

Design, Performance Evaluation and Optimization for Intensive Care Networks based on Non-Hierarchical Overflow Loss Systems

Jingjin Wu, *Member, IEEE*, Wei Tian, *Student Member, IEEE*, Anqi Wang, Yin-Chi Chan, *Member, IEEE*, Eric W. M. Wong, *Senior Member, IEEE*, Kenny Chan, and Gavin Joynt

Abstract—In design and optimization of intensive care unit (ICU) networks, one common practice is to prioritize the treatment for patients of higher emergency levels, while ensuring fairness to other patients by guaranteeing a certain Quality of Service (QoS) level. One common approach to realize such priority arrangement is bed reservation policy, which designates a certain number of last occupied beds in each hospital to be exclusively used by certain patient classes. In this paper, we propose an approach that can significantly improve the computational efficiency in obtaining the optimal reservation thresholds for each patient class given their respective requirements, in a non-hierarchical ICU model (where the external emergency patients can possibly be allocated to any ICU hospital) which has been shown to be computationally challenging in performance evaluation and optimization. Specifically, we apply the Information Exchange Surrogate Approximation (IESA) to analytically approximate the key QoS metrics under given reservation thresholds, and the Integer Particle Swarm Optimization (PSO) algorithm to search for the optimal threshold based on the approximation results by IESA. We demonstrate numerically, with the real data from ICUs in Hong Kong, that IESA approximation can obtain reasonably accurate results for QoS metrics, and thus lead to accurate optimal reservation thresholds. In addition, our proposed approach combining IESA and PSO can significantly reduce the computation time by more than four orders of magnitude, compared to the state-of-the-art evaluation and optimization methods in existing research for similar problems, especially for ICU networks with practical sizes.

Index Terms—Intensive care units; overflow loss model; approximation; optimization; computational efficiency

I. INTRODUCTION

Intensive care units (ICUs) are well-known in the healthcare industry for their high cost, with the daily staffing cost for an ICU bed being up to six times that for a general ward bed [1]. This high cost contributes to the limited availability of ICU beds all over the world. In major urban areas with dense population like Hong Kong, a particular challenge is to utilize the relatively limited supply to meet the highly variant demands under various scenarios. That is, effective and equitable utilization of these scarce and valuable ICU resources becomes crucial.

Meanwhile, despite the need for efficient and fair treatment of all patients, there is an observed trend in many hospitals where local elective patients are often prioritized over incoming external ICU patients, leading to a disproportionately high rejection rate of the latter [2]. While this practice can be justified for logistic and operational reasons, it contradicts a societal consensus that all patients should be treated fairly, as it cannot provide a service guarantee to external patients. Therefore, one of the key challenges in ICU resource allocation studies is to balance efficiency with fairness in the management of ICU admissions.

To meet these goals, we need to first divide patients into appropriate classes by capturing the characteristics of different patients. The classification can be based on the nature of specific diseases, as well as patients' medical conditions and needs. Existing examples of patient classification include a four-class model in [3] and three-class models in [2] and [4].

queueing-theory models with single or multiple server groups (with each server group representing one hospital) have been commonly used to analyze, evaluate, or optimize the operations of ICU hospitals or networks [5], [6]. As these models typically require arrival processes and length-of-service (LoS) distributions to be known in order to calculate key performance metrics such as patient rejection rates, it is also important to examine real patient data and match appropriate distributions to the patient traffic flows.

For models considering multiple patient classes, data collection and analysis are usually required for each class individually. Given the statistical characteristics of each patient class,

This work was supported by the Health and Medical Research Fund of Hong Kong (16171921), the Research Grants Council (RGC) of Hong Kong under the General Research Fund (11104620, 11102421, and 11101422), the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, Project code 2022B1212010006, and Guangdong Higher Education Upgrading Plan (2021-2025) UIC R0400001-22. Corresponding author: E. W. M. Wong.

J. Wu, W. Tian, and A. Wang are with the Guangdong Key Laboratory IRADS, Department of Statistics and Data Science, BNU-HKBU United International College, Zhuhai, P.R. China (e-mail: jj.wu@ieee.org, s230202702@mail.uic.edu.cn, s230202603@mail.uic.edu.cn).

Y. C. Chan is with the Institute for Manufacturing, University of Cambridge, Cambridge CB3 0FS, UK (e-mail: ycc39@cam.ac.uk).

E. W. M. Wong is with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China (e-mail: eee-wong@cityu.edu.hk).

K. Chan is with the Department of Intensive Care, Tuen Mun Hospital, Hong Kong SAR, China (e-mail: chankck@ha.org.hk).

G. Joynt is with the Department of Anaesthesia and Intensive Care, Chinese University of Hong Kong, Hong Kong SAR, China (e-mail: gavinmjoynt@cuhk.edu.hk).

one then can design appropriate policies that differentiate the service among classes, in order to achieve goals related to efficiency and/or fairness. In this paper, we will focus on a family of policies where the last few unoccupied beds in each ICU are reserved for certain types of patients, as in [4]. This approach has been well justified in existing studies for practical considerations [4], [7], [8]. For example, patients that develop life-threatening medical conditions during their hospital stay (internal emergencies) should have higher priority than other patients, whereas critical-care patients originating from outside the hospital (external emergencies, e.g., vehicular accident victims) may be directly admitted to another nearby hospital instead, and patients scheduled for complex but non-urgent surgery (electives) may have their surgeries deferred. This ensures continuity of care for patients already admitted to the hospital and minimizes unnecessary transfers. Other reasons for requiring internal emergency and elective ICU patients be admitted at the hospital at which they originate include ethical and logistical ones.

In contrast, the ability of external emergency patients to be admitted among more than one ICU provides flexibility to the ICU network under our proposed policies. Specifically, when the first hospital contacted (typically the nearest) has no available beds to accommodate an external emergency patient, alternative hospital(s) can be sought for possible admission. This feature fits a classical multi-server queueing/teletraffic theory model, called the *overflow loss system* (OLS) model, where requests (patients) are allowed to attempt secondary server groups (hospitals) in the case that their primary server group is unavailable. As demonstrated in Fig. 1, OLSs allow for better load balancing across the ICU network and optimize resource usage across the ICU network compared to conventional Erlang loss models without overflow. The “overflow” mechanism reduces the loss of care opportunities for external emergency patients by giving them the opportunity to seek admission to more ICUs in the network. The model is not only closer to actual ICU operations where external emergency patients can seek admissions to all hospitals in the network, but also has enhanced flexibility to adapt to more dynamic patient distribution and changing circumstances such as sudden spikes in demand or varying levels of resource availability in different ICUs.

A. Comparison of hierarchical and non-hierarchical overflow loss systems

We focus on a specific class of OLS model called the *non-hierarchical overflow loss system* (NH-OLS). That is, all hospitals can be served as both primary and secondary server groups to certain patients. In other words, all hospitals in the network can receive both initial and overflow patient traffic.

One key difference between hierarchical overflow loss systems and NH-OLSs is that the former stratifies server groups into tiers, where overflows only occur from lower to upper tiers [2]. On the contrary, the latter allows *mutual overflow* between any pairs of server groups in both directions. In ICU networks, hierarchical overflow is more appropriate for the (*inter-hospital*) *patient transfer* scenario [9], where a patient

who is *already* admitted to a hospital bed is transferred to another hospital (thus vacating the original bed and in turn occupying a bed in the receiving hospital) due to clinical reasons. Such transfers are usually uni-directional or hierarchical and performed so that the patient can receive better treatment at the receiving hospital, which is considered to be able to provide more advanced medical facilities and/or specialized care relative to the originating hospital. On the other hand, the overflow mechanism for external emergency patients in this paper is invoked when an accident occurs *outside* the hospital, and involves a *centralized* coordinator for ICU capacity management and load balancing. If this coordinator finds that the preferred hospital (e.g., the one closest to the accident scene) does not have an available bed upon request, it will check each remaining hospital in order of preference until an available bed is found, upon which the patient is transported *directly* to that hospital, without the need to physically visit any of the preceding hospitals on the preference list. Under this policy, overflow may occur between any two hospitals in either direction, i.e., the system is non-hierarchical. Fig. 2 illustrates the comparison between hierarchical transfer and non-hierarchical overflow situations.

As pointed out in existing studies [4], [10]–[12], the NH-OLS model enables better resource sharing among different hospitals in ICU networks than the hierarchical OLS model, especially when demands are bursty and/or the resources are unbalanced across hospitals. Note that this feature is also applicable in emerging applications where central management of heterogeneous resources is beneficial, such as Video-on-Demand [13] and industrial Internet of Things [14].

However, NH-OLSs are known for their significantly higher complexity in evaluating performance metrics compared to hierarchical systems, especially when the system scale is relatively large. Specifically, the exponentially increasing number of system states and the interdependencies among them make exact analytical evaluation infeasible for all but the smallest and simplest NH-OLSs, even when the evolution of system state can be expressed as a Markov process. This has been the key obstacle preventing researchers from applying NH-OLSs in performance evaluation and optimization in such systems despite their advantages.

Discrete-Event Simulation (DES) has been adopted as an alternative tool for performance evaluation in OLSs [10], [15], [16]. A major advantage of DES is its generality, as no restrictions are required for the arrival process or service-time distributions, unlike methods based on Markov processes, e.g., Markov-chain simulation. On the other hand, although DES remains a versatile and popular approach, its high running time may constitute a bottleneck in evaluating the local or system blocking probability (patient rejection rate in the context of this paper), especially when this probability is low [16], or in optimization applications where a considerable number of evaluations are required to be carried out [11]. In this paper, we aim to determine the optimal reservation thresholds (an optimization problem) in an ICU network (modeled as a multi-class OLS) where rejection rates are expected to be relatively low. Therefore, more efficient and robust evaluation approaches are preferred.

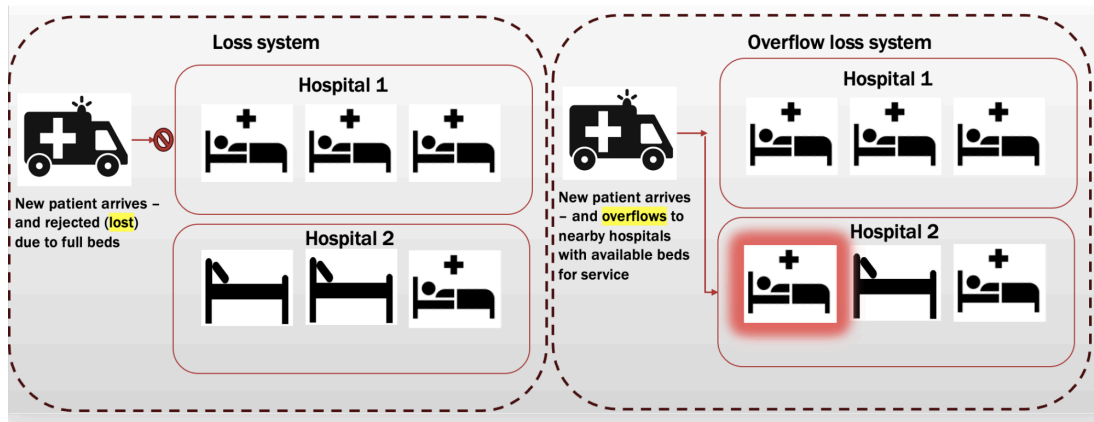


Fig. 1. A comparison of the loss model and the OLS model in ICU.

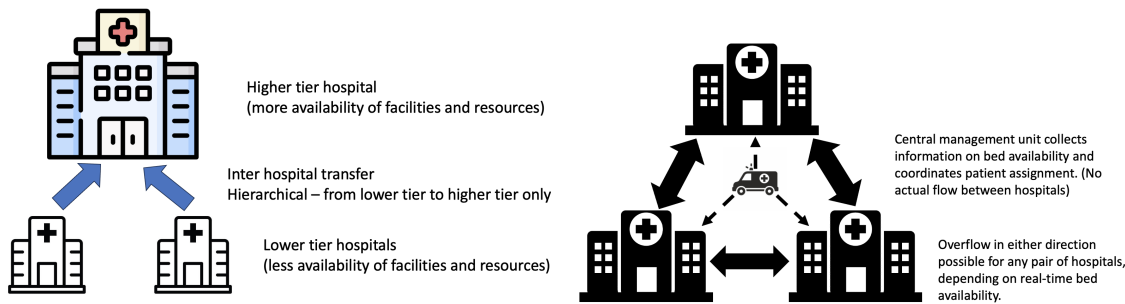


Fig. 2. Graphical comparison between (left) hierarchical overflow in inter-hospital transfer and (right) non-hierarchical overflow of external emergency patients.

B. Contributions of this paper

In this paper, we aim to fill the gap in existing research by proposing a computationally efficient, robust, and scalable approach to overcome the above-mentioned difficulties in design, evaluation, and optimization of ICU networks based on OLS models. We will first extend the work of [4] by analyzing real data from several hospitals in the Hong Kong ICU network, focusing on the dynamics of patient arrivals in order to identify appropriate arrival processes to fit the data such that we can utilize the available tools for further evaluation and optimization. Note that in [4], Poisson arrivals and exponential LoS distributions were assumed without analyzing real data.

Then, we apply an efficient evaluation and optimization approach to minimize the overall rejection and/or deferral rates in the system, subject to certain performance requirements by individual patient classes. Our proposed approach integrates the Information Exchange Surrogate Approximation (IESA) framework with Particle Swarm Optimization (PSO) to establish an integrated approach that can efficiently determine the optimal bed reservation policy for certain classes of ICU patients. Compared with existing similar studies (e.g., [17]–[19]) that use DES for evaluation, our approach significantly reduces the average time for each evaluation, thus improving the overall computational efficiency.

We will also demonstrate the effectiveness of our approach with real data collected from ICU networks in Hong Kong, to verify the applicability of our approach on actual operations of ICU networks. Whereas most results using the IESA frame-

work (e.g., [4], [16], [20]–[22]) are based on an assumption that all LoS distributions are exponential, in this paper, we discuss the effect of different LoS distributions, especially those with higher variance and peakedness, on the evaluation and optimization results.

Compared with existing studies, our novel contributions in this paper can be summarized as follows:

- We demonstrate, using real data from Hong Kong ICUs, that, although the actual situation can be highly dynamic and fluid, they can still be fit using fundamental probability distributions (e.g., exponential, normal, or lognormal) and random processes (e.g., homogeneous Poisson Process or Interrupted Poisson Process (IPP)) for performance evaluation and optimization purposes. This finding can justify the application of our evaluation and optimization tools in real situations.
- We apply and compare different tools, including DES and IESA, in terms of their accuracy and efficiency in evaluating the rejection rate for different classes of patients in OLS-model-based ICU networks under different bed reservation policies. We observe that IESA achieves the best balance in terms of approximation accuracy and computational efficiency. This is consistent with existing studies on ICU networks [4] and other OLSs [22], showing the versatility of IESA in performance evaluations in various applications with different challenges.
- We formulate the problem of identifying the optimal bed reservation policy as a non-linear integer programm

TABLE I
SUMMARY OF ACRONYMS

Acronym	Definition
DES	Discrete Event Simulation
ICU	Intensive Care Unit
IESA	Information Exchange Surrogate Approximation
IPP	Interrupted Poisson Process
LoS	Length of Service
MCS	Markov Chain Simulation
(NH-)OLS	(Non Hierarchical) Overflow Loss System
NLIP	Non-Linear Integer Program
PSO	Particle Swarm Optimization
QoS	Quality of Service

(NLIP). Although similar problem formulations can be found in existing studies, in this paper we take a more efficient approach, using IESA rather than DES to evaluate the rejection rates of all patient classes under a particular bed reservation policy. PSO is then used to search for the optimal policy as evaluated using IESA.

- We demonstrate, using extensive numerical results, that compared with mainstream approaches that use simulation for performance evaluation to identify the optimal reservation policies, our proposed *IESA-based* approach can identify the same optimal policy (or one with similar performance) with greatly improved computational efficiency. The benefit of our proposed IESA-based optimization approach becomes even more obvious as the number of beds in each hospital and/or the number of hospitals increase, leading to very high computational time provided by DES. Therefore, our proposed approach has potential in significantly improving the efficiency in design and planning of city or state wide ICU networks.

A summary of key acronyms used in this paper is presented in Table I.

II. RELATED WORK

A. Bed reservation and optimization in ICU networks

Bed reservation is a commonly adopted resource allocation policy in ICU networks to preserve priority to certain class of patients, especially during times when the demand for ICU beds is high [7], [23], [24]. Many studies have been conducted within the past few years based on the data from the COVID-19 outbreak (e.g., [25]–[27]), while some of them also jointly considered the case where the original or default distribution of medical resources may also not match the demand [28].

Concerns about fairness in ICU priority treatment policies have grown recently, especially since the COVID-19 pandemic [29]. Specifically, Dijkstra *et al.* [30] demonstrated that increasing ICU bed reservations for autonomous same-day admissions benefits these patients, but simultaneously increases the number of required re-allocations of other patients to external regions.

Existing studies mainly made decisions on reserving beds for certain class of patients, based on the accurate prediction of patient demand and/or the division of patients into classes [25], [30], [31]. There have also been studies on the decision of transferring patients to another hospital based on the assessed

risk [32]. Although traditional statistical or machine learning-based models are adept at predicting the *average* demand for ICU beds from different classes of patients, relying solely on these predicted means for bed reservation or resource allocation can be inadequate, particularly in scenarios where demand is subject to high variability [33]. This is especially true during extraordinary events such as pandemics, where the demand for ICU resources can fluctuate drastically and unpredictably. Models that focus only on average demand may not adequately account for sudden surges or declines in patient numbers, leading to either a shortage or under-utilization of critical resources [34].

Our work in this paper, from the model formulation perspective, is closer to that in [2] and [4], where dynamics of the patient flow beyond the mean are taken into account. This approach ensures that the healthcare system remains robust and flexible enough to handle unexpected increases in ICU requirements, while also maintaining efficient operation during periods of lower demand.

B. Performance evaluation of non-hierarchical OLSs

While the non-hierarchical OLS forms a good model for the ICU network considered in this paper, it also introduces the well-known “curse of dimensionality” problem, namely, that accurate and efficient performance evaluation in such systems is extremely difficult. Traditional Markov-Chain evaluation has been shown to be intractable, as the state space grows exponentially with the number of server groups (i.e., hospitals in the current context) [11]. Meanwhile, due to the non-hierarchical nature of the model, with “mutual overflow” of patients among server groups, key assumptions of the classical Erlang Fixed Point Approximation (EFPA) [35] do not hold. Therefore, although EFPA was once considered the state-of-the-art technique for approximation in systems with multiple closed queues, it cannot give accurate results in this case.

In contrast, a number of existing studies (e.g., [10], [15]) relied on simulations to evaluate key performance metrics in ICU networks. Although simulation can give accurate evaluations, the approach is not scalable for optimization in large-scale systems where a large number of measurements are needed to check feasibility and optimality. Discussions on the accuracy and efficiency of MCS, DES, and IESA-based methods in different systems can be found in [21], [22], [36].

The IESA framework [11], [20] was proposed to fix the approximation errors in EFPA. While the effectiveness and efficiency of IESA have been demonstrated in evaluating blocking probabilities in ICU networks [4] and other overflow loss models such as mobile networks [22], [36], there are less studies utilizing this powerful tool in optimization problems. An initial attempt was presented in [4], where IESA (along with EFPA) was used to evaluate patient rejection rates in small-scale ICU networks with up to 5 hospitals.

C. Heuristic algorithms for optimization

Although IESA-based optimization was previously applied to bed reservation in a small cluster of three ICUs [4], the exhaustive search method used therein, which evaluates the

optimality and feasibility of every point in the search space, is not feasible for large-scale ICU networks with more hospitals and/or resources, especially in time-critical situations such as emergencies or pandemics when the demand often exceeds supply. To improve search efficiency in systems with relatively larger sizes, a number of heuristic optimization algorithms have been proposed.

In this paper, we focus on one such heuristic, namely Particle Swarm Optimization (PSO), which was originally derived from basic mechanisms of bird flocking and fish schooling [37]. PSO has been successfully applied in various scenarios including control problems in ICUs [38]. More recently, variants of the PSO have been proposed to accommodate the features of specific problems [39]. One variant that suits the problem in this paper is PSO for integer programming [40], which focuses on situations where one or more decision variables are discrete, by modifying the velocity and position-updating procedures from the original PSO. In fact, it is common for optimization problems in healthcare-related applications to involve integer decision variables. Recent studies (e.g., [41], [42]) have also shown that PSO-based algorithms can improve the efficiency in solving large scale NLIPs or MNLIPs (Mixed Nonlinear Integer Programs) by reducing the number of required evaluations compared to exhaustive search. In addition, Gupta *et al.* [43] demonstrated that the incorporation of an artificial neural-inspired whale optimization algorithm would enhance computational efficiency and resource utilization in e-health applications. Chen and Zeng [44] also illustrated that heuristic algorithms combining the decision tree method and PSO, would achieve superior solutions with higher efficiency.

In Table II, we briefly summarize related studies on resource allocation policies that accounted for the stochastic variability in customer (patient) demands with queueing models. Among these, our work is the first that applies an NH-OLS system model for bed reservation with analytical approximation methods combined with heuristic optimization to address resource allocation in this specific context.

III. MODEL

A. ICU network

We consider an ICU network consisting of N ICU hospitals, with all hospitals “connected” to each other. That is, patients originally destined for a certain hospital can directly overflow to any other hospital in the case that the originally destined hospital is operating at full capacity. This is analogous to the concept of *fully-connected network* in network science and similar to the network model originally proposed in [12] for burn wards in New York State (however, this model did not contain multiple patient types like our model). We denote the set of hospitals by $\mathcal{N} = \{0, 1, 2, \dots, N-1\}$, where hospital $i \in \mathcal{N}$ has b_i beds.

Similar to [2], we consider three types of patients, namely internal emergencies, external emergencies, and electives. We consider *reservation thresholds* for internal emergencies, external emergencies, and electives at each hospital $i \in \mathcal{N}$, denoted by non-negative integers r_i^{in} , r_i^{ex} , and r_i^{el} . That means,

an incoming internal emergency patient will be rejected for admission from hospital i if the number of available beds at its arrival is less than last r_i^{in} . Similarly, an incoming elective patient to hospital i will be deferred for treatment if the number of available beds at its arrival is less than last r_i^{el} . We denote $\mathbf{r}^{\text{in}} = \{r_0^{\text{in}}, r_1^{\text{in}}, \dots, r_{N-1}^{\text{in}}\}$, $\mathbf{r}^{\text{ex}} = \{r_0^{\text{ex}}, r_1^{\text{ex}}, \dots, r_{N-1}^{\text{ex}}\}$ and $\mathbf{r}^{\text{el}} = \{r_0^{\text{el}}, r_1^{\text{el}}, \dots, r_{N-1}^{\text{el}}\}$.

By definition, internal emergency and elective patients would only seek bed spaces in their own hospitals, while external emergency patients are allowed to seek admissions to all hospitals in the network. Different from [4], we do not consider the deployment of overbeds which can accommodate extra internal emergency patients, in accordance with the actual situation in Hong Kong.

If an external emergency patient is rejected at a certain hospital due to a lack of bed spaces, it will then seek admission at another hospital that it has not sought before, until it successfully finds an available bed or is rejected by all hospitals in the network. For convenience, we define $\Gamma_i = (\gamma_{i,0}, \gamma_{i,1}, \dots, \gamma_{i,N-1})$ as the overflow sequence of a patient seeking initial admission at hospital i . In particular, we consider a special mechanism where $\gamma_{i,0} = i$, and $\gamma_{i,k} = (i+k) \% N$ for all $k > 0$. Here, $a \% b$ represents the modulo operation that returns the remainder of dividing a by b . While IESA was originally proposed for approximating the performance of OLSs under random routing, this *Round Robin* routing rule has been used in past IESA work such as [20], [22], [36] as an alternative to the random routing rule to make the implementation computationally feasible. Also note that this Round Robin rule only applies to the IESA implementation, but not the real ICU operations, which intuitively adopt sequential routing based on distances and/or traffic conditions between hospitals. We will demonstrate that the Round Robin-based IESA can achieve a high level of accuracy compared to sequential routing-based simulation results numerically in the results section.

We further define the following Quality of Service (QoS) measurements for the patients:

- R^{I} : the proportion of internal emergency patients that are rejected from intensive care due to lack of bed spaces in their own hospital.
- R^{E} : the proportion of external emergency patients that are rejected from intensive care due to lack of bed spaces in all hospitals.
- D : the proportion of elective patients that have their scheduled surgeries deferred due to lack of bed spaces in the scheduled hospital.

A graphical illustration of the three patient classes, bed reservation policy, and the overflow mechanism in our model is depicted in Fig. 3. A summary of notations defined in this subsection is provided in Table III.

B. Dataset description

We collected data from three public ICU hospitals, including North District Hospital (NDH), Alice Ho Miu Ling Nethersole Hospital (AHNH), and Pok Oi Hospital (POH) in New Territories, Hong Kong. We refer the three hospitals as Hospital A,

TABLE II
SUMMARY OF RELATED STUDIES

Research	Resource allocation policy	System model	Evaluation method	Optimization method
[3]	Admission/discharge control	Single loss queue	Exact analytical solution	N/A
[4]	Bed reservation	NH-OLS	Analytical approximation	Exhaustive search
[6]	Triage and patient transfer	Network of delay queues	Exact analytical solution	Heuristic
[7]	Admission/discharge control	Single loss queue	Exact analytical solution	Exhaustive search
[8]	Bed reservation	Hierarchical OLS	Simulation (large scale cases)	Heuristic
[10]	Hospital bed allocation	Independent queues	Simulation	Heuristic
[18]	Patient assignment	Hierarchical OLS	Simulation	Exhaustive search
[19]	Admission/discharge control	Hierarchical OLS	Simulation	N/A
[22]	Mobile channel borrowing	NH-OLS	Analytical approximation	N/A
[27]	Admission control	Single loss queue	Heuristics	N/A
[31]	Bed reservation	Single delay queue	Exact analytical solution	N/A
[36]	Mobile base station sleeping	NH-OLS	Analytical approximation	N/A
This work	Bed reservation	NH-OLS	Analytical approximation	Heuristic

TABLE III
KEY NOTATIONS USED IN THIS PAPER

Symbol	Definition
$\mathcal{N}$	The set of hospitals in the ICU network, enumerated as $\{0, 1, 2, \dots, N-1\}$.
b_i	The number of ICU beds at hospital i .
$r_i^{in}, r_i^{ex}, r_i^{el}$	The number of reserved beds that cannot be used by internal emergency/external emergency/elective patients.
$\mathbf{r}_i^{in}, \mathbf{r}_i^{ex}, \mathbf{r}_i^{el}$	The sets $\{r_i^{in} \mid i \in \mathcal{N}\}$, $\{r_i^{ex} \mid i \in \mathcal{N}\}$, $\{r_i^{el} \mid i \in \mathcal{N}\}$, of internal emergency/external emergency/elective reservation thresholds.
Γ_i	The sequence $(\gamma_{i,0}, \gamma_{i,1}, \dots, \gamma_{i,N-1})$, where $\gamma_{i,k}$ is the chosen hospital for external emergency patients originally destined for hospital i that have overflowed k times, i.e., rejected from k other hospitals.
R^I, R^E	Rejection rate of internal/external emergency patients.
D	Deferral rate of elective patients.

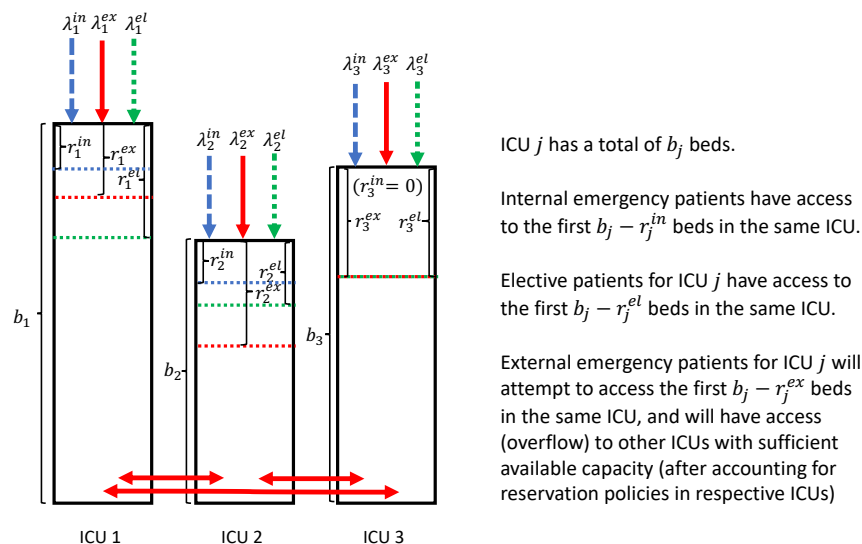


Fig. 3. An illustration of patient arrivals and bed reservation policy in the ICU network.

B, and C (not necessarily corresponding to the same order of hospitals as in the previous sentence) in the rest of the paper, to preserve the confidential characteristics of the hospitals. The dataset contains information including (but not limited to) referral source, patient age, arrival time, assessed urgency, and treatment outcome for patients admitted to ICU in these three hospitals from 2016 to 2017. There are in total 1429, 385, and 1991 record entries from Hospital A, B, and C, respectively.

To match our patient classifications, we consider that

- All patients classified as “non-urgent” in the dataset are considered as **elective** patients.
- Patients classified as “urgent” with a referral source of “Ward” or “OT” (outpatient) of the same hospital are considered as **internal emergency** patients.
- All other patients (e.g., “urgent” patients from “Accident

and Emergency Department” or “other hospitals”) are considered as **external emergency** patients.

C. Optimization problem

In general, we aim to minimize the overall probability C , that a patient cannot be treated due to insufficient ICU beds, by adjusting the reservation thresholds $r_j^{\text{in}}, r_j^{\text{ex}}$ and r_j^{el} based on network parameters including arrival rates and the total number of beds available at each hospital. The relative priority of different patient classes can be adjusted by introducing weighting parameters. Specifically, the optimization problem can be formulated as follows,

$$\begin{aligned} \min_{\mathbf{r}^{\text{in}}, \mathbf{r}^{\text{ex}}, \mathbf{r}^{\text{el}}} C &= w_1 R^{\text{I}} + w_2 R^{\text{E}} + w_3 D \\ \text{s.t. } R^{\text{I}} &\leq R_{\text{req}}^{\text{I}}, \\ R^{\text{E}} &\leq R_{\text{req}}^{\text{E}}, \\ D &\leq D_{\text{req}}, \\ r_j^{\text{in}} &\in \{0, 1, \dots, r_{\text{max}}^{\text{in}}\}, \quad \forall j \in \mathcal{N} \\ r_j^{\text{ex}} &\in \{0, 1, \dots, r_{\text{max}}^{\text{ex}}\}, \quad \forall j \in \mathcal{N} \\ r_j^{\text{el}} &\in \{0, 1, \dots, r_{\text{max}}^{\text{el}}\}, \quad \forall j \in \mathcal{N} \\ w_1 + w_2 + w_3 &= 1, \end{aligned} \quad (1)$$

where $R_{\text{req}}^{\text{I}}, R_{\text{req}}^{\text{E}}$, and D_{req} represent the required thresholds that $R^{\text{I}}, R^{\text{E}}$ and D cannot exceed, respectively. Also, w_1, w_2 and w_3 represent relative weights on the rejection rates which can be adjusted by hospital management depending on specific situations and requirements. If w_1, w_2 and w_3 are set to be the arrival rates of the corresponding patient kinds, the formulation provides the minimum solution for the overall rejection rate. Note that this formulation can also accommodate the situations where the rejection/deferral rates for only one or two patient classes are considered, by setting the weights on the irrelevant class to 0. In addition, we consider that the reservation thresholds for internal emergencies, external emergencies, and elective patients at all hospitals cannot exceed $r_{\text{max}}^{\text{in}}, r_{\text{max}}^{\text{ex}}$, and $r_{\text{max}}^{\text{el}}$, respectively.

Current approaches that attempt to solve (1) (or its related variants) suffer from two key issues that prevent them from being feasible in real ICU networks of practical sizes.

The first issue is **the computational infeasibility in evaluating $R^{\text{I}}, R^{\text{E}}$, and D in NH-OLSS** for a given combination of reservation thresholds $\mathbf{r}^{\text{in}}, \mathbf{r}^{\text{ex}}$, and $\mathbf{r}^{\text{el}}$. Specifically, obtaining the exact rejection rates for an ICU network with N hospitals and b beds in each hospital involves solving for the steady-state probabilities of an N -dimensional Markov chain, which has a computational complexity of $O(b^{3N})$ [45]. This exponential complexity renders exact computation infeasible even for moderate values of N and b .

The second issue is **the NP-hardness of (1)**. The nonlinear relationships between the reservation thresholds $\mathbf{r}^{\text{in}}, \mathbf{r}^{\text{ex}}, \mathbf{r}^{\text{el}}$ and the rejection/deferral rates $R^{\text{I}}, R^{\text{E}}$, and D make (1) an NLIP problem. As a result, the exhaustive search method commonly used in current ICU studies requires a total of $(r_{\text{max}}^{\text{in}} \times r_{\text{max}}^{\text{ex}} \times r_{\text{max}}^{\text{el}})^N$ evaluations to identify the optimal feasible solution of (1).

While most studies have resorted to simulation-based methods to partially address the first issue by avoiding exact analytical computations, this approach remains undesirable for optimization in moderate to large-scale ICU networks. The enormous number of evaluations required due to the NP-hardness and the time-consuming nature of simulations pose significant challenges to obtaining the optimal solution efficiently (see our numerical results in Section VII for more information). We will describe the methods that we propose to address these two issues in the next two sections.

IV. COMPUTATIONALLY EFFICIENT EVALUATIONS OF REJECTION RATES

IESA [20] is a decomposition-based approach that has been demonstrated to achieve a reasonable level accuracy in blocking probability estimation in OLSSs. IESA considers hierarchical surrogate system, where each request has two specifically designed attributes, denoted Δ (a set of the server groups/hospitals that the request has overflowed from) and Ω (estimation busy server groups/hospitals based on overflow history), and applies EFPA on the surrogate model.

In the ICU network context, all external emergency patients start with $\Delta = \emptyset$, and $\Omega = 0$. We use the term (Δ, Ω) -patient to denote an external emergency patient who has attempted all hospitals in Δ and has a congestion estimate of Ω . In addition, an (Δ, Ω) -patient will be considered immediately rejected without attempting the remaining available hospital with a probability of

$$P_{k,n,j} = \begin{cases} 0, & \text{if } j < N, \\ \frac{\binom{j-N}{k-n}}{\binom{N-n}{k-n}}, & \text{if } j \geq N, \end{cases} \quad (2)$$

where k is a parameter denoting the maximum allowed Ω value of patients. In this paper, as we consider an ICU network model where any external emergency patient has access to all hospitals, we set k as a constant equal to the total number of hospitals as in the original IESA design [20]. Note that in partial availability systems such as mobile networks, an optimal value of k can be estimated, for example, by neural network techniques, to further improve the approximation accuracy (see e.g., [16], [22], [36]).

We briefly describe the updating rule for Ω and Δ after each overflow. Consider a (Δ_1, Ω_1) -patient attempting ICU i . If a non-reserved bed is available, the patient is admitted. Otherwise, we assume that the external emergency patient with the highest Ω residing at the same hospital is a (Δ_2, Ω_2) -patient. If $\Omega_1 < \Omega_2$, the information exchange mechanism is activated and the incoming patient overflows as an $(\Delta_1 \cup \{i\}, \Omega_2 + 1)$ -patient, while the residing patient becomes an (Δ_2, Ω_1) -patient. Otherwise, namely when $\Omega_1 \geq \Omega_2$, the incoming patient is rejected and overflows normally, becoming a $(\Delta_1 \cup \{i\}, \Omega_1 + 1)$ -patient. Note that the updating rule also implicitly guarantees that $\Omega \geq |\Delta|$ for all external emergency patients. In this sense, IESA forms an hierarchical traffic structure based on Ω , where level j of the hierarchy includes all patients with $\Omega \leq j$, and the rejection probability of patients at lower hierarchy is not affected by the existence of patients at higher hierarchies.

We introduce the following notations for a better understanding of IESA evaluation procedures

- $a_{i,n,j}$: the offered traffic of external emergency patients to hospital i composed of (n, j) -patients;
- $e_{i,n,j}$: the overflow traffic of external emergency patients from hospital i composed of (n, j) -patients;
- $\tilde{a}_{i,n,j}$: the offered traffic of external emergency patients to hospital i with $|\Delta| = n$ and $\Omega \leq j$;
- $\tilde{e}_{i,n,j}$: the overflow traffic of external emergency patients from hospital i with $|\Delta| = n$ and $\Omega \leq j$;
- $A_{i,j} = \sum_{n=0}^{N-1} \tilde{a}_{i,n,j}$: the offered traffic of external emergency patients to hospital i with $\Omega \leq j$;
- $b_{i,j}$: the probability that all non-reserved beds at hospital i for external emergency patients are occupied by external emergency patients with $\Omega \leq j$ or other types of patients.

For simplicity without loss of generality, we assume the mean LoS for all patients to all hospitals is 1. We can adjust the respective arrival rates proportionally such that all QoS metrics are not changed.

By definition, we have

$$a_{i,0,j} = \begin{cases} \lambda_{i*}^{\text{ex}}, & j = 0, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where λ_{i*}^{ex} denotes the offered traffic of external emergency patients to hospital i .

If the arrival processes of all patients classes to each hospital are Poisson, we can use a one-dimensional Markov chain representation to describe the states with respect to the number of occupied beds in each hospital. Let state m denote the state in which there are m patients in service, and $q_{m,\ell}$ be the transition rate from state m to state ℓ . Then for hospital i ,

$$\begin{aligned} q_{m,m+1} &= A_{i,k} \mathbb{1}\{m < b_i - r_i^{\text{ex}}\} \\ &\quad + \lambda_{i*}^{\text{in}} \mathbb{1}\{m < b_i - r_i^{\text{in}}\} + \lambda_{i*}^{\text{el}} \mathbb{1}\{m < b_i - r_i^{\text{el}}\}, \\ q_{m,m-1} &= m, \\ q_{m,\ell} &= 0, \quad |m - \ell| \neq 1, \end{aligned} \quad (4)$$

where λ_{i*}^{in} and λ_{i*}^{el} are similarly defined as λ_{i*}^{ex} above.

The transition rates $q(m, \ell)$ can be used to calculate the probabilities of all states (the total number of patients in the hospital) including $0, 1, \dots, b_i$. We denote π_m as the steady state probability of state m . Then, the approximated rejection/deferral rates for internal emergency and elective patients at the network level under IESA are

$$\hat{R}^I = \frac{\sum_{i=0}^{N-1} \left(\lambda_{i*}^{\text{in}} \sum_{m=b_i-r_i^{\text{in}}}^{b_i} \pi_m \right)}{\sum_{i=0}^{N-1} \lambda_{i*}^{\text{in}}}, \quad (5)$$

and

$$\hat{D} = \frac{\sum_{i=0}^{N-1} \left(\lambda_{i*}^{\text{el}} \sum_{m=b_i-r_i^{\text{el}}}^{b_i} \pi_m \right)}{\sum_{i=0}^{N-1} \lambda_{i*}^{\text{el}}}. \quad (6)$$

For external emergency patients, due to the information exchange mechanism in IESA and the hierarchical structure in the surrogate system, we have

$$e_{i,n,j} = e_{i,n-1,j-1} b_{i,j-1} + \tilde{e}_{i,n-1,j-2} (b_{i,j-1} - b_{i,j-2}), \quad (7)$$

where $b_{i,j}$ can be calculated by following a similar manner as (4) to obtain the steady state probabilities, while replacing $A_{i,k}$ with $A_{i,j}$ due to the priority of traffic at lower hierarchies. We use the notation $\pi^{m,j}$ to denote the results corresponding to hierarchy $\Omega = j$. In addition, according to the abandonment mechanism described in (2) and the Round Robin routing rule in IESA, the relationship between overflow traffic from one hospital and traffic that will attempt the next hospital in sequence can be described as

$$a_{\gamma_{i,m+1},n,j} = \sum_{m=b_i-r_i^{\text{ex}}}^{b_i} e_{\gamma_{i,m},n,j} (1 - P_{n,j}). \quad (8)$$

Note that values with negative indices in (7) are by definition zero, and that $P_{n,j} = 1$ if $j = N$ according to (2). For $b_{i,j}$, we obtain

$$b_{i,j} = \sum_{m=b_i-r_i^{\text{ex}}}^{b_i} \pi_{m,j}. \quad (9)$$

The approximated rejection rates of external emergency patients in the network can then be obtained by summing all “abandoned” traffic, that is,

$$\hat{R}^E = \sum_{i=1}^n \sum_{n=1}^N \sum_{j=n}^N e_{i,n,j} P_{n,j}. \quad (10)$$

V. HEURISTIC ALGORITHM

As (1) is an NLIP, we consider a variant of PSO proposed in [40] that incorporates special mechanisms such as discretization of position updates.

We describe the key steps of the PSO-based algorithm used in this paper in Algorithm 1, where X_i represents the position of the i -th particle, corresponding to a vector of decision variables $\mathbf{r}^{\text{in}}$, $\mathbf{r}^{\text{ex}}$ and $\mathbf{r}^{\text{el}}$ from the integer program.

For the stopping criteria, we consider that either the algorithm has reached a pre-determined number of iterations, or that the change in fitness $f(X_i)$ in a certain number of consecutive rounds are less than a pre-determined threshold.

VI. DATA ANALYSIS

As mentioned in Section III-B, our data are collected from three ICU hospitals with over 3400 entries in total. However, we will exclude the records that never attempt for ICU admission before conducting further analysis. Specifically, we will only consider records marked with “Admit to ICU” or “Not admitted to ICU due to tight beds”, but not other status such as “patient refusal” and “dead before ICU arrival”, in order to capture the demand of ICU beds more accurately. After filtering out ineligible records, the numbers of valid records for all patient classes at all hospitals are given in Table IV. Key statistical properties, including mean, variance, skewness, and peakedness (variance-to-mean-ratio) by patient type and

Algorithm 1 PSO-based algorithm for determining reservation thresholds in ICU networks

```

1: Initialize  $n_{\text{particles}}$ ,  $X_i$ , and other algorithmic parameters
2: Evaluate the fitness of each particle at their origin positions (using the objective function in (1) as the fitness function  $f(X)$ )
3: for each particle  $i = 1$  to  $n_{\text{particles}}$  do
4:    $pBest_i \leftarrow X_i$ 
5:   if  $f(pBest_i) < f(gBest)$  or  $gBest$  is undefined then
6:      $gBest \leftarrow pBest_i$ 
7:   end if
8: end for
9: while stopping criteria not met do
10:  for each particle  $i = 1$  to  $n_{\text{particles}}$  do
11:    for each dimension  $j$  do
12:      Update and discretize velocity of particles;
13:      Update position  $X_{ij}$ ;
14:    end for
15:    Evaluate fitness:  $f(X_i)$  considering all constraints;
16:    Update  $pBest_i$  and  $gBest$ , if needed;
17:  end for
18:  Update algorithmic parameters.
19: end while
20: return  $gBest$ 

```

TABLE IV
THE NUMBER OF ELIGIBLE RECORDS

Hospital	Patient Class		
	Internal Emergency	External Emergency	Elective
A	119	50	57
B	420	132	200
C	250	97	143
Total	789	279	400

hospital are summarized in Table V. We can observe that the peakedness (variance-to-mean) ratio of selected patient arrival processes can be quite high, which necessitates consideration of more general arrival processes than the classical Poisson process considered in existing studies using queueing models.

We also examined the annual statistical report complied by the Hong Kong Hospital Authority [46]. As of 31 March 2023, public record show that the ICUs at Hospitals A, B, and C contained 9, 11, and 17 ICU beds, respectively.

The homogeneous Poisson process (commonly referred simply as Poisson Process) is characterized by a constant rate of event occurrence, implying that events are equally likely to occur at any time, leading to an exponential distribution of inter-arrival times. The Poisson process is straightforward and mathematically tractable. Also, for a single queue or queueing network model, if the arrival process is Markovian and service time distribution is exponential, the Markov Chain Simulation (MCS), a more time-efficient approach compared to the DES, can be used to evaluate the key metrics. However, the Poisson process often lacks the flexibility to capture more complex real-world scenarios where the event rate might vary due to external factors. On the other hand, one of the reasons

for using an IPP to approximate or fit a process lies in its connection with the homogeneous Poisson process. By considering the homogeneous Poisson process as a special case of the IPP (where the system is always in the active/“on” state), one can leverage the mathematical simplicity and well-understood properties of the HPP while introducing additional flexibility to model real-world processes more accurately. The benefit of using an IPP is that it can better capture the dynamics of systems where the rate of event occurrence is subject to change, offering a more realistic and adaptable model. This increased adaptability does not come at the cost of losing the foundational principles of the Poisson process, as the IPP essentially builds upon the Poisson process framework by adding a mechanism to modulate the event rate.

Using the two-moment matching methodology in [47], we can construct an IPP to fit the mean, variance, and peakedness of traces of patients to each hospital. The key idea of the method is to consider a single server queue with infinite many servers and exponentially distributed LoS with unit mean, identify the following parameters of an IPP:

- λ : The arrival rate of the IPP in the active state;
- ω : The transition rate of the IPP from the inactive to the active state;
- η : The transition rate of the IPP from the active to the inactive state,

such that the first two moments of the distribution of the number of busy servers in the queue generated from the IPP arrival process match those generated from the real trace.

It has been demonstrated in [21] that although the fitting methodology only directly deal with the first two moments (mean and variance), it can actually also fit the third moment (skewness) quite well in most cases. Our results demonstrated in Table VI are consistent with the observation in [21], where the relative differences of skewness between fitted IPP and actual trace arrivals are all within $\pm 6\%$. The results also confirm that all statistics related to the first two moments, i.e., mean, variance, and peakedness, are fitted well.

The fitted parameters of the IPPs are shown in Table VII. Note that for internal emergency patients at Hospital B and external emergency patients at Hospital C, as the variances are less than the means in the original traces, we use the homogeneous Poisson processes to fit the arrival processes. Therefore, ω and γ are not applicable for them.

On the other hand, for the LoS distributions of all types of patients, we obtain from the Clinical Data Analysis and Reporting System (CDARS) database maintained by the Hong Kong Hospital Authority [48] that the mean LoS for internal emergency, external emergency, and elective patients are 5.492, 4.852, and 1.645 days, respectively. We will demonstrate numerically later in this paper that the performance metrics concerned are not very sensitive to the LoS distributions beyond the mean.

VII. NUMERICAL RESULTS

We now present the numerical results for demonstrating the accuracy and efficiency of our proposed approach. The 95% confidence interval of all simulation results presented in this

TABLE V

STATISTICAL PROPERTIES OF THE ARRIVAL PROCESSES BY PATIENT TYPE AND HOSPITAL BASED ON THE REAL TRACES (UNIT: DAYS)

Hospital	Patient Type	Statistic			
		Mean	Variance	Skewness	Peakedness
A	Int. Emerg.	0.327	0.351	1.998	1.075
	Ext. Emerg.	0.136	0.139	2.858	1.027
	Elective	0.150	0.172	3.038	1.145
B	Int. Emerg.	1.145	1.122	0.887	0.980
	Ext. Emerg.	0.360	0.380	1.790	1.055
	Elective	0.542	0.580	1.458	1.069
C	Int. Emerg.	0.633	0.653	1.313	1.032
	Ext. Emerg.	0.251	0.246	1.975	0.981
	Elective	0.371	0.428	1.863	1.153

TABLE VI

RELATIVE DIFFERENCE OF STATISTICAL PROPERTIES OF THE ARRIVAL PROCESSES BY PATIENT TYPE AND HOSPITAL BY THE REAL TRACES AND FITTED IPP ARRIVALS (UNIT: DAYS)

Hospital	Patient Type	Relative difference			
		Mean	Variance	Peakedness	Skewness
A	Int. Emerg.	0.05%	-0.02%	-0.07%	-4.32%
	Ext. Emerg.	-0.02%	-0.01%	0.01%	-1.51%
	Elective	0.02%	0.03%	0.01%	0.69%
B	Int. Emerg.	-0.03%	1.99%	2.02%	5.46%
	Ext. Emerg.	0.07%	0.11%	0.04%	-0.48%
	Elective	-0.09%	-0.13%	-0.04%	0.63%
C	Int. Emerg.	-0.03%	-0.03%	-0.00%	-0.78%
	Ext. Emerg.	-0.05%	1.90%	1.96%	-1.17%
	Elective	-0.03%	-0.04%	-0.01%	4.12%

section, based on Student's t-distribution, are within $\pm 3\%$ of the observed mean.

A. Small-scale case ($N = 3$) for validating the accuracy of our proposed analytical method.

We first consider a cluster of three ICU hospitals, with 9, 11, and 17 beds, respectively, mirroring the actual setting of the three hospitals. We also set the mean LoS of internal emergency, external emergency, and elective patients as 5.492, 4.852, and 1.645 days, respectively, in accordance with the statistics from the CDARS. External emergency patients from any of the three hospitals are allowed to overflow to either of the other two hospitals, with random overflow sequence. For demonstration purpose, we set the reservation thresholds as $r^{\text{in}} = 0$, $r^{\text{ex}} = 1$, $r^{\text{el}} = 2$ across all hospitals.

As presented in Table V, the maximum peakedness (variance-to-mean) ratio in the real patient traces among all hospitals and classes is 1.153. To demonstrate the effect of a reasonable range of arrival peakedness on the rejection and deferral rates, we first demonstrate the results with peakedness of the IPP arrivals ranging from 1 (homogeneous Poisson arrivals) to 1.2, along with the peakedness value fitted from the real trace as in Section VI. We also consider the impact of LoS distribution, specifically the variance, when the mean is kept constant. According to existing analysis [49], [50], lognormal is an appropriate distribution for describing the LoS

TABLE VII

PARAMETERS OF THE FITTED IPP ARRIVAL PROCESSES

Hospital	Patient Type	Parameter		
		λ	ω	γ
A	Int. Emerg.	0.519	0.971	0.570
	Ext. Emerg.	0.189	0.723	0.286
	Elective	0.516	0.444	1.084
B	Int. Emerg.*	1.144	N/A	N/A
	Ext. Emerg.	0.500	1.109	0.431
	Elective	0.740	1.364	0.498
C	Int. Emerg.	0.726	1.678	0.247
	Ext. Emerg.*	0.251	N/A	N/A
	Elective	0.802	0.837	0.970

* These arrival processes are fitted using homogeneous Poisson processes, as both variances are slightly less than the means in the actual traces.

of patients.

As the real patient traffic were relatively light during the period (leading to internal rejection, external rejection, and deferral rates less than 0.1%, 0.01%, and 5%), we proportionally scale the traffic up to examine scenarios where these rates are more significant and thus bed reservation policies become meaningful. The results are presented in Fig. 4, where the horizontal axis denotes the scale parameter, indicative of the factor by which the real traffic volume is multiplied.

The results show that, under reasonable range of parameters including the arrival peakedness z (where $z = z_{\text{real}}$ means each arrival process follows the peakedness value obtained from the IPP fitting in Tables V and VI) and the variance of the LoS distribution σ , changes in R^I , R^E and D are not very significant given that the means of the patient interarrival times and LoS do not change. Meanwhile, we observe that IESA can obtain quite accurate approximation results of all three measurements. The relative errors of IESA approximation are all within 20% under scenarios where $z = z_{\text{real}}$.

We also consider another case where the total number of beds in each hospital is adjusted. In Fig. 5, the horizontal axis represents the number of beds adjusted with respect to the actual data at the three hospitals. For example, an adjusted number of 1 corresponds to a hypothesized scenario where the hospitals have 10, 12, and 18 ICU beds (+1 bed from the original numbers 9, 11, and 17). The observation is similar to that in Fig. 4, where R^I , R^E , and D are all not very sensitive to the changes in the arrival processes and/or LoS distributions beyond the mean. On the other hand, IESA still gives reasonably accurate approximations in all scenarios, especially for those with $z = z_{\text{real}}$.

B. Large-scale case ($N = 17$) for accuracy validation and demonstrating the benefit of non-hierarchical OLS design

We now consider a larger ICU network with a total of 17 hospitals, which is the same as the total number of public hospitals with ICU in Hong Kong. Due to the lack of real data, we consider the range of the relevant parameters in Hospitals A, B, and C, and uniformly generate the mean arrival rates of all three patient classes from all hospitals in $(0.1, 1)$, the total number of beds in each hospital in $\{9, 10, \dots, 25\}$, and

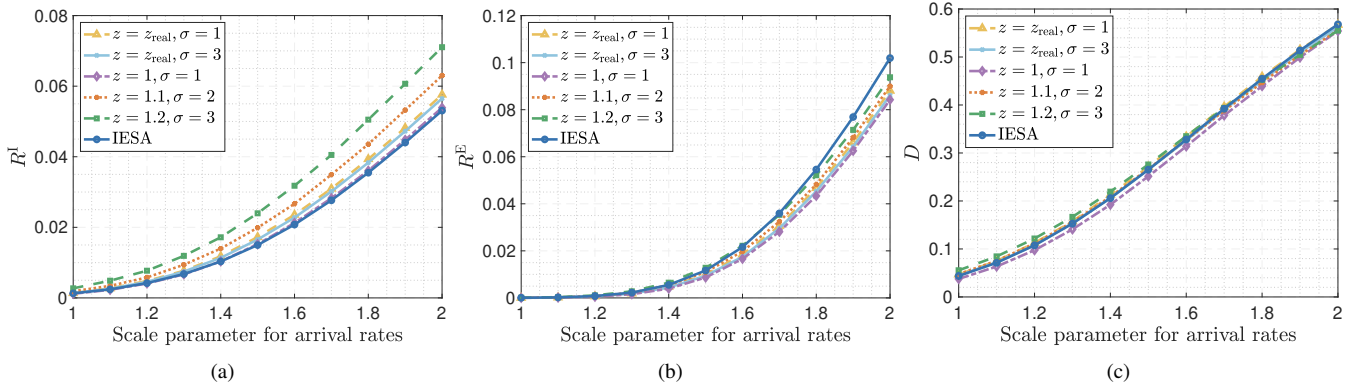


Fig. 4. IESA and simulation results for (a) R^I , (b) R^E , and (c) D , with different arrival processes (z denotes the peakedness), LoS distributions (σ denotes the standard deviation), and changing arrival rates (3 hospitals).

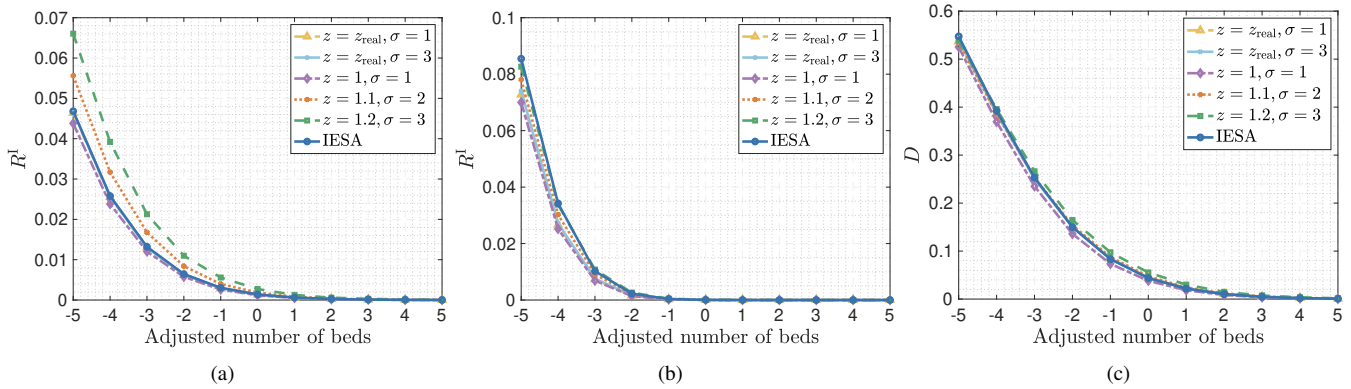


Fig. 5. IESA and simulation results for (a) R^I , (b) R^E , and (c) D , with different arrival processes (z denotes the peakedness), LoS distributions (σ denotes the standard deviation) and adjusted numbers of bed spaces (3 hospitals).

the number of reserved beds for each class in $\{0, 1, 2, 3, 4, 5\}$. The mean LoS for each patient class are kept the same as in the last subsection.

Firstly, as in the small-scale case, we demonstrate the sensitivity of key QoS metrics to the peakedness in arrival processes and LoS distributions beyond the mean, to justify the application of IESA under homogeneous Poisson arrivals and exponential LoS distributions. Note that R^E is extremely close to 0 for all distributions and evaluation methods, as the external emergency patients are allowed to use beds from all 17 hospitals under this setting, and hence are rarely rejected. Therefore, we do not present the results for R^E separately. The results of R^I and D are shown in Fig. 6. Similar to the case with 3 hospitals, IESA is sufficiently accurate in estimating R^I and D , with the deviation consistently less than 20%.

Then, we demonstrate another set of comparisons in terms of the QoS metrics between the non-hierarchical OLS model and a model that only allows each external emergency patient to attempt one hospital (i.e., no overflows to hospitals other than the initially attempted one), assuming Poisson arrivals and exponential LoS distributions for both cases. The results are shown in Fig. 7. It is indicated that by allowing overflows of external emergency patients, the rejection rate of such patients would significantly decrease, at the expense of slightly increase in the rejection rate of internal emergency and elective

patients. More importantly, the overall probability that patients (of all classes) cannot be admitted due to lack of bed space, denoted by O , is decreased by more than 55%, due to better utilization of bed spaces.

We would like to mention that, as a side-note and discussed in [4], applying another round of EFPA to evaluate the rejection rate of internal emergency patients and deferral rate of elective patients can improve compared to relying on IESA to approximate all three metrics. However, as the accuracy of IESA is sufficient for the purpose in this paper, we shall not discuss the approach involving both IESA and EFPA, but refer the interested readers to [4] for more details.

C. General case ($3 \leq N \leq 17$) for validating the solutions and comparing computational efficiency

We now consider more general city-wide ICU systems. As mentioned before, in the case of Hong Kong, there are a total of 17 ICU hospitals in the territory. Therefore, we will take 17 as the maximum possible number of hospitals in the network throughout this subsection. Meanwhile, the total number of available beds at each ICU hospital is randomly generated from integers from 9 to 25 (both inclusive) with equal probability. We use different combinations of evaluation and optimization approaches to solve (1). We set the relevant parameters as $R_{\text{req}}^I = R_{\text{req}}^E = 0.1$, $D_{\text{req}} = 0.3$, and $r_{\text{max}}^{\text{in}} =$

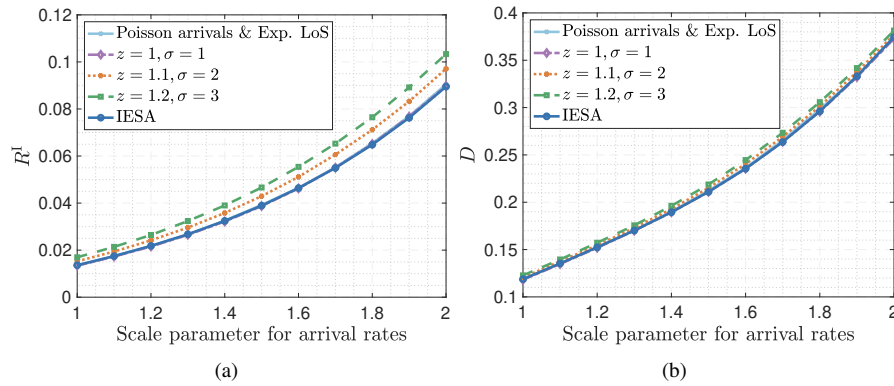


Fig. 6. IESA and simulation results for (a) R^I , and (b) D , with different arrival processes (z denotes the peakedness), LoS distributions (σ denotes the standard deviation), and changing arrival rates (17 hospitals).

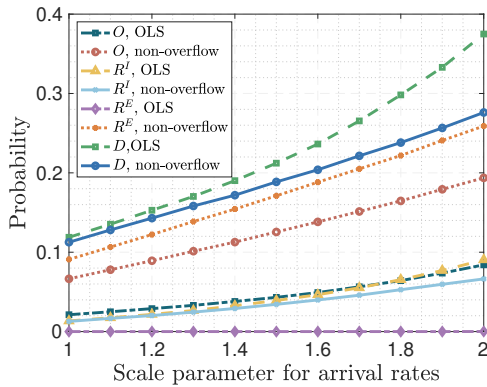


Fig. 7. A comparison of QoS metrics between OLS and non-overflow models (O : overall probability that patients of all class cannot be admitted due to lack of bed spaces).

$r_{\max}^{\text{ex}} = r_{\max}^{\text{el}} = 5$. For PSO-related parameters, we set $c_1 = c_2 = 1.4962$, $w = 0.999^t$ where t is the number of iterations executed. As we mentioned previously in the paper, our optimization formulation can flexibly cater for different situations in ICU networks. The parameters used in this subsection only represent an example for demonstrating the time efficiencies of different evaluation and optimization approaches and can be adjusted accordingly in practical scenarios.

Recall that, as demonstrated earlier in this paper, the rejection rate is not very sensitive to whether the arrival processes is homogeneous Poisson or IPP, and LoS distributions beyond the mean. We will consider Poisson arrivals and exponential LoS distributions in all experiments in this subsection. Therefore, it is possible to adopt the more efficient MCS instead of DES for simulation. We present the results for both simulation methods.

We first present the convergence performance of PSO with respect to different sizes of particle population, $n_{\text{particles}}$. In Fig. 8, we consider three cases with $N = 5, 10$ and 17 , and demonstrate the ratio of global best value of the objective function (1) by in the t -th iteration to the optimal solution C^* obtained by the exhaustive search method, denoted by ϕ . A ϕ value close to 1 means that the PSO solution approaches the optimal solution. The results show that, even for an ICU network with relatively large scale ($N = 17$), PSO with a modest size of particle populations ($n_{\text{particles}} = 50$) can

approach the optimal solution generally within 15 iterations. This finding indicates that PSO-based methods require neither increasing particle population size nor additional iterations for convergence to the optimal solution for an increase in system size within practical ranges, which is a significant improvement compared to the exponential computational complexity for the exhaustive search method.

Specifically, in terms of computational time required for different combinations of evaluation and optimization approaches, are presented in Fig. VII-C, where curves with “ES” represents the theoretical running time of using the Exhaustive Search method for identifying the optimal thresholds, based on the number of evaluations required and the average running time of each evaluation approach for the given network size. For PSO-based methods in this set of experiments, we consider $n_{\text{particles}} = 150$ for a better balance between the number of iterations required to converge and the computation time in each iteration. In addition, we also present the results for a generalized version of the branch-and-cut (BC) method for solving NP-hard NILPs [51]. We observe from the results that Heuristics PSO + IESA can achieve significant improvement in terms of computational efficiency over the benchmark methods (ES/BC+DES/MCS), especially for medium-to-large system sizes, indicating that PSO and IESA are the most effective combination in finding the optimal solution. Again, we would like to mention that the only viable approach in existing similar studies is to use simulation to evaluate the QoS metrics under a specific set of system parameters, and then exhaustively determine the feasibility and optimality of every setting. The exhaustive search nature of this method makes it practically infeasible as shown in Fig. VII-C for practical sizes of ICU networks, as the number of required computations increases exponentially with the network size. Approaches involving the PSO, on the other hand, due to the fast convergence shown in Fig. 8, are relatively insensitive in terms of running time against the system size. In addition, in all cases, IESA+PSO can reduce the running time by no less than four orders of magnitudes compared to MCS+PSO (the second best method), which justifies the practicability of IESA as an efficient evaluation tool. Finally, the optimal thresholds obtained by IESA in all experiments carried out in this paper

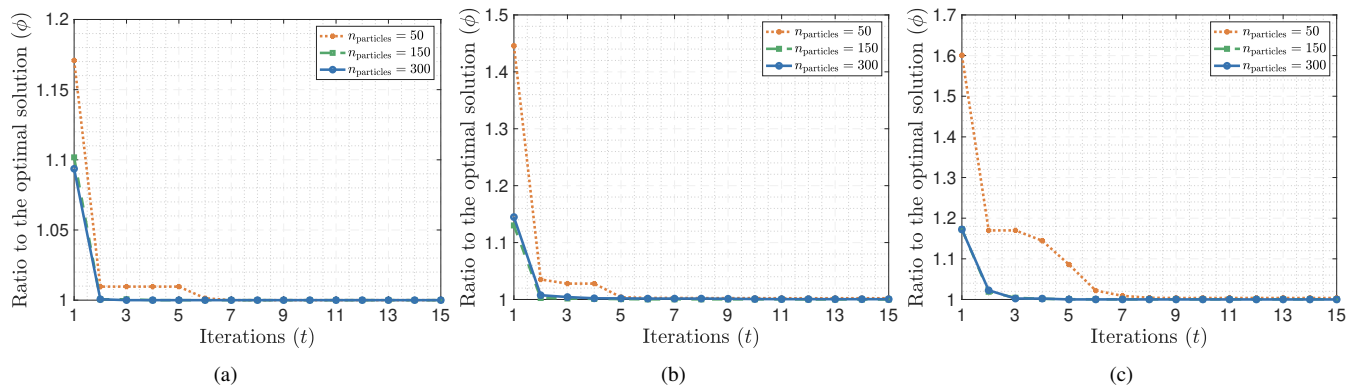


Fig. 8. PSO convergence performance with respect to the optimal solution with (a) $N = 5$, (b) $N = 10$, and (c) $N = 17$.

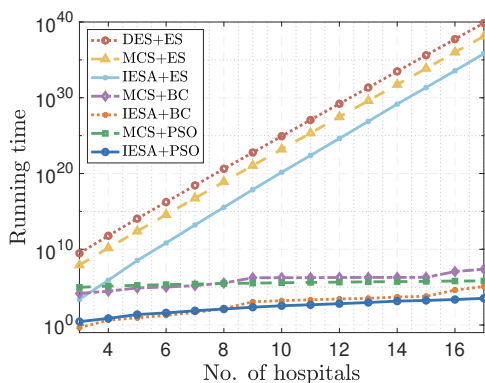


Fig. 9. A comparison of computational efficiency of different optimization approaches as the number of ICU hospitals in the system increases (ES = theoretical running time of Exhaustive Search method; BC = Branch and Cut method).

are the same as those obtained by DES or MCS, due to the high accuracy of IESA in evaluating the QoS metrics demonstrated in the last subsection.

All these observations well justify the practical value of the evaluation and optimization tools we proposed in this paper, especially during emergencies when policies are required to be determined and carried out in a limited period of time.

VIII. CONCLUSIONS

We studied an ICU network model composing of multiple hospitals and three patient classes, and investigated the effect of threshold-based bed reservation policies on the QoS metrics of each patient class. In particular, patients in one class (external emergency patients) are allowed to overflow to other hospitals in a non-hierarchical setting, leading to the *mutual overflow* phenomenon that complicates the performance evaluation, design, and optimization processes.

To tackle the complexity issue in determining the optimal thresholds for each class that can guarantee both performance and fairness requirements, we proposed an approach that uses IESA to evaluate the performance for each set of thresholds, and PSO to search for the optimal set based on the approximation results by IESA. Experiment results showed that the newly proposed approach can reduce the total time in determining the optimal thresholds significantly compared to

the exhaustive search-simulation combination commonly used in similar existing studies, while attaining satisfactorily close results, for ICU networks with practical sizes.

We also analyzed the real data on patient arrival times from three ICU hospitals in Hong Kong, and identified that IPP can fit the first three moments of real data traces satisfactorily. In addition, our numerical results showed that approximating IPP by the homogeneous Poisson process, and lognormal LoS distribution (shown to be appropriate in existing studies) by exponential LoS distribution will not significantly affect the QoS metrics. As IESA is based on the assumptions of Poisson arrivals and exponential LoS distributions, this observation further validated that IESA is applicable to the multi-hospital multi-patient-class ICU network problem.

In addition, as demonstrated in Figs. 4 – 6, the relationship between QoS metrics and input parameters (the number of beds or the intensity of offered traffic) is convex under both the simulation and IESA. Therefore, if necessary, it is possible to consider optimization algorithms that converge faster than PSO but require the convexity condition, to integrate with IESA and further improve the computational efficiency.

To summarize, our proposed approach has been demonstrated to be effective and efficient in design, performance evaluation, and optimization of non-hierarchical ICU networks. This will allow academia and practitioners to deal with such problems during critical periods such as pandemics when timely decisions must be made.

DATA ETHICS STATEMENT

The use of data in this paper does not require ethical approval.

REFERENCES

- [1] J. D. Griffiths, N. Price-Lloyd, M. Smithies, and J. E. Williams, "Modelling the requirement for supplementary nurses in an intensive care unit," *J. Oper. Res. Soc.*, vol. 56, no. 2, pp. 126–133, 2005.
- [2] N. Litvak, M. Van Rijsbergen, R. J. Boucherie, and M. van Houdenhoven, "Managing the overflow of intensive care patients," *Eur. J. Oper. Res.*, vol. 185, no. 3, pp. 998–1010, 2008.
- [3] S.-C. Kim, I. Horowitz, K. K. Young, and T. A. Buckley, "Analysis of capacity management of the intensive care unit in a hospital," *Eur. J. Oper. Res.*, vol. 115, no. 1, pp. 36–46, 1999.

- [4] Y.-C. Chan, E. W. Wong, G. Joynt, P. Lai, and M. Zukerman, "Overflow models for the admission of intensive care patients," *Health Care Manag. Sci.*, vol. 21, pp. 554–572, 2018.
- [5] J. Wang, X. Zhong, J. Li, and P. K. Howard, "Modeling and analysis of care delivery services within patient rooms: A system-theoretic approach," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 2, pp. 379–393, 2014.
- [6] J. González, J.-C. Ferrer, A. Cataldo, and L. Rojas, "A proactive transfer policy for critical patient flow management," *Health Care Manag. Sci.*, vol. 22, no. 2, pp. 287–303, 2019.
- [7] S.-C. Kim, I. Horowitz, K. K. Young, and T. A. Buckley, "Flexible bed allocation and performance in the intensive care unit," *J. Oper. Manag.*, vol. 18, no. 4, pp. 427–443, 2000.
- [8] R. Bekker, G. Koole, and D. Roubos, "Flexible bed allocations for hospital wards," *Health Care Manag. Sci.*, vol. 20, no. 4, pp. 453–466, 2017.
- [9] A. Kulshrestha and J. Singh, "Inter-hospital and intra-hospital patient transfer: Recent concepts," *Indian J. Anaesth.*, vol. 60, no. 7, p. 451, 2016.
- [10] B. e Oliveira, J. De Vasconcelos, J. Almeida, and L. Pinto, "A simulation-optimisation approach for hospital beds allocation," *Int. J. Med Inform.*, vol. 141, p. 104174, 2020.
- [11] E. W. Wong and Y.-C. Chan, "A century-long challenge in teletraffic theory: Blocking probability evaluation for overflow loss systems with mutual overflow," *IEEE Access*, 2023.
- [12] E. L. Blair and C. E. Lawrence, "A queueing network approach to health care planning with an application to burn care in New York state," *Socio-economic planning sciences*, vol. 15, no. 5, pp. 207–216, 1981.
- [13] X. Wei, P. Ding, L. Zhou, and Y. Qian, "QoE oriented chunk scheduling in P2P-VoD streaming system," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8012–8025, 2019.
- [14] W. Sun, J. Liu, Y. Yue, and Y. Jiang, "Social-aware incentive mechanisms for D2D resource sharing in IIoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5517–5526, 2020.
- [15] F. Mallor and C. Azcárate, "Combining optimization with simulation to obtain credible models for intensive care units," *Annals of Operations Research*, vol. 221, pp. 255–271, 2014.
- [16] Y.-C. Chan, J. Wu, E. W. Wong, and C. S. Leung, "Integrating teletraffic theory with neural networks for Quality-of-Service evaluation in mobile networks," *Applied Soft Computing*, p. 111208, 2023.
- [17] M. Lu, H.-C. Lam, and F. Dai, "Resource-constrained critical path analysis based on discrete event simulation and particle swarm optimization," *Autom. Constr.*, vol. 17, no. 6, pp. 670–681, 2008.
- [18] H. Zhang, T. J. Best, A. Chivu, and D. O. Meltzer, "Simulation-based optimization to improve hospital patient assignment to physicians and clinical units," *Health Care Manag. Sci.*, vol. 23, pp. 117–141, 2020.
- [19] J. Bai, J. O. Brunner, and S. Gerstmeier, "Simulation and evaluation of ICU management policies," in *2020 Winter Simulation Conference (WSC)*, 2020, pp. 864–875.
- [20] E. W. Wong, J. Guo, B. Moran, and M. Zukerman, "Information exchange surrogates for approximation of blocking probabilities in overflow loss systems," in *Proceedings of the 2013 25th International Teletraffic Congress (ITC)*, 2013, pp. 1–9.
- [21] Y.-C. Chan, E. W. Wong, and C. S. Leung, "Evaluating non-hierarchical overflow loss systems using teletraffic theory and neural networks," *IEEE Communications Letters*, vol. 25, no. 5, pp. 1486–1490, 2021.
- [22] J. Wu, E. W. Wong, J. Guo, and M. Zukerman, "Performance analysis of green cellular networks with selective base-station sleeping," *Performance Evaluation*, vol. 111, pp. 17–36, 2017.
- [23] D. Bertsimas, J. Pauphilet, J. Stevens, and M. Tandon, "Predicting inpatient flow at a major hospital using interpretable analytics," *Manuf. Serv. Oper. Manag.*, vol. 24, no. 6, pp. 2809–2824, 2022.
- [24] T. Zhu, L. Luo, X. Zhang, Y. Shi, and W. Shen, "Time-series approaches for forecasting the number of hospital daily discharged inpatients," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 2, pp. 515–526, 2017.
- [25] S.-Y. Lee, R. B. Chinnam, E. Dalkiran, S. Krupp, and M. Nauss, "Prediction of emergency department patient disposition decision for proactive resource allocation for admission," *Health Care Manag. Sci.*, vol. 23, pp. 339–359, 2020.
- [26] E. H. Kaplan, "OM Forum — COVID-19 scratch models to support local decisions," *Manuf. Serv. Oper. Manag.*, vol. 22, no. 4, pp. 645–655, 2020.
- [27] R. Liu, J. Xu, and Y. Liu, "Dynamic patient admission control with time-varying and uncertain demands in COVID-19 pandemic," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 2, pp. 620–631, 2021.
- [28] F. Parker, H. Sawczuk, F. Ganjkanloo, F. Ahmadi, and K. Ghobadi, "Optimal resource and demand redistribution for healthcare systems under stress from COVID-19," *arXiv preprint arXiv:2011.03528*, 2020.
- [29] L. Nielsen and A. Albertsen, "Pandemic justice: fairness, social inequality and COVID-19 healthcare priority-setting," *Journal of Medical Ethics*, vol. 49, no. 4, pp. 283–287, 2023.
- [30] S. Dijkstra, S. Baas, A. Braaksma, and R. J. Boucherie, "Dynamic fair balancing of COVID-19 patients over hospitals based on forecasts of bed occupancy," *Omega*, vol. 116, p. 102801, 2023.
- [31] S. Qiu, R. B. Chinnam, A. Murat, B. Batarsee, H. Neemuchwala, and W. Jordan, "A cost sensitive inpatient bed reservation approach to reduce emergency department boarding times," *Health Care Manag. Sci.*, vol. 18, pp. 67–85, 2015.
- [32] W. Hu, C. W. Chan, J. R. Zubizarreta, and G. J. Escobar, "An examination of early transfers to the ICU based on a physiologic risk score," *Manuf. Serv. Oper. Manag.*, vol. 20, no. 3, pp. 531–549, 2018.
- [33] C. Pagel, V. Banks, C. Pope, P. Whitmore, K. Brown, A. Goldman, and M. Utley, "Development, implementation and evaluation of a tool for forecasting short term demand for beds in an intensive care unit," *Operations Research for Health Care*, vol. 15, pp. 19–31, 2017.
- [34] C. W. Seymour, T. J. Iwashyna, W. J. Ehlenbach, H. Wunsch, and C. R. Cooke, "Hospital-level variation in the use of intensive care," *Health Serv. Res.*, vol. 47, no. 5, pp. 2060–2080, 2012.
- [35] F. P. Kelly, "Blocking probabilities in large circuit-switched networks," *Advances in applied probability*, vol. 18, no. 2, pp. 473–505, 1986.
- [36] J. Wu, E. W. Wong, Y.-C. Chan, and M. Zukerman, "Power consumption and GoS tradeoff in cellular mobile networks with base station sleeping and related performance studies," *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 4, pp. 1024–1036, 2020.
- [37] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, 1995, pp. 1942–1948.
- [38] Y. Wang, H. Xie, X. Jiang, and B. Liu, "Intelligent closed-loop Insulin delivery systems for ICU patients," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 1, pp. 290–299, 2014.
- [39] J. Sun, C.-H. Lai, and X.-J. Wu, *Particle swarm optimisation: classical and quantum perspectives*. CRC Press, 2016.
- [40] E. C. Laskari, K. E. Parsopoulos, and M. N. Vrahatis, "Particle swarm optimization for integer programming," in *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02*, vol. 2. IEEE, 2002, pp. 1582–1587.
- [41] Z. Li, W. Tian, and J. Wu, "Modeling and joint optimization of security, latency, and computational cost in Blockchain-based healthcare systems," in *2023 IEEE ICC Workshops*, 2023, pp. 1938–1943.
- [42] M. Y. N. Attari, M. Ahmadi, A. Ala, and E. Moghadamnia, "RSDM-AHSnet: Designing a robust stochastic dynamic model to allocating health service network under disturbance situations with limited capacity using algorithms NSGA-II and PSO," *Computers in Biology and Medicine*, vol. 147, p. 105649, 2022.
- [43] P. Gupta, S. Bhagat, D. K. Saini, A. Kumar, M. Alahmadi, and P. C. Sharma, "Hybrid whale optimization algorithm for resource optimization in cloud e-healthcare applications," *Computers, Materials & Continua*, vol. 71, no. 3, 2022.
- [44] P.-S. Chen and Z.-Y. Zeng, "Developing two heuristic algorithms with metaheuristic algorithms to improve solutions of optimization problems with soft and hard constraints: An application to nurse rostering problems," *Applied Soft Computing*, vol. 93, p. 106336, 2020.
- [45] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*. Princeton, NJ: Princeton University Press, 1994.
- [46] Hong Kong Hospital Authority, "Statistical report 2022-2023," 2023. [Online]. Available: <https://www3.ha.org.hk/data/HASStatistics/DownloadCluster/49>
- [47] D. L. Jagerman, "Methods in traffic calculations," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 7, pp. 1283–1310, 1984.
- [48] M. Cheng, Y. Tong, T. Kwok, I. Cheng, A. Chung, J. Leung, M. Fung, and N. Cheung, "Development journey of Clinical Data Analysis and Reporting System (CDARS) in Hospital Authority of Hong Kong," in *13th World Congress on Medical Informatics*, 2010, p. 1648.
- [49] X. Zhang, S. Barnes, B. Golden, M. Myers, and P. Smith, "Lognormal-based mixture models for robust fitting of hospital length of stay distributions," *Oper. Res. for Health Care*, vol. 22, p. 100184, 2019.
- [50] F. Mallor, C. Azcárate, and J. Barado, "Optimal control of ICU patient discharge: from theory to implementation," *Health Care Manag. Sci.*, vol. 18, pp. 234–250, 2015.
- [51] P. Kesavan and P. I. Barton, "Generalized branch-and-cut framework for mixed-integer nonlinear optimization problems," *Computers & Chemical Engineering*, vol. 24, no. 2, pp. 1361–1366, 2000.