Health Care Manag Sci (2018) 21:554–572
DOI 10.1007/s10729-017-9412-8

Overflow models for the admission of intensive care patients

Yin-Chi Chan[1] · **Eric W. M. Wong**[1] · **Gavin Joynt**[2] · **Paul Lai**[3] · **Moshe Zukerman**[1]

Received: 19 December 2016 / Accepted: 11 July 2017 / Published online: 28 July 2017

Abstract An earlier article, inspired by overflow models in telecommunication systems with multiple streams of telephone calls, proposed a new analytical model for a network of intensive care units (ICUs), and a new patient referral policy for such networks to reduce the blocking probability of external emergency patients without degrading the quality of service (QoS) of canceled elective operations, due to the more efficient use of ICU capacity overall. In this work, we use additional concepts and insights from traditional teletraffic theory, including resource sharing, trunk reservation, and mutual overflow, to design a new patient referral policy to further improve ICU network efficiency. Numerical results based on the analytical model demonstrate that our proposed policy can achieve a higher acceptance level than the original policy with a smaller number of beds, resulting in improved service for all patients. In particular, our proposed policy can always achieve much lower blocking probabilities for external emergency patients while still providing sufficient service for internal emergency and elective patients. In addition, we provide new accurate and computationally efficient analytical approximations for QoS evaluation of ICU networks using our proposed policy. We demonstrate numerically that our new approximation method yields more accurate, robust and conservative results overall than the traditional approximation. Finally, we demonstrate how our proposed approximation method can be applied to solve resource planning and optimization problems for ICU networks in a scalable and computationally efficient manner.

Keywords Intensive care units · Overflow networks · Analytical approximation · Optimization

✉ Yin-Chi Chan
ycchan26@cityu.edu.hk

Eric W. M. Wong
eewong@cityu.edu.hk

Gavin Joynt
gavinmjoynt@cuhk.edu.hk

Paul Lai
paullai@cuhk.edu.hk

Moshe Zukerman
m.zu@cityu.edu.hk

[1] Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Ave., Kowloon Tong, Hong Kong

[2] Department of Anesthesia and Intensive Care, Chinese University of Hong Kong, Prince of Wales Hospital, Sha Tin, Hong Kong

[3] Department of Surgery, Chinese University of Hong Kong, Prince of Wales Hospital, Sha Tin, Hong Kong

1 Introduction

The intensive care unit (ICU) is a crucial and expensive resource, with an ICU bed costing up to six times the cost of a regular hospital bed [18]. As a result, ICUs are frequently under-resourced and over-utilized, with occupancy rates of over 90 percent reported in the literature [27]. In addition, congestion in the ICU has a knock-on effect on the rest of the hospital system, for example in the form of deferred elective operations, and can lead to increased rates of death or ICU readmission due to early discharge from the ICU [10, 53]. Therefore, much research has gone into the efficient management of such units [1].

In particular, various studies in the literature have found that queueing theory is a useful tool for the modeling and resource planning of intensive care resources [19, 44]. In this paper, we apply and extend various concepts from queueing and teletraffic theory to develop a new policy for improving the admission of patients to an ICU network, so that the blocking probability of external emergency patients is reduced, while still fulfilling QoS requirements for other patient types. We also develop new analytical approximation tools for the performance evaluation and resource planning of ICU networks using the proposed policy.

1.1 Classification of ICU patients

In this paper, we consider an analytical model for a network of ICUs in which patients are classified into three types [41]. External emergency patients are those arriving from ambulatory care (including air transport). Internal emergency patients arrive to an ICU from other departments at the same hospital. Finally, elective patients correspond to planned operations requiring post-operative ICU stay.

Due to legal, logistical, and economic concerns regarding patient admission and transfer, efficient resource planning and daily operation of ICU networks, subject to meeting all quality-of-service (QoS) constraints, can form a major challenge. In almost all cases, internal emergency and elective patients must be served at the ICU of the hospital from which they originate. For example, in the United Kingdom, government policy is that patient transfers for non-clinical reasons "should only take place in exceptional circumstances and ideally only during daylight hours" [52]. Studies in the UK and Victoria, Australia have shown that critical care patients having undergone at least one non-clinical transfer remain in critical care for a longer duration on average [2, 12]. In other countries, such as the Netherlands, non-clinical transfers are banned outright and "a patient can only be transferred if it is beneficial for the patient" [41].

Furthermore, for economic reasons, ICUs may reject external emergency patients in favor of elective patients awaiting planned operations. For example, it has been reported in California that some hospitals prefer to divert ambulances in favor of their elective patients, in the interest of improving payer mix and revenue collection [29]. In another example, Litvak et al. [41] studied a group of hospitals in the Netherlands and found that the designation of one of the hospitals in the region as a trauma center was causing capacity problems at that hospital's ICU, as the other hospitals were increasingly referring external emergency patients to the designated trauma center in favor of their own elective patients.

On account of the different requirements for different ICU patient types, and of the inequities in patient outcomes caused by current policies, any new policy for the admission of patients to an ICU network must take fairness into account in addition to the overall patient rejection rate. This may require restricting ICU services to patients most likely to benefit [9]. In fact, maximizing the number of beds available to each patient without reservation does not even necessarily lead to the lowest *overall* patient rejection rate, as demonstrated in Section 3.4.3.

1.2 Cooperation between multiple medical units in a region

One method currently in use for improving the QoS of hospital patients is the pooling of resources from multiple hospitals within a region [41, 48]. For example, public hospitals in Hong Kong are grouped into seven clusters, and the New Territories East cluster (NTEC) has three ICUs with approximately 20, 15, and 8 beds, respectively. The NTEC is frequently full and chronically over-utilized, and ICU patients in the region may be forced to transfer (although this option may not always be available). In response to chronic patient refusal, supported by data published by one of these ICUs [27], the Hong Kong Hospital Authority has introduced an ICU transfer policy to improve resource use.

Another example of inter-hospital cooperation is given by McManus et al. [44], who describe an ICU in the United States where "external requests for transfer" of a patient to an ICU may be "diverted to other institutions in the region" during times of congestion. This corresponds to the concept of *overflow* for external emergency patients in the Litvak et al. [41] model. Additionally, overflow of internal emergency patients is "accommodated in off-service care sites" such as a post-anesthesia care unit or a separate, specialized cardiac ICU.

Despite evidence that cooperation between multiple ICUs in a region can reduce the rejection rate of intensive care patients, many hospital regions currently have no strategy for doing so. Instead, most ICUs currently seek transfer of overflow patients on an ad-hoc basis, without centralized systems or systematic polices (such as the one proposed in this paper) to coordinate capacity and utilization issues. Therefore, the practical behavior of ICUs can and should change for the better. The current lack of centralized systems for ICU coordination may be because places sophisticated enough to introduce such systems are generally well-resourced. However, shortages are predicted in the future due to increasing costs and aging populations, even in well-resourced countries. Systems like this are thus likely to be needed.

Furthermore, cooperation between ICUs to improve resource use may prove vital during periods of unexpected

increases in ICU demand, such as an outbreak of acute infections. For example, there were 1755 cases of severe acute respiratory syndrome in Hong Kong between February and May 2003 [58], with over 20 percent of patients admitted to the ICU at one hospital at one point [40]. The European Society of Intensive Care Medicine recommends that Emergency Executive Control Groups be set up at regional or even national levels to exercise authority over resource use and communications during such outbreaks or other mass disasters [28].

1.3 Improving patient admissions in ICU networks

Litvak et al. [41] studied a network of four Dutch ICUs, with the aim of minimizing the blocking of external emergency patients (defined as the probability that such a patient is refused from all four ICUs due to lack of capacity), subject to maintaining a minimum QoS for elective patients. Inspired by overflow models in telecommunication systems, Litvak et al. [41] designed an analytical model of an ICU network and proposed a policy in which ICUs in a region jointly reserve beds for the admission of external emergency patients only. In practice, these beds will be distributed over the ICUs in the region, thus creating a virtual ICU. This implies that in certain cases hospitals must cancel elective operations despite having an empty operational bed.

Litvak et al. [41] claim that their policy results in improved service for *all* patients, despite reducing the number of beds available to internal emergency and elective patients. This is due to the more efficient use of ICU capacity overall. Nevertheless, the proposed reservation scheme still leads to wastage of valuable ICU resources (i.e. beds) in some cases. For example, an elective patient due to undergo operation at a given hospital may be deferred even if an ICU bed is empty at the same hospital, due to that bed being reserved for the virtual ICU. At the same time, an external emergency patient may be rejected from a regional ICU network despite a bed being available somewhere in the network, due to that bed not forming part of the virtual ICU.

In this paper, we use additional concepts and insights from traditional teletraffic theory, including resource sharing [30, 36], trunk reservation [23, 31, 51], and mutual overflow [22, 37], to design a new patient referral policy for the Litvak et al. [41] model in which no ICU bed is explicitly reserved for a particular patient type or set of patient types. In this way, we increase the flexibility for assigning ICU beds for all types of patients. We demonstrate numerically that our proposed model has better resource sharing than the virtual ICU model, in the sense of reducing the blocking probability of external emergency patients given a fixed QoS requirement for internal emergency and elective patients.

1.4 Analytical approximation methods for the performance evaluation and resource planning of ICU networks

To illustrate the need for approximate QoS evaluation, we note that the number of system states in an ICU network, under both the virtual ICU model and our proposed model, is exponential in the number of ICUs in the network. As an example, when solving a similar resource allocation problem for a burn care network, Blair and Lawrence [5] were only able to optimize a network of four wards and were forced to split their seven-ward network into two fully independent parts.

This paper combines two existing analytical approximation methods in the literature, namely exponential decomposition (ED) [16], also known as the Erlang fixed-point approximation [15, 31, 32], and the Information Exchange Surrogate Approximation (IESA) [7, 8, 56, 57, 59], and extends both methods to apply to the ICU network model. Such extensions are necessary due to special properties of the ICU network model that do not exist in other types of systems.

To illustrate the usefulness of our proposed QoS approximation methods, we apply these methods to the following optimization problem: given a network of ICUs, each with a predetermined capacity, find the optimal reservation thresholds for each ICU so that the overall blocking probability of external emergency patients is minimized, subject to maintaining a minimum QoS for internal emergency and elective patients. We demonstrate that the accuracy and fast running time of our approximations, as compared to simulation, allows for efficient coverage of large search spaces. Numerical results show that our proposed threshold reservation policy, with the reservation thresholds optimized using our new QoS evaluation method, produces much lower blocking for external emergency patients than the virtual ICU policy of Litvak et al. [41] (with the number of virtual ICU beds optimized using their QoS evaluation method).

1.5 Organization

The remainder of this paper is organized as follows. Section 2 provides a brief background on relevant topics regarding ICU management and queueing theory. Section 3 describes the Litvak et al. [41] model of an ICU network and compares two policies for patient referral within such a network: one proposed by Litvak et al. [41] and one which we propose here. In Section 4, we show that the ICU network model is not very sensitive to the patient length-of-stay (LoS) distribution. This allows us to assume an exponential LoS distribution which greatly simplifies analysis. Sections 5 and 6 provide approximation methods for the QoS of an

ICU network under our proposed model. In Section 7, these approximation methods are incorporated into an optimization algorithm for minimizing the blocking probability of external emergency patients in an ICU network. Concluding remarks are made in Section 8.

2 Queueing theory and healthcare: a literature review

Queueing theory has long been used in the field of healthcare [39, 49] not only to analyze system performance, but also in order to facilitate system design. In particular, Lakshmi and Iyer [39] distinguish between mathematical challenges (health care problems for which appropriate queueing network models have not yet been developed) and health care challenges (health care problems which have not been studied yet, but could be studied using *existing* queueing techniques). In this section, we review several concepts from queueing theory which motivate both the ICU network model which we seek to optimize and the patient referral policy that we propose for this model.

2.1 Modeling patient flows

Newell [45] studied arrivals of emergency cases to a teaching hospital in England between the years of 1950 to 1952, and found the daily tallies could be modeled using a Poisson distribution, as long as Sunday and weekday arrivals were counted separately. Similar results were noted by Long and Feldstein [42], Kim et al. [34, 35], and Kim and Whitt [33]. In practice, many analytical models of patient arrivals to hospitals ignore daily or seasonal variations in the arrival rate, thus assuming a simple Poisson process [5, 14, 41, 45, 48]. We will do the same in our model.

Various distributions have been used to model the LoS distribution of hospital patients, including lognormal [20, 41, 43], hyperexponential [20], Weibull [50], and hypergamma [47]. Despite the large number of LoS distribution types used in the literature, it has been found that a lognormal distribution provides a satisfactory approximation of patient LoS [20, 43]. Furthermore, Litvak et al. [41] found that the QoS of their ICU network did not significantly depend on the patient LoS distribution apart from its mean, allowing a much simpler exponential LoS distribution to be used. We shall re-examine this claim in Section 4.

By assuming Poisson arrivals and exponential LoS, the state of an ICU network can be modeled as a continuous-time Markov chain [46], from which QoS measures can be (in theory) obtained via an exact analytical solution. On the other hand, since the number of states grows exponentially as the number of ICUs increases, exact analysis of the resulting state space is not a scalable analytical approach.

2.2 Resource sharing, closed chains, and mutual overflow

Resource sharing is motivated by the fact that pooled resources can be used more efficiently than dedicated resources: for example, a single queue for a group of servers (i.e. cashiers, bank tellers, ICU beds, etc.) results in shorter waiting times than a separate queue for each server. While resource sharing in hospital systems has been linked to economies of scale [4], Kleinrock [36] explains that resource sharing leads to gains beyond simple unit cost discounts for resource acquisition, management, and maintenance. Instead, these gains are related to the statistical nature of the demand. Simply put, the law of large numbers dictates that any statistical fluctuations in an individual's demand for a resource is smoothed out in the larger population, so that the total demand approaches deterministic as the population size increases [36, p. 275].

A closely related concept to resource sharing is that of *mutual overflow* [22, 37]. Mutual overflow arises when congestion in a queue causes overflow to other queues, which in turn become congested and yield overflow back to the original queue. One way of achieving mutual overflow is the *closed chain*, in which requests attempt each queue in the chain in cyclic order. For example, consider a set of G server groups (ICUs) labeled 1 to G. In a closed-chain configuration, requests of type n will attempt, in order, server groups $n_{(G)}, (n+1)_{(G)}, \ldots (n+k-1)_{(G)}$, where k represents the maximum number of server groups a request may attempt and $x_{(G)} = ((x-1) \bmod G) + 1$. Closed chains have been shown to improve QoS compared to systems without closed chains [17, 21, 26].

2.3 Resource allocation in systems with multiple patient types

Bekker et al. [3] listed four possible policies for allocating multiple types of patients to multiple wards in the same hospital:

- A *separate ward* for each patient type
- *Simple merging*: all wards serve all patient types
- *Earmarking*: each patient type has a number of dedicated beds, with the remaining beds shared among all patients (the previous two policies can be seen as the limiting cases of this policy)
- *Threshold policy*: All beds may serve all patient types, but each ward refuses certain patients if the number of vacant beds falls below a certain threshold. An example of threshold reservation in a healthcare context is given by Esogbue and Singh [14], who considered a single hospital ward with N beds serving emergency and elective patients, with the last $N - m$ beds reserved for

emergency patients (i.e elective patients are not admitted if more than m beds are occupied at the ward). An exact method of computing the blocking probabilities of both types of patients was presented and was used to optimize the reservation threshold m.

Bekker et al. [3] proposed that the threshold policy be used for small hospital systems, in which the policy was shown to be nearly optimal, whereas the earmarking policy is preferable for larger-scale systems in order to cut down on cross-training costs, as only a few shared beds (and therefore medical staff) are required to be shared among all patient types. On the other hand, in our ICU network model, there is only one type of bed which is distributed among *multiple* hospitals, and restrictions on overflow are based on geographical considerations instead.

2.4 Resource allocation and threshold reservation in a network of physically separate ICUs

In the proposed policy of Litvak et al. [41], each ICU reserves several beds exclusively for external emergency patients. which are then pooled together in a *virtual ICU*. This is similar to Bekker et al.'s [3] earmarking policy for a single hospital with multiple wards, but differs in that internal emergency and elective patients do not have access to the virtual ICU. The virtual ICU policy is inefficient as the strict reservation of certain beds for external emergency patients means that other types of patients may be rejected even when there is a large number of vacant beds available in the network.

To resolve this, we propose using a *threshold reservation* scheme as described in Section 2.3. The threshold policy is such that external emergency patients cannot use the last few remaining beds of each ICU. Such a policy bares similarities to trunk reservation in telecommunications networks [38], in which the last few circuits of each trunk are reserved for direct traffic in order to prevent overflow traffic (which use longer and more resource-intensive alternate routes) from dominating the network. Note that in trunk reservation, no individual circuit is explicitly reserved for direct traffic; likewise, in our threshold reservation policy, no individual bed is explicitly reserved for a particular patient type or set of patient types. Threshold reservation thus maximizes resource sharing in periods of non-congestion, while protecting the QoS of internal emergency and elective patients, which cannot overflow, during periods of occasional congestion.

2.5 QoS approximation in overflow loss systems

The ICU network model which we consider in this paper belongs to a broad class of stochastic models known as *overflow loss systems*. In an overflow loss system, there is a set of request types and a set of server groups, each server group serves some subset of the request types in the system, and a *routing policy* determines the order which requests of each type attempt the set of accessible server groups for that request type. In this paper, we will consider only routing policies where these orderings are fixed, as opposed to state-dependent or random.

The classical analytical approximation approach for performance evaluation in overflow loss systems, known by various names such as the reduced load approximation [54], Erlang fixed-point approximation [15, 31, 32], and exponential decomposition [16], has a long history in teletraffic theory; see Cooper and Katz [11] for an early example of its use. In this paper, we will use the term exponential decomposition (ED). ED decomposes the system into independent Erlang B subsystems [13] by adding two simplifying assumptions to the analytical model: (i) that the offered traffic to each subsystem, composed of both direct and overflow traffic, is Poisson, and (ii) that the offered traffic to each subsystem is independent of all other traffic streams. Due to these two simplifying assumptions, ED dramatically reduces the computing time compared to exact analysis of the full state space. However, as these two assumptions are generally not valid, they can also lead to large approximation errors in various scenarios [56, 57].

Several publications [11, 16, 24] have proposed *moment matching* for reducing errors caused by the Poisson assumption; once again, see Cooper and Katz [11] for an early example. This approach was used effectively by Litvak et al. [41] for performance evaluation of their ICU network model under the virtual ICU policy. In this paper, we use moment matching to provide conservative QoS estimates for the two patient types without overflow, namely internal emergency and elective patients.

On the other hand, moment matching provides only marginal improvement over traditional ED in systems involving mutual overflow, where the independence assumption forms the main source of error [56]. Therefore, ED with moment matching is not adequate for QoS evaluation of external emergency patients under out proposed patient referral policy. Other publications have proposed ways to reduce errors caused by the independence assumption. One approach is to apply the technique used in traditional ED, i.e. decomposing the systems into independent Erlang B subsystems, on a *surrogate* of the original system. Ideally, the QoS of the surrogate closely approximates that of the original system, but the surrogate possesses certain properties which greatly reduce its approximation error caused by decomposition. The estimated QoS of the surrogate is then used as a QoS estimate for the original system.

In this paper, we adapt and extend one such surrogate-based approximation framework, the Information Exchange

Surrogate Approximation (IESA) framework [7, 8, 57, 59], to our proposed ICU network model, where it is used to evaluate the QoS of external emergency patients. IESA features an information exchange mechanism in which incoming calls/requests may exchange certain congestion information with calls/requests in service. This mechanism can capture traffic dependence in the system, and hence it can provide significantly improve the accuracy over ED. Numerical results demonstrate that IESA is more accurate and robust than traditional ED for the QoS evaluation of external emergency patients, while remaining much more computationally efficient than exact analysis or simulation. In fact, IESA has a closed form solution, whereas ED does not, due to the hierarchical nature of the information exchange mechanism.

3 Model

We consider the three-patient-type model of Litvak et al. [41] with external emergency patients, internal emergency patients, and elective patients, as depicted in Fig. 1. In this model, there are G ICUs, each with its own catchment zone. Let Zone i denote the catchment zone for ICU i. External emergency patients to each ICU i arrive from Zone i according to a Poisson process with rate $\lambda_{i,1}$ and may be admitted to any ICU in the network. Such patients are *blocked* (transferred to another hospital network or demoted to a lower level of care) if and only if every bed in the entire ICU network is either occupied or reserved for other patient types.

Internal emergency patients and elective patients arrive directly at each ICU i in accordance to Poisson processes with rates $\lambda_{i,2}$ and $\lambda_{i,3}$, respectively, and are not allowed to overflow. An internal emergency patient which cannot be admitted to a regular ICU bed will trigger the creation of a temporary *overbed*; in a physical ICU network, this may be a bed in another hospital department such as a post-anesthesia care ward or a separate, specialized cardiac ICU [44]. Elective patients, on the other hand, generally correspond to non-time-critical surgical operations; if no ICU is available for such patients, the operation is *deferred*. For simplicity, we will not model retrials; instead, any subsequent attempt of an elective patient to obtain an ICU bed is modeled as a new arrival.

Let C_i denote the number of regular beds in ICU i, i.e. the rated capacity of that ICU. As in Litvak et al. [41], we assume that patient LoS is exponentially distributed with equal mean (except in Section 4, where we show via simulation that the QoS is not very insensitive to the shape of the LoS distribution apart from its mean). Without loss of generality, we assume this mean to be one.

As an analytical model, the Litvak et al. [41] model contains several simplifications compared to a physical ICU network, as listed in Table 1. Nevertheless, the model forms a good environment for testing new concepts and methodologies before they are applied to more complex real-world systems.

3.1 Notation for QoS evaluation

For measuring the QoS of an ICU network, let B_i denote the blocking probability of external emergency patients from catchment zone i, defined as the probability that such a patient is refused by all the ICUs in the network and thus rejected from the ICU network entirely. Let D_i denote the deferral probability of elective patients arriving at ICU i, defined as the probability that the planned operation of an elective patient is deferred due to a lack of beds at ICU i. Let T_i denote the mean number of overbeds at ICU i for

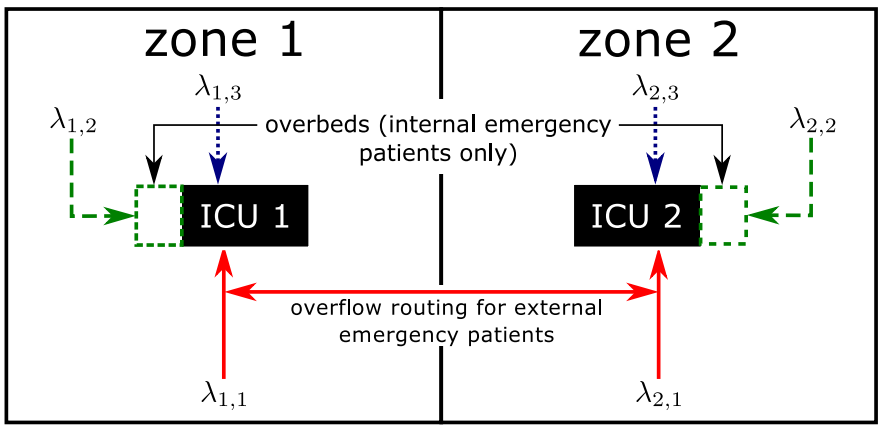

Fig. 1 An example ICU network with two ICUs. Solid arrows, dashed arrows, and dotted arrows represent external emergency, internal emergency, and elective patients, respectively

Table 1 Comparison of our ICU network model to a physical network

Physical network	Analytical model	Justification
Patients arrive to the ICU network from other hospital departments, including the AED and surgical units. The arrival process to the ICU network is unknown.	The ICU network is treated as an isolated system to which patients arrive directly, according to a Poisson processes with constant rate.	Poisson processes are well suited to modeling events that are rare from an individual point of view, but which occur within a large population. Isolating the ICUs from the rest of the hospital network simplifies analysis.
Patient LoS has an unknown distribution.	Patient LoS is modeled using an exponential distribution.	The sensitivity of the QoS to the LoS distribution is demonstrated in [41] and Section 4 to be low.
Certain ICUs may be better-equipped to deal with certain patients, based on the types of specialists required.	All ICU beds are considered identical. Any penalty incurred by serving an external emergency patient at a non-preferred ICU (in the form of transportation costs, decreased quality of care, etc.) is ignored.	Simplification of the analytical model.
Elective patients are deferred if no ICU bed is available and will reattempt their planned operation a later time.	Subsequent service attempts by elective patients are treated as new arrivals.	Simplification of the analytical model.
Internal emergency patients may be referred to another hospital unit, e.g. the post-anathesia care unit, if the ICU is full.	Internal emergency patients arriving at an ICU create temporary overbeds *within* the ICU itself when the ICU is full.	The patient may require increased resources compared to a regular patient despite being referred to an non-ICU unit. Additionally, it is expected that such patients will be transferred back to the ICU as soon as an ICU bed becomes available.

internal emergency patients. Let B, D, and T represent the overall blocking probability, deferral probability, and mean number of overbeds for the entire network; thus

$$B = \frac{\sum_{z=1}^{G} \lambda_{z,1} B_i}{\sum_{z=1}^{G} \lambda_{z,1}}$$

$$D = \frac{\sum_{i=1}^{G} \lambda_{i,3} D_i}{\sum_{i=1}^{G} \lambda_{i,3}},$$

and

$$T = \sum_{i=1}^{G} T_i.$$

Finally, let b_i denote the congestion probability of ICU i for external emergency patients, defined as the probability that an external emergency patient arriving at ICU i will be refused by that ICU. The notation defined above is summarized in Table 2.

3.2 Virtual ICU policy

The virtual ICU policy was introduced by Litvak et al. [41] as a more efficient policy than a set of G fully independent ICUs, demonstrating that through resource sharing, improvements in QoS could be obtained for all patient types. Under the virtual ICU policy, each ICU i, $i = 1, \ldots, G$, reserves r_i^{V} beds exclusively for external emergency patients. These reserved beds form a *virtual ICU* which only serves external emergency patients. An external emergency patient arriving from Zone i will first attempt to obtain one of the $C_i - r_i^{\mathrm{V}}$ unreserved beds at ICU i. If none of these beds are available, the patient will attempt to obtain a bed at the virtual ICU. If all virtual ICU beds are also occupied, then the patient is *blocked*. A graphical depiction of the virtual ICU model is shown in Fig. 2.

Litvak et al. [41] provided a moment-matched version of ED for QoS evaluation under the virtual ICU policy, and demonstrated that this method produces accurate QoS results for this policy. In general, moment-matched ED is effective for hierarchical overflow models [16]. On the other hand, the virtual ICU policy is sub-optimal in terms of maximizing QoS: the purely hierarchical structure of the virtual ICU model means that the level of resource sharing remains far from ideal.

3.3 Threshold reservation policy

Let $\Gamma_{z,n}$ denote the ICU to which external emergency patients from Zone z and with n previous service attempts

Table 2 Table of notations for the ICU network model

Symbol	Definition
G	Number of ICUs in the system
$\lambda_{i,1}$	Arrival rate of external emergency patients from catchment zone i
$\lambda_{i,2}$	Arrival rate of internal emergency patients to ICU i
$\lambda_{i,3}$	Arrival rate of elective patients to ICU i
C_i	Rated capacity of ICU i
B	Overall blocking probability of external emergency patients in the ICU network
B_i	Blocking probability of external emergency patients from catchment zone i
b_i	Probability that an external emergency patient attempting ICU i will be refused by that ICU
T	Mean number of temporary overbeds in the ICU network for internal emergency patients
T_i	Mean number of temporary overbeds for internal emergency patients in ICU i
D	Overall deferral probability of elective patients in the ICU network
D_i	Deferral probability of elective patients at ICU i

are referred. Under the threshold policy, external emergency patients arriving from Zone z will attempt each bed in $\Gamma_z = \left(\Gamma_{z,0}, \Gamma_{z,1}, \ldots \Gamma_{z,G-1}\right)$ in order until an available ICU bed is found. We call Γ_z the *overflow policy* of external emergency patients from zone z. However, unlike in the virtual ICU model, no beds are explicitly set aside for any patient type. Instead, we impose a set of thresholds, $r^{\mathrm{R}}_{i,1}$ and $r^{\mathrm{R}}_{i,3}$, so that for each ICU i, $i = 1, \ldots, G$, external emergency patients are barred from last $r^{\mathrm{R}}_{i,1}$ beds and elective patients barred from last $r^{\mathrm{R}}_{i,3}$ beds. In other words, these patients will not be admitted if the number of vacant beds at ICU i falls below the specified threshold. A graphical depiction of the threshold policy is shown in Fig. 3. Finally, let $\Gamma = (\Gamma_1, \Gamma_2, \ldots, \Gamma_G)$ denote the overflow policy of the entire network.

3.4 Numerical comparison of reservation policies

We consider an ICU network with 3 ICUs, with 20 beds in each ICU. The offered load for external emergency patients from Zone i is $\lambda_{i,1} = \lambda$ and the offered load for internal emergency and elective patients to ICU i is $\lambda_{i,2} = \lambda_{i,3} = \lambda$. The overflow policy for external emergency patients is $\Gamma = ((1, 2, 3), (2, 3, 1), (3, 1, 2))$. The network is thus symmetrical in both offered load and overflow policy. We shall also restrict the reservation policy to be the same for each ICU.

For each reservation setting, the QoS of the ICU network is evaluated using Markov-chain simulation. Simulation is terminated when either the 95% confidence interval, as computed using Student's t-distribution, lies within 1% of the simulation mean, or when thirty simulation runs have

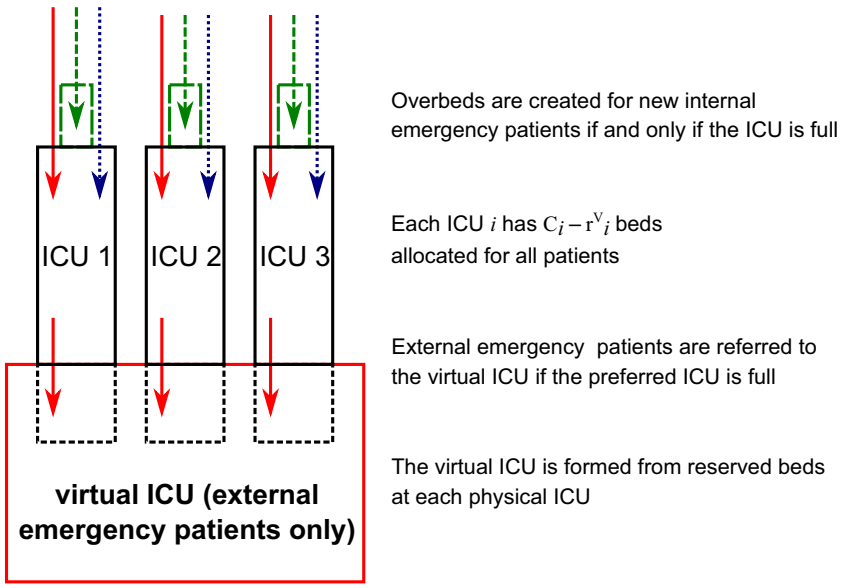

Fig. 2 Graphical depiction of the virtual ICU policy. *Solid arrows*, *dashed arrows*, and *dotted arrows* represent external emergency, internal emergency, and elective patients, respectively

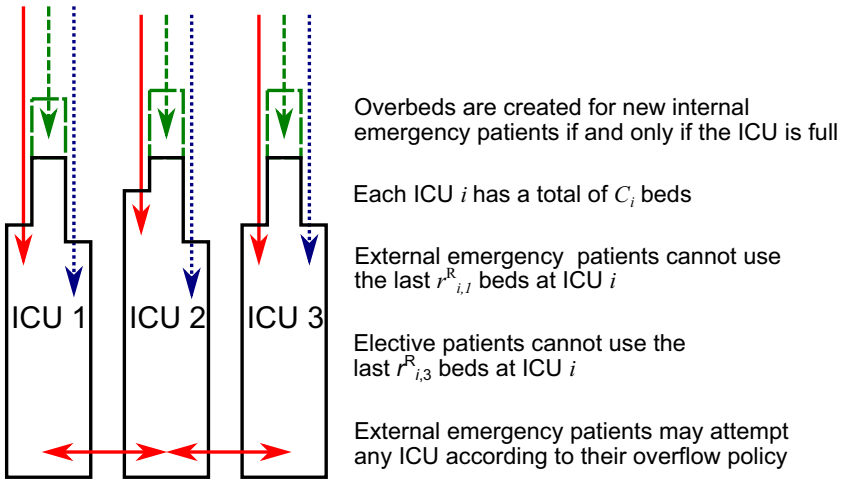

Fig. 3 Graphical depiction of the threshold policy. *Solid arrows*, *dashed arrows*, and *dotted arrows* represent external emergency, internal emergency, and elective patients, respectively

been completed. The few cases where the confidence interval does *not* fall within 1% of the simulation mean, even after thirty runs, all have the property of $B < 10^{-4}$. Such cases do not affect the results of this subsection, as the simulation error is dominated by the difference in QoS between the different reservation settings.

3.4.1 Minimizing the blocking probability of external emergency patients

For each λ in $\{5, 5.2, \ldots, 6\}$, we determine via simulation the optimal reservation settings to minimize the blocking probability B of external emergency patients, subject to $T < 0.3$ and $D < 0.25$. The results, shown in Table 3, demonstrate that the threshold policy reduces the blocking probability of external emergency patients from 60 to 85% compared to the virtual ICU policy. In addition, when we restrict $r^{R}_{i,1} = r^{R}_{i,3}$ for better comparison with the virtual ICU policy (which has only one reservation setting for each ICU), the threshold policy still results in lower blocking probability than the virtual ICU policy for $\lambda > 5.2$, with the benefits of the threshold policy increasing with λ. This demonstrates the increased level of resource sharing in the threshold policy, compared to the virtual ICU policy, has a large effect on the QoS of the network.

Table 3 Optimal reservation settings for minimizing the blocking probabilities of external emergency patients in a symmetric 3-ICU network

	2 thresholds	1 threshold	
λ	Threshold policy	Threshold policy	Virtual ICU policy
5	7.07×10^{-5}	0.00138	0.00048
	$r^{R}_{i,1} = 0, r^{R}_{i,3} = 3$	no reservation	$r^{V}_{i} = 4$
5.2	0.00015	0.00259	0.00101
	$r^{R}_{i,1} = 0, r^{R}_{i,3} = 3$	no reservation	$r^{V}_{i} = 4$
5.4	0.00067	0.00453	0.00558
	$r^{R}_{i,1} = 0, r^{R}_{i,3} = 2$	no reservation	$r^{V}_{i} = 3$
5.6	0.00281	0.00752	0.00934
	$r^{R}_{i,1} = 0, r^{R}_{i,3} = 1$	no reservation	$r^{V}_{i} = 3$
5.8	0.00455	0.0172	0.0323
	$r^{R}_{i,1} = 0, r^{R}_{i,3} = 1$	no reservation	$r^{V}_{i} = 2$
6	0.0174	0.0174	0.0441
	no reservation	no reservation	$r^{V}_{i} = 2$

Each entry shows the blocking probability B of external emergency patients and the corresponding reservation settings

Note that the optimal blocking probability for the two-threshold policy and virtual ICU policy is not continuous in λ, as would be the case if all solutions for a particular policy used the same reservation settings. As λ increases, certain reservation settings that are viable for lower λ become no longer viable as the constraints on T and D are no longer met. Note also that the optimal blocking probability for all policies is quite sensitive to the value of λ.

3.4.2 Minimizing the overall rejection rate

For each λ in $\{5, 5.2, \ldots, 6.0\}$, we determine via simulation the optimal reservation settings to minimize the overall rejection rate, subject to $B < 0.05$, $T < 0.3$ and $D < 0.25$. The overall rejection rate is defined as the proportion of patients that are either blocked or deferred:

$$\frac{\text{mean rejection rate}}{\text{total offered load}} = \frac{B\lambda + D\lambda}{3\lambda} = \frac{B + D}{3}.$$

The results, shown in Table 4, demonstrate a 32-44% decrease in rejection rate by adopting the threshold policy instead of the virtual ICU policy. In addition, comparison of the QoS demonstrates that the threshold policy can result in improved service for *all three* patient types compared to the virtual ICU policy. Finally, when we restrict $r^{R}_{i,1} = r^{R}_{i,3}$

Table 4 Optimal reservation settings for minimizing the overall rejection rate of a symmetric 3-ICU network

λ	Threshold policy	Virtual ICU policy
5	0.02391	0.04314
	$B = 0.00133$	$B = 0.00552$
	$T = 0.06127$	$T = 0.1158$
	$D = 0.06774$	$D = 0.1129$
	no reservation	$r_i^{\mathrm{V}} = 2$
5.2	0.03144	0.05391
	$B = 0.00255$	$B = 0.00937$
	$T = 0.08248$	$T = 0.1441$
	$D = 0.08669$	$D = 0.1336$
	no reservation	$r_i^{\mathrm{V}} = 2$
5.4	0.04069	0.06670
	$B = 0.00453$	$B = 0.0149$
	$T = 0.1083$	$T = 0.1762$
	$D = 0.1085$	$D = 0.1554$
	no reservation	$r_i^{\mathrm{V}} = 2$
5.6	0.05174	0.08177
	$B = 0.00752$	$B = 0.0225$
	$T = 0.1390$	$T = 0.2116$
	$D = 0.1327$	$D = 0.1779$
	no reservation	$r_i^{\mathrm{V}} = 2$
5.8	0.06471	0.09911
	$B = 0.01172$	$B = 0.03215$
	$T = 0.1741$	$T = 0.2504$
	$D = 0.1590$	$D = 0.2009$
	no reservation	$r_i^{\mathrm{V}} = 2$
6	0.07977	0.11885
	$B = 0.01472$	$B = 0.04409$
	$T = 0.2143$	$T = 0.2927$
	$D = 0.1870$	$D = 0.2243$
	no reservation	$r_i^{\mathrm{V}} = 2$

Each entry shows the overall rejection rate of the ICU network, the QoS of each patient type, and the corresponding reservation settings

for better comparison with the virtual ICU policy (which has only one reservation setting for each ICU), the threshold policy still gives a lower overall rejection rate than the virtual ICU policy.

3.4.3 Example where restricting overflow lowers the overall rejection rate

In [41], it is found that internal emergency patients have a lower arrival rate than the other patient types. We thus consider the same optimization problem as in Section 3.4.1 but with $\lambda_{i,1} = \lambda_{i,3} = \lambda$ and $\lambda_{i,2} = \lambda - 1$. The results, shown in Table 5, demonstrate a 20-43% decrease in rejection rate by adopting the threshold policy instead of the virtual ICU policy. They also demonstrate that setting $r_{i,1}^{\mathrm{R}} = r_{i,3}^{\mathrm{R}} = 0$, i.e. no reservation, does not necessarily result in the lowest rejection rates, unlike in Section 3.4.1. This is despite maximal resource sharing in the sense that external patients have access to all beds in all ICUs in the network and no patient is ever barred from an ICU if there is at least one bed available. Instead, for $\lambda = 5$ or 5.2, we obtain the counter-intuitive result of reduced overall rejection rate when the overflow of external emergency patients is restricted. This is because overflowing external emergency patients adversely affect the QoS of internal emergency and elective patients. A similar effect was observed by Gurumurthi and Benjaafar [21]; however, in their work, the authors control overflow by changing the routing policy itself and do not consider reservation.

Table 5 Optimal reservation settings for minimizing the overall rejection rate of a 3-ICU network, with reduced arrival rates of internal emergency patients

λ	Threshold policy	Virtual ICU policy
5	0.01200	0.02119
	$B = 0.00246$	$B = 0.00902$
	$T = 0.01971$	$T = 0.03973$
	$D = 0.02862$	$D = 0.05454$
	$r_{i,1}^{\mathrm{R}} = 1, r_{i,3}^{\mathrm{R}} = 0$	$r_i^{\mathrm{V}} = 1$
5.2	0.01789	0.02775
	$B = 0.00492$	$B = 0.01419$
	$T = 0.02842$	$T = 0.05340$
	$D = 0.03893$	$D = 0.06905$
	$r_{i,1}^{\mathrm{R}} = 1, r_{i,3}^{\mathrm{R}} = 0$	$r_i^{\mathrm{V}} = 1$
5.4	0.02542	0.03550
	$B = 0.00160$	$B = 0.02128$
	$T = 0.05521$	$T = 0.06987$
	$D = 0.07144$	$D = 0.08523$
	no reservation	$r_i^{\mathrm{V}} = 1$
5.6	0.03325	0.04439
	$B = 0.00298$	$B = 0.03039$
	$T = 0.07406$	$T = 0.08907$
	$D = 0.09079$	$D = 0.1028$
	no reservation	$r_i^{\mathrm{V}} = 1$
5.8	0.04275	0.05441
	$B = 0.00517$	$B = 0.04162$
	$T = 0.09702$	$T = 0.1111$
	$D = 0.1127$	$D = 0.1216$
	no reservation	$r_i^{\mathrm{V}} = 1$
6	0.05406	0.06807
	$B = 0.00839$	$B = 0.02555$
	$T = 0.1241$	$T = 0.1831$
	$D = 0.1370$	$D = 0.1786$
	no reservation	$r_i^{\mathrm{V}} = 2$

Each entry shows the overall rejection rate of the ICU network, the QoS of each patient type, and the corresponding reservation settings

3.4.4 Robustness to increases in the offered load

Certain events such as the outbreak of an infectious disease may cause short-term spikes in the arrival rate of patients to an ICU network. In order to demonstrate that the benefits of the threshold policy over the virtual ICU policy are not dependent on the offered load, we consider the $\lambda = 5.4$ case from Table 3. Using the optimal reservation settings for both policies for $\lambda = 5.4$, we examine the effect on B, T, and D as the offered load is increased by up to 20%. The results are shown in Fig. 4. As the offered load increases, the gap in B between the threshold policy and the virtual ICU policy also increases. On the other hand, T and D are about the same for both policies. In other words, the threshold policy is robust to increases in the offered load in the sense that the threshold policy continues to achieve a better QoS than the virtual ICU policy when the arrival rates are increased (with the reservation settings fixed).

4 Sensitivity to the patient length-of-stay distribution

Litvak et al. [41] demonstrated via simulation that their ICU network model, using their virtual ICU policy, is not very sensitive to the shape of the patient LoS distribution apart from its mean. In this section, we show that this near insensitivity also applies to the threshold policy. We consider the same 3-ICU network as Section 3.4 and generate 1000 random configurations, with $5.0 \leq \lambda_{i,t} \leq 6.0$, $0 \leq r^{\mathrm{R}}_{i,1} \leq 3$, and $0 \leq r^{\mathrm{R}}_{i,3} \leq 3$ for each $i = 1, 2, \ldots G$ and $t = 0, 1, 2$. The number of simulation runs is such that the 95% confidence interval, as computed using Student's t-distribution, lies within 1% of the simulation mean.

Let B^x, T^x, and D^x denote the blocking probability of external emergency patients, mean number of overbeds for internal emergency patients, and deferral probability of elective patients, respectively, for a lognormal LoS distribution with mean 1.0 and variance x, as found by simulation. Let B, T, and D denote the same values for an exponential LoS distribution, also with a mean of 1.0. The distributions of the ratios B^x/B, T^x/T and D^x/D are shown in Fig. 5 for $x \in \{0.5, 2.0, 4.0\}$. The results suggest that the QoS of our ICU network is not very sensitive to the patient LoS distribution, with all results within the interval [0.98, 1.02].

5 Estimating the QoS of an ICU network

While accurate approximations exist for ICU networks using the virtual ICU policy [41], the presence of mutual overflow under the threshold policy means that estimation of QoS becomes considerably more difficult [56]. Additionally, although there are similarities between the ICU network and other overflow systems such as telecommunications systems and call centers, there are also some fundamental differences. For example, our ICU network considers three different patient types, of which only one type may overflow. Furthermore, the concept of an over-bed is unique to the current ICU network model. These differences make the problem in this paper challenging. In this section, we examine and compare several approximations for QoS in an ICU network under the threshold policy, and show how they can be extended to apply to the current ICU network model.

5.1 Markov chain representation of a single ICU

We start by making the simplifying assumption that all traffic offered to an ICU, including overflow traffic, follows a Poisson process. Let $a_{i,n}$ denote the offered traffic of external emergency patients from Zone i which have overflowed n times in the network. Then

$$\begin{aligned} a_{z,0} &= \lambda_{z,1} \\ a_{z,n} &= a_{z,n-1} b_{\Gamma_{z,n-1}}, \quad n > 0. \end{aligned} \tag{1}$$

Let x_i denote the total offered traffic of external emergency patients to ICU i. Then

$$x_i = \sum_{z=1}^{G} \sum_{n:\Gamma_{z,n}=i} a_{z,n}. \tag{2}$$

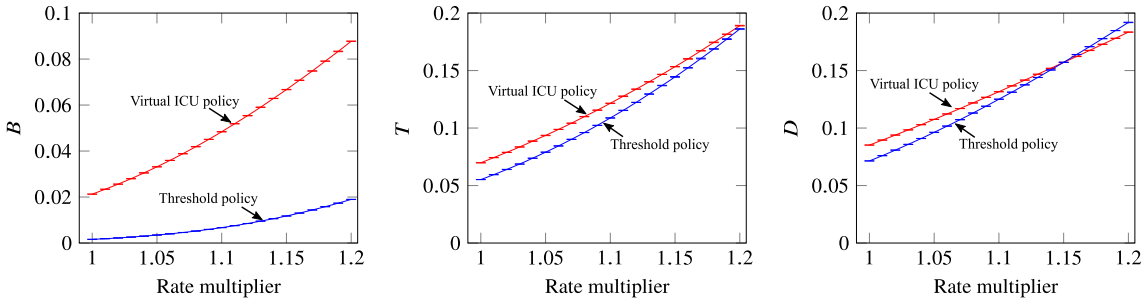

Fig. 4 QoS of a three-ICU network with respect to increases in the offered load

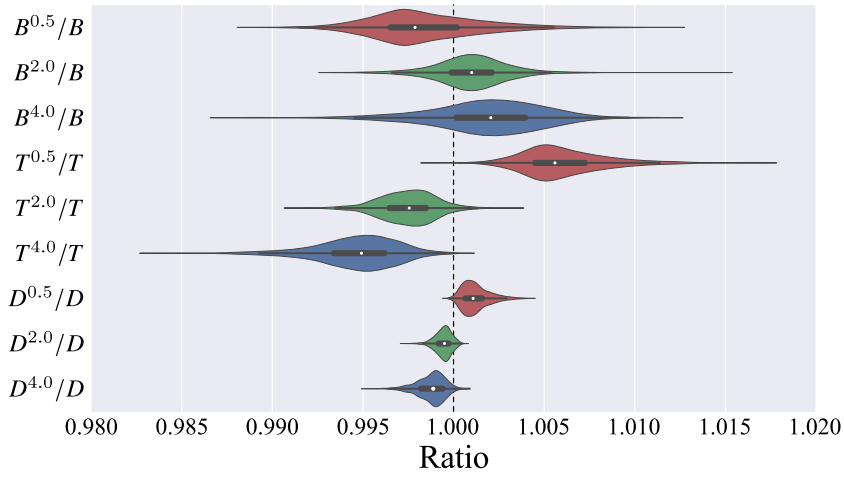

Fig. 5 Sensitivity of B, T, and D to the patient LoS distribution. The superscript represents the variance of a lognormal LoS distribution, whereas no superscript represents an exponential LoS distribution

Thus ICU i receives a total offered load of $x_i + \lambda_{i,2} + \lambda_{i,3}$ Erlangs.

By assuming that the arrival process to each ICU is a Poisson process, we obtain a one-dimensional Markov chain representation for each ICU $i, i = 1, \ldots, G$, as follows. Let state j denote the state in which there are j patients in service, and $q_{j,k}$ be the transition rate from state j to state k. Then

$$\begin{aligned} q_{j,j+1} &= x_i \mathbf{1}\left\{j < C_i - r_{i,1}^{\mathrm{R}}\right\} + \lambda_{i,2} + \lambda_{i,3}\mathbf{1}\left\{j < C_i - r_{i,3}^{\mathrm{R}}\right\} \\ q_{j,j-1} &= j \\ q_{j,k} &= 0, \qquad |j-k| \neq 1, \end{aligned}$$

where $\mathbf{1}\{\cdot\}$ represents the indicator function.

From the transition rate matrix $q_{[j,k]}$, we can obtain the probability of each state j, $j \in \mathbb{N}$, which we denote as π_j. Then

$$\begin{aligned} b_i &= \sum_{j=C_i - r_{i,1}^{\mathrm{R}}}^{\infty} \pi_j \\ T_i &= \sum_{j=C_i+1}^{\infty} \pi_j \, (j - C_i) \\ D_i &= \sum_{j=C_i - r_{i,3}^{\mathrm{R}}}^{\infty} \pi_j . \end{aligned} \tag{3}$$

For a through discussion on Markov chains, see Norris [46].

5.2 ED

ED can be applied to the ICU network by treating x_i for each ICU i, $i = 1, 2, \ldots, G$, as mutually independent. This results in a system of fixed-point equations involving $(x_i)_{i=1}^{G}$ and $(b_i)_{i=1}^{G}$, which can be solved via iterative substitution [6] using (1)–(3). The stopping criterion is defined as follows. Let $b_i^{(k)}$ denote the k^{th}-iteration estimate of b_i. The fixed-point iteration is terminated when $\left|b_i^{(k)} - b_i^{(k-1)}\right| < 10^{-8}$ for all $i = 1, 2, \ldots, G$.

After obtaining x_i and b_i for each ICU $i, i = 1, 2, \ldots, G$, the quantity B_i can be obtained as the product of the congestion probabilities for each ICU in Γ_i, i.e. $B_i = \prod_{j \in \Gamma_i} b_j$.

5.3 IESA

IESA [7, 8, 57, 59] is based on the applying the underlying methodology of ED, namely decomposition of the ICU network into a set of independent queues with Poisson input, to a *surrogate* model of the original network, so that the dependencies between ICUs are represented in a manner that is preserved when decomposition is applied. In the IESA surrogate model, each external emergency patient carries three attributes: z, the originating zone, Δ, the set of attempted ICUs, and Ω, an estimate of the number of ICUs in the network currently refusing external emergency patients. All new patients start with $\Delta = \emptyset$ and $\Omega = 0$. We will use the term (z, Δ, Ω)-patient to denote a external emergency patient from Zone z which has attempted each ICU in Δ and has a congestion estimate of Ω. Unlike the "true" model of the ICU network, in addition to blocking if all ICUs have been attempted unsuccessfully, external emergency patients in the IESA model will also *abandon* the network if Ω reaches G.

Consider a $(z_1, \Delta_1, \Omega_1)$-patient attempting ICU i. If a bed is available at ICU i for external emergency patients, the patient is admitted. Otherwise, the patient is compared to the external emergency patient with the highest Ω values among all external emergency patients residing at ICU i, which we denote as an $(z_2, \Delta_2, \Omega_2)$-patient. Ties are broken arbitrarily. If $\Omega_1 \geq \Omega_2$, then the incoming patient overflows normally and becomes a $(z_1, \Delta_1 \cup \{i\}, \Omega_1 + 1)$-patient. On the other hand, if $\Omega_1 < \Omega_2$, then *exchange* of Ω occurs and the incoming patient overflows as an

$(z_1, \Delta_1 \cup \{i\}, \Omega_2 + 1)$-patient, while the admitted patient becomes an $(z_2, \Delta_2, \Omega_1)$-patient. Note that due to these rules, $\Omega \geq |\Delta|$ for all incoming patients.

IESA thus forms an hierarchical traffic structure based on Ω, where level j of the hierarchy includes all patients with Ω less than or equal to j. Due to abandonment when $\Omega = G$, the hierarchy has a total of exactly G layers, from 0 to $G-1$. Due to this hierarchy, IESA does not require fixed-point iterations when applied to our ICU network model, unlike EFPA.

Let:

- $e_{z,n,j}$ denote the offered traffic to ICU $\Gamma_{z,n}$ composed of external emergency patients from Zone z which have overflowed n times in the network and have a congestion estimate of j;
- $\tilde{e}_{z,n,j}$ denote the offered traffic to ICU $\Gamma_{z,n}$ composed of external emergency patients from Zone z which have overflowed n times in the network and have a congestion estimate of $0, 1, \ldots$ or j;
- $a_{i,n,j}$ denote the offered traffic to ICU i composed of all external emergency patients which have overflowed n times in the network and have a congestion estimate of j;
- $\tilde{a}_{i,n,j}$ denote the offered traffic to ICU i composed of all external emergency patients which have overflowed n times in the network and have a congestion estimate of $0, 1, \ldots$ or j;
- $A_{i,j}$ denote the offered traffic to ICU i composed of all external emergency patients with congestion estimate $0, 1, \ldots$ or j; and
- $b_{i,j}$ denote the congestion probability of ICU i for external emergency patients with congestion estimate $0, 1, \ldots$ or j.

By definition,

$$e_{z,0,j} = \begin{cases} \lambda_{z,1}, & j = 0 \\ 0, & \text{otherwise,} \end{cases}$$

$\tilde{e}_{z,n,j} = \sum_{k=n}^{j} e_{z,n,k}$, and $\tilde{a}_{i,n,j} = \sum_{k=n}^{j} e_{i,n,k}$. Summing over all possible z,

$$a_{i,n,j} = \sum_{z:\Gamma_{z,n}=i} e_{z,n,j}.$$

Summing over all possible n,

$$A_{i,j} = \sum_{n=0}^{G-1} \tilde{a}_{i,n,j}.$$

From $A_{i,j}$, $\lambda_{i,2}$, and $\lambda_{i,3}$, $b_{i,j}$ can be computed via Markov-chain analysis as described in Section 5.1 . In accordance with the information exchange mechanism, we obtain,

$$\begin{aligned} e_{z,n,j} = \; & e_{z,n-1,j-1} b_{\Gamma_{z,n-1},j-1} \\ & + e_{z,n-1,j-2} \left(b_{\Gamma_{z,n-1},j-1} - b_{\Gamma_{z,n-1},j-2} \right). \end{aligned} \tag{4}$$

The above values can be obtained iteratively for $j = 0, 1, \ldots, G-1$. Finally, the blocking probability of external emergency patients in zone i is

$$B_i = \sum_{n=1}^{G-1} e_{i,n,G}. \tag{5}$$

Note that (5) is a slight abuse of notation as patients with a congestion estimate of G are never offered to any ICU; however, defining $e_{z,n,G}$ as per (4) yields the correct result for (5).

The values of T_i and D_i can be estimated from the last (i.e. $G - 1^{\text{th}}$) level of the IESA hierarchy using the same Markov-chain analysis as for $b_{i,G-1}$.

5.4 Numerical comparison of ED and IESA

We consider ICU networks of $G = 3, 4$, or 5 ICUs. External emergency patients are referred to an ICU in a round-robin manner: thus an external emergency patient from zone i will attempt ICUs $i, i+1 \ldots, G, 1, 2, \ldots i-1$, in that order. For each value of G , we generated 500 random configurations with the following parameters:

- 15-20 beds in each ICU ($15 \leq C_i \leq 20$ for $i = 1, 2, \ldots G$),
- reservation thresholds $r^{\mathrm{R}}_{i,1}$ and $r^{\mathrm{R}}_{i,3}$ of 0 to 3 for each ICU i, and
- arrival rates $\lambda_{i,t}$ of $0.25C_i$ to $0.3C_i$ for each ICU i and for each patient type t.

The configurations were then filtered according to the following conditions:

- a blocking probability B of between 0.1 and 5% for external emergency patients, as estimated by IESA;
- at most $0.1G$ overbeds ($T \leq 0.1G$), as estimated by IESA; and
- a deferral probability D of at most 25% for elective patients, as estimated by IESA.

The number of valid configurations were 427, 452, and 372 for $G = 3, 4$, and 5, respectively. For each valid configuration, B, T,and D were evaluated using Markov chain simulation, ED, and IESA. The number of simulation runs is such that the 95% confidence interval, as computed using Student's t-distribution, lies within 1% of the simulation mean. The relative errors of ED and IESA are shown in Fig. 6. The results demonstrate that IESA is much more accurate than ED when estimating the blocking probability of external emergency patients. On the other hand, both approximations are fairly accurate for internal emergency and elective patients, with ED being slightly more accurate than IESA.

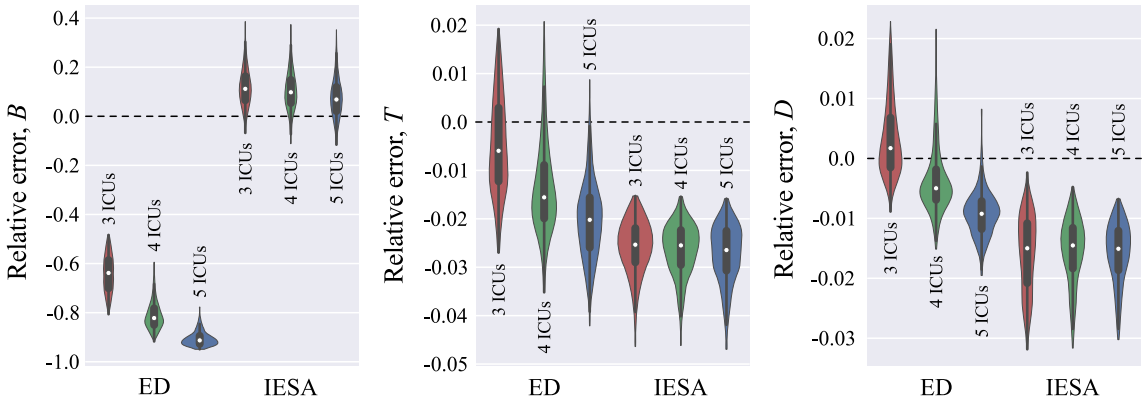

Fig. 6 Relative errors for B, T, and D for ED and IESA. Each shaded area is truncated to show the extremums of the observed data

6 Obtaining a conservative estimate for patient QoS

When dimensioning an ICU network, it is generally necessary to ensure that the QoS estimates for some patients are conservative. For example, if one of the optimization constraints is that the deferral probability D of elective patients must not exceed $D_{\max}$, then any estimation of D must be equal to or greater than the actual value of D. In this section, we demonstrate a method of obtaining conservative estimates of T, the mean number of overbeds in the network, and D, the deferral probability of elective patients.

6.1 Hayward's approximation

It has long been recognized that overflow traffic in overflow loss systems has a higher peakedness (variance-to-mean ratio) than fresh traffic, and that such peakedness increases the blocking probability of requests offered to the system. For a G/M/N/N queue offered traffic with mean m and variance v, with $z = v/m$, a simple but effective blocking probability approximation is provided by Hayward:

$$B(m, v, N) = B\left(\frac{m}{z}, \frac{N}{z}\right).$$

This is equivalent to splitting the system into z independent G/M/$\frac{N}{z}$/$\frac{N}{z}$ queues, thus raising the blocking probability of the system as servers in different queues now cannot coordinate to reduce congestion in the system. In many cases, N/z will not be an integer; Jagerman [25] gives an analytic continuation of the Erlang B function for such cases.

To adapt Hayward's approximation to an ICU network model with threshold reservation, we construct a Markov chain as follows. Let x_i be the offered load of external emergency patients to ICU i and let v_i be the corresponding variance. Then the total offered traffic to ICU i has mean $M_i = x_i + \lambda_{i,2} + \lambda_{i,3}$ and variance $V_i = v_i + \lambda_{i,2} + \lambda_{i,3}$. Define $z_i = V_i/M_i$.

We split the ICU into z_i independent parts so that each part contains C_i/z_i beds and the offered traffic to each part composed of external emergency, internal emergency, and elective patients is Poisson with means $a_{i,1} = x_i/z_i$, $a_{i,2} = \lambda_{i,2}/z_i$, and $a_{i,3} = \lambda_{i,3}/z_i$, respectively. The reservation thresholds for external emergency and elective patients become $r^{\mathrm{R}}_{i,1}/z_i$ and $r^{\mathrm{R}}_{i,1}/z_i$, respectively.

Non-integer ICU sizes and reservation thresholds are handled as follows. As in Section 5.1, let state j denote the state in which there are j patients in service, and $q_{j,k}$ be the transition rate from state j to state k. Let $c_{i,1} = \left(C_i - r^{\mathrm{R}}_{i,1}\right)/z_i$, $c_{i,2} = C_i/z_i$, and $c_{i,3} = \left(C_i - r^{\mathrm{R}}_{i,3}\right)/z_i$. Let $n_{i,t}$ and $f_{i,t}$ be the integer and fractional parts of $c_{i,t}$, respectively, for $t = 1$, 2, or 3. Furthermore, define

$$u_{i,j,t} = \begin{cases} a_{i,t}, & j < n_{i,t} \\ a_{i,t} f_{i,t}, & j = n_{i,t} \\ 0, & j > n_{i,t}. \end{cases}$$

Then

$$\begin{aligned} q_{j,j+1} &= u_{i,j,1} + u_{i,j,2} + u_{i,j,3} \\ q_{j,j-1} &= j \\ q_{j,k} &= 0, \qquad |j-k| \neq 1, \end{aligned}$$

from which the steady-state probability of each state j can be obtained. Finally,

$$b_i = (1 - f_1)\pi_{n_{i,1}} + \sum_{j=n_{i,1}+1}^{\infty} \pi_j$$

$$T_i = \sum_{j=\lceil n_{i,2} \rceil}^{\infty} \pi_j (j - n_{i,2})$$

$$D_i = (1 - f_3)\pi_{n_{i,3}} + \sum_{j=n_{i,3}+1}^{\infty} \pi_j.$$

6.2 Overflow variance of external emergency patients

Let $a = x_{i,k}$ denote the offered traffic to ICU i composed of external emergency patients that have overflowed k times in the system, and let v denote the corresponding variance. Let $z = v/a$. To estimate the overflow traffic of patients corresponding to this input stream, we construct an M/M/n/n queue offered $a' = a/z$ Erlangs of Poisson traffic so that $E\left(a', n\right) = b_i$. A method of computing n is given by Jagerman [25].

The overflow mean and variance of the imaginary queue are $a'_{out} = a'b_i$ and

$$v'_{out} = a'_{out}\left[1 - a'_{out} + \frac{a'}{n - a' + a'_{out} + 1}\right],$$

respectively, with the latter formula given by Riordan [55, Appx. I]. Finally, the overflow variance from ICU i composed of emergency patients that have overflowed $k + 1$ times is estimated as $v'_{out}z$.

6.3 Numerical results

By using the methods described in Sections 6.1 and 6.2, we can create a moment-matched version of ED, which we call EDm. Using the same configurations as in Section 5.4, we obtain the results shown in Fig. 7, which demonstrate that EDm gives conservative estimates of T and D for the ICU network.

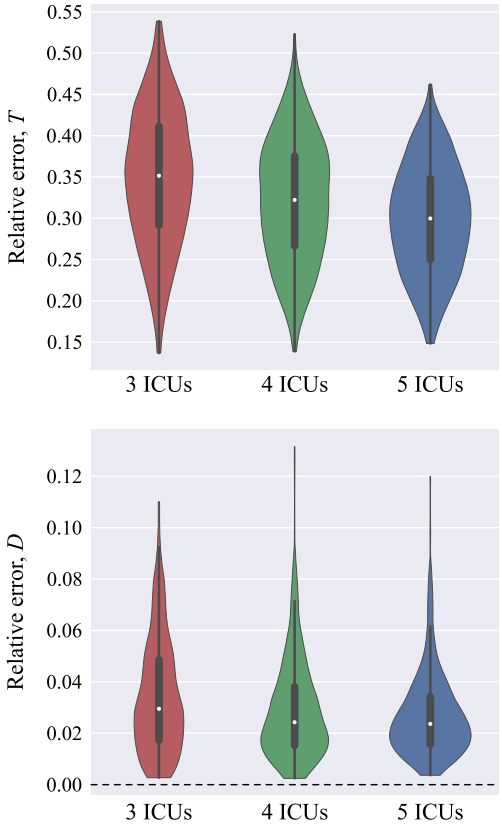

Fig. 7 Relative errors for T, and D for EDm. Each shaded area is truncated to show the extremums of the observed data

7 Minimizing the blocking of external emergency patients

From the randomly generated configurations from Section 5.4, we select 100 configurations for $G = 3$ and $G = 4$,and 48 configurations for $G = 5$. For each network, we use exhaustive search to solve the following problems:

$$\begin{aligned} \underset{\left\{r^V_{i,1}, |i=1,\ldots G\right\}}{\arg\min} \quad & B_V \\ \text{s.t.} \quad & T_V < 0.1G \\ & D_V < 0.25 \\ & \forall i,\ r^V_{i,1} \le 10 \end{aligned} \tag{P1}$$

and

$$\begin{aligned} \underset{\left\{r^R_{i,1}, r^R_{i,3} |i=1,\ldots G\right\}}{\arg\min} \quad & B_R \\ \text{s.t.} \quad & T_R < 0.1G \\ & D_R < 0.25 \\ & \forall i,\ r^R_{i,1} \le 5 \quad \forall i,\ r^R_{i,3} \le 5, \end{aligned} \tag{P2}$$

where the subscripts V and R represent the virtual ICU policy and our proposed threshold reservation policy, respectively. We use IESA to approximate B_R, and EDm, which we showed in Section 6.3 to be conservative, to approximate T_R and D_R. For the virtual ICU policy, we use the approximation method defined in Litvak et al. [41]. Let B^*_V and B^*_R denote the optimal values of $B_V$and B_R, respectively.

In Table 6, we show both the mean and standard deviation of B^*_V/B^*_R, which represents the reduction in blocking probability of external emergency patients, using both approximation and Markov-chain simulation. The results demonstrate a much lower blocking probability of external emergency patients when the threshold reservation policy is used (i.e. $B^*_R \ll B^*_V$), with the difference in performance between the two policies increasing with the number of ICUs in the network. Note that simulation in used here *only* to evaluate the QoS of the final configuration returned by the optimization process.

Table 6 Ratio of B^*_V to B^*_R for the optimal reservation settings for an ICU network under the virtual ICU and threshold policies

Number of ICUs	B^*_V/B^*_R (estimated)		B^*_V/B^*_R (simulated)	
	Mean	St. dev.	Mean	St. dev.
3	3.8063	0.7557	4.7747	0.7487
4	8.7156	1.6009	11.793	1.6914
5	19.3404	4.3810	28.236	4.4632

Table 7 Running times for optimizing reservation thresholds for networks of 3, 4, and 5 ICUs

Number of ICUs	Running time (s)	
	Mean	St. dev.
3	29.706	1.8479
4	1957.9	83.538
5	114243	3732.8

The results in Table 6 show that IESA is conservative for estimating B_R^* for the optimal reservation setting in all cases considered. As ERM is very accurate at estimating B_V in ICU networks using the virtual ICU policy, the end result is that the estimation of B_V^*/B_R^* is also conservative for all cases considered. In addition, since EDm is conservative for estimating T_R and D_R, all solutions found for the threshold policy are valid. In other words, our approximate approach provides not only a valid solution in all cases to optimization problem (P2), but also conservative estimates of the QoS and the amount of improvement achieved over the virtual ICU policy.

7.1 Running times

For our optimization algorithm for the threshold reservation policy, the running times for $G = 3$, 4, and 5 are shown in Table 7. It is demonstrated that due to the speed of EDm and IESA, it is possible to perform exhaustive search for networks of up to 5 ICUs. The average speed as calculated for $G = 5$ is 529.3 QoS evaluations per second (where one evaluation includes B, T, and D).

8 Concluding remarks

We have proposed a new threshold-based patient referral policy for the admission of patients to a network of ICUs and shown that our new policy can achieve a higher patient acceptance level than the previously proposed policy [41] using a smaller number of beds, resulting in improved service for all patients. Our proposed policy incorporates important concepts and insights from traditional teletraffic theory, including the overflow loss model with multiple streams of calls (i.e. patients), resource sharing improved by allowing mutual overflow for overflow calls (i.e. external emergency patients), and trunk reservation for reserving the last unused amount of resource at each node (i.e. the last few unused beds at each ICU that external emergency patients cannot use) for providing sufficient service level for non-overflow calls (i.e. internal emergency and elective patients).

In particular, we focus on the problem of minimizing the blocking probability of external emergency patients in an ICU network subject to meeting minimum QoS requirements for internal emergency and elective patients. This is achieved in three parts: (1) our new threshold-based policy for patient referral, (2) accurate and computationally efficient analytical approximation methods for estimating the QoS of an ICU network under our proposed policy, and (3) the incorporation of these approximation methods into an algorithm for quickly determining the optimal reservation thresholds for each ICU in the network. In addition to the combining existing concepts and methods to construct a comprehensive design, QoS evaluation, and optimization framework, new contributions include the construction of a new moment-matching method specific to the ICU network model with a threshold reservation policy and the unique combination of IESA and EDm to create a conservative analytical approximation method for all three patient types.

Numerical results demonstrate that our proposed threshold policy is more efficient than the virtual ICU policy of Litvak et al. [41]. This is because our proposed threshold policy enhances the level of resource sharing by allowing an external emergency patient to be assigned to any ICU bed in the network as long as none of the vacancy thresholds are violated. On the other hand, under the policy of Litvak et al. [41], an external emergency patient is only allowed to overflow to a dedicated group of ICU beds, thus limiting the level of resource sharing. Therefore, our proposed policy can achieve a lower blocking probability for external emergency patients than the previous policy given the same QoS requirements for internal emergency and elective patients. Alternatively, the threshold policy can reduce the *overall* rejection rate of an ICU network compared to virtual ICU policy (a reduction of 20–44% was achieved in our numerical examples). Enhanced cooperation between hospitals by improving resource sharing can achieve a higher acceptance level with a smaller number of beds resulting in improved service for *all* patients in this scenario. On the other hand, we have also shown a new interesting result: *maximizing* resource sharing, by allowing external emergency patients to attempt any ICU and setting no reservation thresholds whatsoever, does not necessarily lead to the lowest overall patient rejection rate.

Numerical results also demonstrate that the QoS of ICU networks using the threshold policy is not very sensitive to the patient LoS distribution apart from its mean, meaning that the QoS approximations developed in the second part of this paper are applicable to a wide range of QoS networks with different patient LoS distributions.

Numerical results also demonstrate that IESA can provide a much more accurate estimate of the blocking probability of external emergency patients than the classical method, ED. On the other hand, while ED by itself can achieve relatively accurate estimates of both the mean number of overbeds required for internal emergency patients and

the mean deferral probability of elective patients, such estimates are not conservative, a requirement of our proposed optimization algorithm. We therefore presented a version of ED incorporating moment-matching, namely EDm, which was demonstrated to be conservative in all our numerical tests.

Optimization of ICU networks is performed in this paper using a unique combination of IESA for the QoS evaluation of external emergency patients and EDm for the QoS evaluation of non-overflow patients (i.e. internal emergency and elective patients). The speed of IESA and EDm allows us to use exhaustive search for our optimization algorithm for ICU networks of up to five ICUs. Numerical results demonstrate much better QoS (e.g. an average of 4.7, 11.7, and 28.2 times blocking probability reduction for external emergency patients in a 3-ICU, 4-ICU, and 5-ICU network, respectively) can be achieved by our proposed threshold policy, using our approximation-based optimization algorithm, than for the virtual ICU policy, using the optimization method of Litvak et al. [41]. Additionally, our approximate approach provides not only a valid solution in all cases for the threshold policy, but also conservative estimates of the QoS and amount of improvement achieved over the virtual ICU policy.

In conclusion, our proposed patient referral policy, QoS approximation methods, and optimization algorithm combined together form an effective and computationally efficient new framework for achieving much better service for all patients while meeting the QoS requirements for each individual patient type in an ICU network.

8.1 Challenges for implementing the threshold policy in a real ICU network

In order to locate a suitable ICU bed under the threshold policy, a centralized system will be required to track the occupancy level of each ICU in the network. This system may also be made accessible to ambulatory services so that patients that are likely to require ICU stay can be transported to a suitable hospital before even being admitted to the hospital system. The cost of such a system would depend on ICU and hospital electronic systems that may already be present for monitoring patient flow. Many modern ICUs (and hospitals) already have integrated electronic clinical information systems that would allow a platform for a program such as this.

Another challenge towards implementation of the threshold policy is that the analytical model itself is not completely realistic, as detailed in Table 1. In particular, while our current model assumes that external emergency patients may be assigned to any ICU, in reality certain ICUs may be more suitable than others based on the availability of specialist care (e.g. cardiothoracic, neurosurgery, multi-trauma). Future work is required to extend the model for such cases before it can be applied to a real ICU network.

8.2 Future work

Future work may involve extending the ICU network model to more closely resemble a physical system, addressing some of the differences outlined in Table 1. For example, the optimization problem may be modified to factor in the costs of patients being admitted to a non-primary choice ICU, as well as the costs associated with implementing and maintaining the new policy. The interactions between the ICU network, AEDs, and transport services may also be incorporated into the model. Nevertheless, it is expected that the cost of the proposed threshold policy will be substantially less than the savings and quality of care gained by switching to the proposed policy.

Finally, heuristic optimization techniques can be used to both reduce the computational time needed to set the reservation parameters of each ICU, and to allow the problem to be solved for larger ICU networks, possibly covering an entire city. Although the use of heuristics may lead to a sub-optimal solution, it is expected that the configurations produced will still outperform existing policies for ICU patient referral. In addition, in this paper we only consider optimization of the reservation parameters; future work may involve dimensioning of the ICUs themselves (i.e. determining the optimal number of beds). The use of heuristics will allow us to deal with the additional number of decision variables that need to be handled in such an optimization problem in a computationally efficient manner.

References

1. Bai J, Fügener A, Schoenfelder J, Brunner JO (2016) Operations research in intensive care unit management: a literature review. Health Care Manag Sci http://doi.org/10.1007/s10729-016-9375-1
2. Barratt H, Harrison DA, Rowan KM, Raine R (2012) Effect of non-clinical inter-hospital critical care unit to unit transfer of critically ill patients: a propensity-matched cohort analysis. Crit Care 16(R179):1–10
3. Bekker R, Koole G, Roubos D (2016) Flexible bed allocations for hospital wards. Health Care Manag Sci, pp 1–14, http://doi.org/10.1007/s10729-016-9364-4
4. Berry Jr. RE (1967) Returns to scale in the production of hospital services. Health Serv Res 2(2):123–139
5. Blair EL, Lawrence CE (1981) A queueing network approach to health care planning with an application to burn care in New York State. Socio-Econ Plann Sci 15(5):207–216
6. Browder FE, Petryshyn WV (1966) The solution by iteration of nonlinear functional equations in Banach spaces. Bullet Amer Math Soc 72(3):571–575
7. Chan YC, Guo J, Wong EWM, Zukerman M (2015) Performance analysis for overflow loss systems of processor-sharing queues. In: Proceedings of the IEEE INFOCOM '15, pp 1409–1417

8. Chan YC, Guo J, Wong EWM, Zukerman M (2016) Surrogate models for performance evaluation of multi-skill multi-layer overflow loss systems. Perform Eval 104:1–22
9. Christian MD, Joynt GM, Hick JL, Colvin J, Danis M, Sprung CL (2010) Critical care triage. In: Recommendations and standard operating procedures for intensive care unit and hospital preparations for an influenza epidemic or mass disaster. Summary report of the European Society of Intensive Care Medicine's Task Force for intensive care unit triage during an influenza epidemic or mass disaster, Intensive Care Medicine, vol 36 (Suppl 1), chap 7, pp 55–64
10. Chrusch CA, Olafson KP, McMillan PM, Roberts DE, Gray PR (2009) High occupancy increases the risk of early death or readmission after transfer from intensive care. Crit Care Med 37(10):2753–2758
11. Cooper RB, Katz SS (1964) Analysis of alternate routing networks with account taken of nonrandomness of overflow traffic. Memo. MM64-3122-2, Bell Telephone Laboratories
12. Duke GJ, Green JV (2001) Outcome of critically ill patients undergoing interhospital transfer. Med J Aust 174(3):122–125
13. Erlang AK (1917) Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. In: Brockmeyer E, Halstrøm HL, Jensen A (eds) The life and works of A.K. Erlang, no. 2 in transactions of the Danish Academy of Technical Sciences, pp 138–155
14. Esogbue AO, Singh AJ (1976) A stochastic model for an optimal priority bed distribution problem in a hospital ward. Oper Res 24(5):884–898
15. Everitt DE (1994) Traffic engineering of the radio interface for cellular mobile networks. Proc IEEE 82(9):1371–1382
16. Franx GJ, Koole G, Pot A (2006) Approximating multi-skill blocking systems by hyperexponential decomposition. Perform Eval 63(8):799–824
17. Graves SC (2008) Flexibility principles. In: Building intuition, international series in operations research & management science, vol 115. Springer, chap 3, pp 33-49
18. Griffiths JD, Price-Lyold N, Smithies M, Williams JE (2005) Modelling the requirement for supplementary nurses in an intensive care unit. J Oper Res Soc 56(2):126–133
19. Griffiths JD, Price-Lyold N, Smithies M, Williams J (2006) A queueing model of activities in an intensive care unit. IMA J Manag Math 17(3):277–288
20. Griffiths JD, Jones M, Read MS, Williams JE (2010) A simulation model of bed-occupancy in a critical care unit. J Simul 4(1):52–59
21. Gurumurthi S, Benjaafar S (2004) Modeling and analysis of flexible queueing systems. Naval Res Logist 51(5):755–782
22. Hennion B (1979) Feedback methods for calls allocation on the crossed traffic routing. In: Proceedings of the 9th international teletraffic congress (ITC 9)
23. Hong D, Rappaport SS (1986) Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. IEEE Trans Vehicular Tech 35(3):77–92
24. Huang Q, Ko KT, Iversen VB (2008) Approximation of loss calculation for hierarchical networks with multiservice overflows. IEEE Trans Commun 56(3):466–473
25. Jagerman DL (1984) Methods in traffic calculations. AT&T Bell Labor Tech J 63(7):1283–1310
26. Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. Manag Sci 41(4):577–594
27. Joynt G, Gomersall C, Tan P, Lee A, Cheng C, Wong E (2001) Prospective evaluation of patients refused admission to an intensive care unit: triage, futility, and outcome. Intensive Care Med 27(9):1459–1465
28. Joynt GM, Loo S, Taylor BL, Margalit G, Christian MD, Sandrock C, Davis M, Leoniv Y, Sprung CL (2010) Coordination and collaboration with interface units. In: Recommendations and standard operating procedures for intensive care unit and hospital preparations for an influenza epidemic or mass disaster. Summary report of the European Society of Intensive Care Medicine's Task Force for intensive care unit triage during an influenza epidemic or mass disaster, Intensive Care Medicine, vol 36 (Suppl 1), chap 3, pp 21–31
29. Kahn CA, Stratton SJ, Anderson CL (2014) Characteristics of hospitals diverting ambulances in a California EMS system. Prehospital Disaster Med 29(1):27–31
30. Kaufman JS (1981) Blocking in a shared resource environment. IEEE Trans Commun 29(10):1474–1481
31. Kelly FP (1986) Blocking probabilities in large circuit-switched networks. Adv Appl Probab 18(2):473–505
32. Kelly FP (1989) Fixed point models of loss networks. J Aust Math Soc Ser B: Appl Math 31(2):204–218
33. Kim S, Whitt W (2014) Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? Manuf Serv Oper Manag 16(3):464–480
34. Kim SC, Horowitz I (2002) Scheduling hospital services: the efficacy of elective-surgery quotas. Omega 30(5):335–346
35. Kim SC, Horowitz I, Young KK, Buckley TA (1999) Analysis of capacity management of the intensive care unit in a hospital. Eur J Oper Res 115(1):36–46
36. Kleinrock L (1976) Queueing systems, vol 2. Wiley
37. Krieger UR (1988) Analysis of a loss system with mutual overflow. In: Proceedings of the ITC-seminar. Peking
38. Krupp RS (1982) Stabilization of alternate routing networks. In: Proceedings of the IEEE international conference on communications. Philadelpha
39. Lakshmi C, Iyer SA (2013) Application of queueing theory in health care: a literature review. Oper Res Health Care 2(1–2):25–39
40. Lee N, Hui D, Wu A, Chan P, Cameron P, Joynt GM, Ahuja A, Yung MY, Leung CB, To KF, Lui SF, Szeto CC, Chung S, Sung JJY (2003) A major outbreak of severe acute respiratory syndrome in Hong Kong. Engl J Med 348(20):1986–1994
41. Litvak N, Van Rijsbergen M, Boucherie RJ, Van Houdenhoven M (2008) Managing the overflow of intensive care patients. Eur J Oper Res 185(3):998–1010
42. Long MF, Feldstein PJ (1967) Economics of hospital systems: peak loads and regional coordination. Amer Econ Rev 57(2):119–129
43. Lowery JC (1993) Multi-hospital validation of critical care simulation model. In: Proceedings of the 25th conference on winter simulation, pp 1207–1215
44. McManus ML, Long MC, Cooper A, Litvak E (2004) Queuing theory accurately models the need for critical care resources. Anesthesiology 100(5):1271–1276
45. Newell DJ (1954) Provision of emergency beds in hospitals. Br J Prev Soc Med 8(2):77–80
46. Norris JR (1997) Markov Chains. No. 2 in Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge
47. Papi M, Pontecorvi L, Setola R (2016) A new model for the length of stay of hospital patients. Health Care Manag Sci 19(1):58–65
48. Parker RD (1968) Variation of the occupancy of two medical units with the amount of sharing between the units. Health Serv Res 3(3):214–223
49. Preater J (2002) Queues in health. Health Care Manag Sci 5(4):283
50. Qiu S, Chinnam RB, Murat A, Batarse B, Neemuchwala H, Jordan W (2015) A cost sensitive inpatient bed reservation approach to reduce emergency department boarding times. Health Care Manag Sci 18(1):67–85

51. Rajaratnam M, Takawira F (2000) Nonclassical traffic modeling and performance analysis of cellular mobile networks with and without channel reservation. IEEE Trans Vehicular Tech 49(3):817–834
52. The Association of Anaesthetists of Great Britain and Ireland (2009) AAGBI safety guidline: interhospital transfer. London, http://www.aagbi.org/sites/default/files/interhospital09.pdf
53. Town JA, Churpek MM, Yuen TC, Huber MT, Kress JP, Edelson DP (2014) Relationship between ICU bed availability, ICU readmission, and cardiac arrest in the general wards. Crit Care Med 42(9):2037–2041
54. Whitt W (1985) Blocking when service is required from several facilities simultaneously. AT&T Tech J 64(8):1807–1856
55. Wilkinson RI (1956) Theories for toll traffic engineering in the U.S.A. Bell Syst Tech J 35(2):421–514
56. Wong EWM, Zalesky A, Rosberg Z, Zukerman M (2007) A new method for approximating blocking probability in overflow loss networks. Comput Netw 51(11):2958–2975
57. Wong EWM, Guo J, Moran B, Zukerman M (2013) Information exchange surrogates for approximation of blocking probabilities in overflow loss systems. In: Proceedings of the 25th international teletraffic congress (ITC 25)
58. World Health Organization (2004) Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. Online, http://www.who.int/csr/sars/country/table2004_04_21/en/
59. Wu J, Wong EWM, Zukerman M (2017) Performance analysis of green cellular networks with selective base-station sleeping. Perform Eval 111:17–36