as constants related to resource r, $r \in X$. We have

$$C_{1,r} = \begin{cases} p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}, & \text{if } r = int, \\ p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res} & \text{if } r = res, \\ p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} \\ +p_{rrt\&res}\tau_{rrt\&res} \\ +p_{rrt\&fel}\tau_{rrt\&fel} + p_{rrt\&atn}\tau_{rrt\&atn} & \text{if } r = rrt, \\ p_{fel}\tau_{fel} + p_{rrt\&fel}\tau_{rrt\&fel} & \text{if } r = fel, \\ p_{atn}\tau_{atn} + p_{rrt\&atn}\tau_{rrt\&atn} & \text{if } r = atn. \end{cases}$$

$$C_{2,r} = \begin{cases} p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}, & \text{if } r = int, \\ p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res} & \text{if } r = res, \\ p_{rrt}^2\tau_{rrt} + p_{rrt\&int}^2\tau_{rrt\&int}^2 \\ +p_{rrt\&res}^2\tau_{rrt\&res} \\ +p_{rrt\&fel}^2\tau_{rrt\&fel} + p_{rrt\&atn}^2\tau_{rrt\&atn} & \text{if } r = rrt, \\ p_{fel}^2\tau_{fel} + p_{rrt\&fel}^2\tau_{rrt\&fel} & \text{if } r = fel, \\ p_{atn}^2\tau_{atn} + p_{rrt\&atn}^2\tau_{rrt\&atn} & \text{if } r = atn. \end{cases}$$

For iteration j, if $p_{2,r}^{(j)} > p_{2,r}^{(j-1)}$, then for patient 1:

$$\tau_{1,r}^{(j)} = T_d + p_{2,r}^{(j-1)}C_{1,r} < T_d + p_{2,r}^{(j)}C_{1,r} = \tau_{1,r}^{(j+1)}, \tag{11}$$

$$p_{1,r}^{(j)} = \frac{C_{1,r}}{\tau_{1,r}^{(j)}} > \frac{C_{1,r}}{\tau_{1,r}^{(j+1)}} = p_{1,r}^{(j+1)}. \tag{12}$$

This leads to, for patient 2,

$$\tau_{2,r}^{(j)} = T_d + p_{1,r}^{(j)} C_{1,r} > T_d + p_{1,r}^{(j+1)} C_{1,r} = \tau_{2,r}^{(j+1)}, \tag{13}$$

$$p_{2,r}^{(j)} = \frac{C_{2,r}}{\tau_{2,r}^{(j)}} < \frac{C_{2,r}}{\tau_{2,r}^{(j+1)}} = p_{2,r}^{(j+1)}. \tag{14}$$

The obtained results in the above four inequations complete the proof. □

Proof of Lemma 2 Induction is used for the proof of the lemma.

Initial Step: When $j = 1$, since $p_{2,r}^{(0)} = 0$, from Eq. (14), we have

$$p_{2,r}^{(1)} > p_{2,r}^{(0)} = 0.$$

Then, from Lemma 1, we obtain

$$\tau_{1,r}^{(2)} > \tau_{1,r}^{(1)}, \quad p_{1,r}^{(2)} < p_{1,r}^{(1)}, \quad \tau_{2,r}^{2)} < \tau_{2,r}^{(1)}, \quad p_{2,r}^{(2)} > p_{2,r}^{(1)}.$$

The base case is proved.

Inductive Step: Assume when $j = k$, we have

$$\tau_{1,r}^{(k+1)} > \tau_{1,r}^{(k)}, \quad p_{1,r}^{(k+1)} < p_{1,r}^{(k)}, \quad \tau_{2,r}^{(k+1)} < \tau_{2,r}^{(k)}, \quad p_{2,r}^{(k+1)} > p_{2,r}^{(k)}.$$

From Lemma 1, this leads to

$$\tau_{1,r}^{(k+2)} > \tau_{1,r}^{(k+1)}, \quad p_{1,r}^{(k+2)} < p_{1,r}^{(k+1)}, \quad \tau_{2,r}^{(k+2)} < \tau_{2,r}^{(k+1)}, \quad p_{2,r}^{(k+2)} > p_{2,r}^{(k+1)}.$$

Thus, the case of $j = k + 1$ also holds.

By induction, we obtain that, when $m = 2$, the sequences $p_{1,r}^{(j)}$ and $\tau_{2,r}^{(j)}$ are monotonically decreasing, while the sequences $p_{2,int}^{(j)}$ and $\tau_{1,int}^{(j)}$ are monotonically increasing, $r \in X$, $j = 1, 2, \ldots$. □

Proof of Proposition 1 From Lemma 2, we obtain the monotonicity of decreasing sequences $p_{1,r}^{(j)}$ and $\tau_{2,r}^{(j)}$ and increasing sequences $p_{2,int}^{(j)}$ and $\tau_{1,int}^{(j)}$, $r \in X$, $j = 1, 2, \ldots$. Next we show that the sequences $\tau_{i,r}^{(j)}$ and $p_{i,r}^{(j)}$, are bounded from above and below. For $p_{i,r}^{(j)}$s, from Eqs. (12) and (14), we have

$$0 < p_{i,r}^{(j)} < 1.$$

For $\tau_{i,r}^{(j)}$s, from Eqs. (11) and (13), since $0 < p_{i,r}^{(j)} < 1$, we obtain

$$T_d < \tau_{i,r}^{(j)} < T_d + C_{i,r}.$$

Since the sequences $\tau_{i,r}^{(j)}$ and $p_{i,r}^{(j)}$, $r \in X$; $j = 1, 2, \ldots$, are monotonic and bounded from above and below, they are convergent. Thus, Level-1 iteration is convergent. □

To prove Proposition 2, Lemma 3 is needed.

Lemma 3 *Under assumptions* (1)–(5), *if* $\rho_i^{(l)} > \rho_i^{(l-1)}$, $i = 1, \ldots, m$, $l = 1, 2, \ldots$, *then* $\rho_i^{(l+1)} > \rho_i^{(l)}$.

Proof of Lemma 3 From Eq. (8), we obtain

$$\lambda_1^{(l+1)} = T_{in}(1 + \rho_1^{(l)} \Sigma_{i=2}^{m} \rho_i^{(l)}) > T_{in}(1 + \rho_1^{(l-1)} \Sigma_{i=2}^{m} \rho_i^{(l-1)}) = \lambda_1^{(l)}.$$

This implies that

$$\rho_1^{(l+1)} = \frac{\lambda_1^{(l+1)}}{\lambda_1^{(l+1)} + T_{normal}} = \frac{1}{1 + \frac{T_{normal}}{\lambda_1^{(l+1)}}} > \frac{1}{1 + \frac{T_{normal}}{\lambda_1^{(l)}}} = \rho_1^{(l)}.$$

When $2 \le k \le m - 1$, from (9), we have

$$\begin{aligned}
\lambda_k^{(l+1)} &= T_{in}(1 + \rho_k^{(l)} \left(\Sigma_{i=1}^{k-1} \rho_i^{(l+1)} + \Sigma_{i=k+1}^{m} \rho_i^{(l)} \right) \\
&> T_{in}(1 + \rho_1^{(l-1)} \left(\Sigma_{i=1}^{k-1} \rho_i^{(l)} + \Sigma_{i=k+1}^{m} \rho_i^{(l-1)} \right) \\
&= \lambda_k^{(l)}, \\
\rho_k^{(l+1)} &= \frac{\lambda_k^{(l+1)}}{\lambda_k^{(l+1)} + T_{normal}} > \frac{1}{1 + \frac{T_{normal}}{\lambda_k^{(l)}}} = \rho_k^{(l)}.
\end{aligned}$$

Finally, for $k = m$, from (10), it follows that

$$\begin{aligned}
\lambda_m^{(l+1)} &= T_{in} \left(1 + \rho_m^{(l)} \Sigma_{i=1}^{m-1} \rho_i^{(l+1)} \right) > T_{in} \left(1 + \rho_m^{(l-1)} \Sigma_{i=1}^{m-1} \rho_i^{(l)} \right) = \lambda_m^{(l)}, \\
\rho_m^{(l+1)} &= \frac{\lambda_m^{(l+1)}}{\lambda_m^{(l+1)} + T_{normal}} > \frac{1}{1 + \frac{T_{normal}}{\lambda_m^{(l-1)}}} = \rho_m^{(l)}.
\end{aligned}$$

The arguments follow directly. □

Proof of Proposition 2 First we prove that the sequences $\lambda_i^{(l)}$ and $\rho_i^{(l)}$, $i = 1, 2, \ldots, m$; $l = 1, 2, \ldots$, are monotonically increasing using mathematical induction.

Initial Step: When $l = 1$, since $\rho_i^{(0)} = 0$, from Lemma 3,

$$\rho_i^{(1)} > \rho_i^{(0)} = 0.$$

This leads to

$$\rho_i^{(2)} > \rho_i^{(1)}, \qquad \lambda_i^{(2)} > \lambda_i^{(1)}.$$

The base case is proved.

Inductive Step: Assume when $l = k$, we have

$$\lambda_i^{(k)} > \lambda_i^{(k-1)}, \qquad \rho_i^{(k)} > \rho_i^{(k-1)}, \qquad i = 1, 2, \ldots, m.$$

Then from Lemma 3, we have

$$\lambda_i^{(k+1)} > \lambda_i^{(k)}, \qquad \rho_i^{(k+1)} > \rho_i^{(k)}.$$

Therefore, the case where $l = k + 1$ also holds. Then, the sequences $\lambda_i^{(l)}$ and $\rho_i^{(l)}$, $i = 1, 2, \ldots, m$; $l = 1, 2, \ldots$, are monotonically increasing.

For boundedness, it is clear that $\rho_i^{(l)}$s are bounded between 0 and 1 from Eqs. (8), (9), and (10), while $\lambda_i^{(l)}$s are also bounded according to equations (8) and (9).

Since the sequences $\lambda_i^{(l)}$ and $\rho_i^{(l)}$, $i = 1, 2, \ldots, m$, are both monotonic and bounded from above and below, they are convergent. □

References

Berwick DM, Calkins DR, McCannon CJ, Hackbarth AD (2006) The 100,000 lives campaign: setting a goal and a deadline for improving health care quality. J Am Med Assoc 295(3):324–327

Brandeau ML, Sainfort F, Pierskalla WP (2004) Operations research and health care: a handbook of methods and applications. Springer, Berlin

Brindley PG (2010) Patient safety and acute care medicine: lessons for the future, insights from the past. Crit Care 14(2):217–221

Buchman TG, Coopersmith CM, Meissen HW, Grabenkort WR, Bakshi V, Hiddleson CA, Gregg SR (2017) Innovative interdisciplinary strategies to address the intensivist shortage. Crit Care Med 45(2):298–304

Chan PS, Jain R, Nallmothu BK, Berg RA, Sasson C (2010) Rapid response teams: a systematic review and meta-analysis. Arch Intern Med 170(1):18–26

Dacey MJ, Mirza ER, Wilcox V, Doherty M, Mello J, Boyer A, Gates J, Brothers T, Baute R (2007) The effect of a rapid response team on major clinical outcome measures in a community hospital. Crit Care Med 35(9):2076–2082

DeVita MA, Bellomo R, Hillman K, Kellum J, Rotondi A, Teres D, Auerbach A, Chen W-J, Duncan K, Kenward G (2006) Findings of the first consensus conference on medical emergency teams. Crit Care Med 34(9):2463–2478

DeVita MA, Hillman K, Bellomo R (2011) Textbook of rapid response systems: concept and implementation. Springer, Berlin

Downey A, Quach J, Haase M, Haase-Fielitz A, Jones D, Bellomo R (2008) Characteristics and outcomes of patients receiving a medical emergency team review for acute change in conscious state or arrhythmias. Crit Care Med 36(2):477–481

Fomundam S, Herrmann J (2007) A survey of queuing theory applications in health care. Technicial report no. 2007-24, the Institute for Systems Research, University of Maryland, College Park, MA

Garg L, McClean S, Meenan B, Millard P (2010) A non-homogeneous discrete time Markov model for admission scheduling and resource planning in a cost or capacity constrained healthcare system. Health Care Manag Sci 13(2):155–169

Green L (2006) Queueing analysis in healthcare. In: Hall RW (ed) Patient flow: reducing delays in healthcare delivery. Springer, Berlin, pp 281–307

Gunal MM, Pidd M (2010) Discrete event simulation for performance modelling in health care: a review of the literature. J Simul 4(1):42–51

Hall RW (2006) Patient flow: reducing delays in healthcare delivery. Springer, Berlin

Hillman K, Bristow P, Chey T, Daffurn K, Jacques T, Norman S, Bishop GF, Simmons G (2001) Antecedents to hospital deaths. Inter Med J 31(6):343–348

Hillman K, Chen J, Cretikos M, Bellomo R, Brown D, Doig G, Finfer S, Flabouris A (2005) Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. Lancet 365(9477):2091–2097

Jacobson SH, Hall SN, Swisher JR (2006) Discrete-event simulation of health care systems. Patient Flow Reducing Delay Healthc Deliv 91:211–252

Kohn LT, Corrigan JM, Donaldson MS (2000) To err is human: building a safer health system. Institute of Medicine, National Academy Press, Washington

Lakshmi C, Iyer SA (2013) Application of queueing theory in health care: a literature review. Oper Res Health Care 2(1):25–39

Leape LL, Berwick DM (2005) Five years after to err is human: What have we learned? J Am Med Assoc 293:2384–2390

Li J, Meerkov SM (2005) On the coefficients of variation of up- and downtime of manufacturing equipment. Math Probl Eng 2005:1–6

Massey D, Aitken LM, Chaboyer W (2010) Literature review: Do rapid response systems reduce the incidence of major adverse events in the deteriorating ward patient? J Clin Nurs 19(23–24):3260–3273

Mayhew L, Smith D (2008) Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments. Health Care Manag Sci 11(1):11–21

Meyers MO, Sarosi GA, Brasel KJ (2017) Perspective of residency program directors on accreditation council for graduate medical education changes in resident work environment and duty hours. JAMA Surg 152(10):905–906

McArthur-Rouse F (2001) Critical care outreach services and early warning scoring systems: a review of the literature. J Adv Nurs 36(5):696–704

McGloin H, Adam SK, Singer M (1999) Unexpected deaths and referrals to intensive care of patients on general wards. Are some cases potentially avoidable? J R Coll Phys Lond 33(3):255–259

Priestley G, Watson W, Rashidian A, Mozley C, Russell D, Wilson J, Cope J, Hart D, Kay D, Cowley K, Pateraki J (2004) Introducing critical care outreach: a ward randmized trial of phased introduction in a general hospital. Intensive Care Med 30(7):1398–1404

Ranji S, Auerbach A, Hurd C, O'Rourke K, Shohania K (2007) Effects of rapid response systems on clinical outcomes: systematic review and meta analysis. J Hosp Med 2(6):422–432

Schaefer AJ, Bailey MD, Shechter SM, Roberts MS (2005) Modeling medical treatment using Markov decision processes. In: Brandeau ML et al (eds) Operations research and health care. Springer, Berlin, pp 593–612

Wang J, Quan S, Li J, Hollis A (2012) Modeling and analysis of work flow and staffing level in a computed tomography division of University of Wisconsin Medical Foundation. Health Care Manag Sci 15(2):108–120

Wang J, Li J, Howard PK (2013) A system model of work flow in the patient room of hospital emergency department. Health Care Manag Sci 16(4):341–351
Wang J, Zhong X, Li J, Howard PK (2014) Modeling and analysis of care delivery services within patient rooms: a system-theoretic approach. IEEE Trans AutomSci Eng 11(2):379–393
Watcher RM (2004) The end of the beginning: patient safety five years after "To err is human". Health Aff W4:534–545
Whitlock J (2017) Doctors, residents, interns, and attendings: What's the difference? The doctors on your healthcare team. https://www.verywell.com/types-of-doctors-residents-interns-and-fellows-3157293 Accessed Jan 2018
Wiler JL, Griffey RT, Olsen T (2011) Review of modeling approaches for emergency department patient flow and crowding research. Acad Emerg Med 18(12):1371–1379
Winters BD, Pham JC, Hunt E, Guallar EA, Berenholtz S, Pronovost PJ (2007) Rapid response systems: a systematic review. Crit Care Med 35(5):1238–1243
Xie X, Li J, Swartz CH, Depriest P (2012) Modeling and analysis of rapid response process to improve patient safety. IEEE Trans Autom Sci Eng 9(2):215–225
Xie X, Li J, Swartz CH, Depriest P (2014) Improving response-time performance in acute care delivery: a systems approach. IEEE Trans Autom Sci Eng 11(4):1240–1249
Xie X, Li J, Swartz C, Dong Y, DePriest P (2016) Modeling and analysis of ward patient rescue process on the hospital floor. IEEE Trans Autom Sci Eng 13(2):514–528
Zhong X, Williams M, Li J, Kraft S, Sleeth J (2015) Primary care redesign: review and a simulation study at a pediatric clinic. In: Yang H, Lee E (eds) Healthcare data analytics, Wiley series on operations research and management science (WORMS). Wiley, Hoboken, pp 399–426

Zexian Zeng received both his master degrees in Industrial and Systems Engineering and Computer Science from University of Wisconsin–Madison in 2013. He received his Ph.D. in Biomedical Informatics from Northwestern University in 2018. He is currently working as post-doc fellow in Dana-Farber Cancer Institute in Harvard University. His research interests include stochastic modeling, machine learning, natural language processing, computational genomics, with a focus on healthcare and biomedical applications.

Zhenghao Fan received his bachelor's degree from the Department of Industrial Engineering at Tsinghua University in 2017 and is currently working towards his Ph.D. degree in the same department. His research interests are in stochastic process, simulation and healthcare analytics. He is the member of Institute of Industrial and Systems Engineers (IISE) and the Institute of Electrical and Electronics Engineers (IEEE).

Xiaolei Xie is an associate professor with the Department of Industrial Engineering at Tsinghua University. He obtained his Ph.D. from Department of Industrial and Systems Engineering at University of Wisconsin, Madison, in 2014. His research interests are healthcare operations management, healthcare data analytics and productions systems engineering. He is a member of the Institute for Operations Research and the Management Sciences (INFORMS) and the Institute of Electrical and Electronics Engineers (IEEE).

Colleen H. Swartz DNP, MSN, MBA, RN, NEA-BC, FNAP, holds a DNP degree, a master's degree in nursing as a Clinical Nurse Specialist in Trauma/Critical Care, as well as an MBA. She has completed the Johnson & Johnson Wharton Fellows Program in Management for Nurse Executives and is a Robert Wood Johnson Foundation Executive Nurse Fellow Alumna, 2011 Cohort. She became chief nurse executive for UK HealthCare in December 2008 and was appointed chief administrative officer in February of 2017. In January 2019, she was appointed VP for Hospital Operations. Her prior experience includes serving as chief nursing officer at a regional community hospital, director of emergency and trauma services, flight nursing and as director of the Capacity Command Center for UK HealthCare.

Paul DePriest MD, MHCM serves as Executive Vice President and Chief Operating Officer of Baptist Memorial Health Care in Memphis Tennessee. He is Board Certified in Obstetrics and Gynecology with sub-specialty board certification in Gynecologic Oncology. He received his medical degree at the University of Kentucky College of Medicine, where he also served his residency and fellowship training. He received a Master of Science Degree in Healthcare Management from the Harvard School of Public Health.

Jingshan Li received the B.S. degree from Tsinghua University, Beijing, China, the M.S. degree from Chinese Academy of Sciences, Beijing, and the Ph.D. degree from University of Michigan, Ann Arbor, in 1989, 1992, and 2000, respectively. He was a Staff Research Engineer at General Motors Research and Development Center from 2000 to 2006, and was with University of Kentucky from 2006 to 2010. He is now a Professor in Department of Industrial and Systems Engineering, University of Wisconsin–Madison. His primary research interests are in modeling, analysis and control of manufacturing and healthcare systems. He is an IEEE Fellow and an IEEE Distinguished Lecturer in robotics and automation. He received 2010 NSF Career Award, 2009 IIE Transactions Best Application Paper Award, 2005 IEEE Transactions on Automation Science and Engineering Best Paper Award, 2006 IEEE Early Industry/Government Career Award in Robotics and Automation, and multiple awards in flagship international conferences. He is a Senior Editor of IEEE Transactions on Automation Science and Engineering and IEEE Robotics and Automation Letters, Department Editor of IIE Transactions, Area Editor of Flexible Service and Manufacturing Journal, and Associate Editor of International Journal of Production Research and International Journal of Automation Technology. He is the Program Chair of 2019 IEEE International Conference on Automation Science and Engineering and was General and Program Co-Chair in 2013 and 2015. He was the founding Chair of IEEE Technical Committee on Sustainable Production Automation (2012–2016) and has been the Chair of the Technical Committee on Automation for Healthcare Management since 2016.

Affiliations

Zexian Zeng[1] · Zhenghao Fan[2] · Xiaolei Xie[2] · Colleen H. Swartz[3] · Paul DePriest[4] · Jingshan Li[5]

Zexian Zeng
zexian.zeng@northwestern.edu

Zhenghao Fan
fanzh17@mails.tsinghua.edu.cn

Colleen H. Swartz
chswar2@uky.edu

Paul DePriest
paul.dePriest@bmhcc.org

Jingshan Li
jingshan.li@wisc.edu

1 Feinberg School of Medicine, Northwestern University, Evanston, IL, USA

2 Department of Industrial Engineering, Tsinghua University, Beijing 100084, People's Republic of China

3 University of Kentucky Chandler Medical Center, Lexington, KY 40506, USA

4 Baptist Memorial Health Care Corporation, Memphis, TN 38120, USA

5 Department of Industrial and Systems Engineering, University of Wisconsin, Madison, WI 53706, USA