Submitted to *INFORMS Journal on Computing*
manuscript (Please, provide the manuscript number!)

Acuity-Based Allocation of ICU-Downstream Beds with Flexible Staffing

Silviya Valeva
Department of Decision & System Sciences, Erivan K. Haub School of Business, Saint Joseph's University, svaleva@sju.edu

Guodong Pang
Department of Computational Applied Mathematics and Operations Research, Rice University, gdpang@rice.edu

Andrew J. Schaefer
Department of Computational Applied Mathematics and Operations Research, Rice University, andrew.schaefer@rice.edu

Gilles Clermont
Critical Care Medicine, University of Pittsburgh, cler@pitt.edu

Intensive care units (ICUs) are crucial resources within hospitals, caring for the most critically ill patients. We propose a novel modeling framework that improves the outflow of ICU patients by anticipating unit interactions and resource sharing within the system. Across an arbitrary bipartite network of units, we consider two types of downstream staffing (baseline and flexible) and a two-stage decision process. In the first stage, we determine the level of flexible bed staffing using existing physical beds at downstream units in anticipation of incoming transfers from the ICUs. In the second stage, we determined the allocation of ICU patients to downstream beds. The goal of the model is to reduce inefficiencies and transfer delays causing ICU bed block due to lack of space in downstream units. We formulate a dynamic multi-period model and analyze the dual of its (relaxed) stationary counterpart. Decomposing the relaxed stationary model into an ICU and downstream subproblems, we calculate the relative values of downstream beds and derive a practical acuity-based policy for the daily operational decisions. Using a large-scale simulation calibrated with historic hospital data, we demonstrate that our acuity-based policy reduces the number of long-run diverted ICU arrivals, particularly in high-demand scenarios, thus improving ICU throughput, when compared to a deterministic, a generalized randomized-most-idle, and static policies.

Key words: ICU management; flexible staffing; acuity-based policy

1. Introduction

Hospitals face extensive demand from patients for improved clinical outcomes and service quality. However, inefficiencies and delays in bed usage as well as unproductive occupancy resulting from waiting for tests, transfer, or discharge often lead to patient blocking, diversions, unit congestions (Green 2005, Hall 2012), increased waiting time, rescheduling of planned procedures, and reduced perceived quality of service for patients (de Bruin et al. 2010). Furthermore, inefficient use of facilities often poses a significant financial strain, as many US hospitals often operate at or near full capacity and regularly face shortages of healthcare workers (Toner and Waldhorn 2020). The situation is further exacerbated during local or global surges in demand for beds. As the COVID-19 pandemic has illustrated, ICU capacity can quickly reach its limits (Leatherby et al. 2020), calling for improved hospital management policies (Landro 2020) and making efficient use of the existing scarce resources even more pressing. Hall (2012) identifies bed management as a crucial step for reducing inefficiencies, as for most hospitals the available beds and the types of patients they serve create bottlenecks. While the whole hospital is a highly complex network that is extremely difficult to study, interactions among units should not be ignored (Armony et al. 2015). Hence, studies of hospital sub-networks provide a more focused and tractable middle ground. We focus our study on the sub-network of intensive care units (ICUs) and downstream units within a hospital.

ICUs provide temporary care to critically ill patients, who usually arrive through emergency departments or immediately after surgery. ICUs are characterized by high nurse-to-patient ratio (1:2 or 1:1) and expensive specialized equipment. Consequently, a day in an ICU bed can be 2.5 times more costly than elsewhere in a hospital (Barrett et al. 2014). Proper management of ICUs is important not only from a financial standpoint but also for patient safety, as increased ICU occupancy is associated with a heightened risk of patients' early death or readmission (Chrusch et al. 2009). Furthermore, delayed or refused patient admission in ICUs from emergency departments is associated with higher length of stay and increased mortality (Metcalfe et al. 1997, Chalfin et al. 2007, Robert et al. 2012). Thus, efficient utilization of ICU beds is crucial for reducing cost, limiting delays or refused admissions, and increasing patient throughput and health outcomes.

Downstream units (or general wards) provide care for stabilized and recovering patients after an ICU stay and patients arriving from other units throughout the hospital. In practice, and depending on their condition, patients may need to be transferred to different units of varying levels of care after an ICU stay, such as step-down units, telemetry units, etc. For the purposes of this study, we consider one layer of lower-level care units with possibly different connections to ICUs, referred to as downstream units (DSUs). DSUs have a lower nurse-to-patient ratio compared to ICUs (about 1:8) and are significantly less expensive to operate. A single ward generally treats patients requiring similar kind of care as determined by staff training. Thus, a particular downstream ward may only accept patients from certain ICUs, resulting in a network of ICU-downstream routes, where ICU patients are transferred to downstream units before being discharged from the hospital. The topology of the network is determined by the existing hospital layout, policies, and staff training. If no restrictions exist, we assume a complete bipartite network where each ICU is connected to each downstream unit. The bed capacities at units are largely determined by staff availability.

This study focuses on two types of decisions in the ICU-DSU network, namely, the flexible staffing of downstream beds and the allocation of ICU patients to downstream beds (see Figure 1).

The discharge of a patient from the ICU happens after his/her status has been updated to *ready* – a decision made by a physician based on health indicators(note that individual patient discharge decisions are not modeled in this paper). The request for a ward bed is relayed to a ward bed manager, who identifies an available bed in a downstream unit where the patient can be transferred. If no downstream bed is available to accommodate a transfer request, the patient will remain in ICU, which in turn could lead to incoming patient diversions and refused admissions. The discharge delay of ICU patients can be defined as the time lag between the moment patients are declared clinically ready for discharge from ICU and the moment when they physically vacate the ICU bed (Perlmutter et al. 1998, Williams and Leslie 2004, Chaboyer et al. 2006). Lack of beds in downstream units is a common reason for delayed discharge of ICU patients (Levin et al. 2003, Lin et al. 2009). As a result, patients often cannot be admitted into ICU because it is full and a number of ICU beds are occupied by patients waiting for ward beds, a situation commonly referred to as "bed-block" or "outflow limitation" (Lin et al. 2009, Zychlinski et al. 2020). One way to mitigate this issue is by staffing additional beds at downstream units during periods of peak demand. Generally, there are fewer staffed beds in a unit than physical beds, however, only staffed beds are available to accept patients (de Bruin et al. 2010). We distinguish between two types of staffing: *baseline staffing* and *flexible staffing.* Baseline staffing refers to the regular bed staffing done with full-time nurses and based on expected demand. Flexible staffing is the additional staffing of existing physical beds (usually with temporary or on-call personnel) when needed. Saville et al. (2020) show that in addition to baseline staffing, flexible staffing can be an effective measure for responding to variation in demand and can reduce both overstaffing and understaffing. In particular, we consider a setting in which (some) downstream units may have extra beds that can be staffed with on-call personnel in order to better meet ICU transfer demand when baseline staffed beds are unavailable. We thus propose a model to determine the number and location of additional operational beds through flexible staffing and the allocation of ready ICU patients to DSU beds. Modeling the two decisions together allows improved bed availability for incoming ICU patients and, to the best of our knowledge, has not been studied before. Note that Zychlinski et al. (2020) study capacity allocation and periodic reallocation of beds at geriatric units, however, their approach considers the long-term setup of beds which remains fixed until the next round of reallocation (e.g., quarterly). On the other hand, we consider the short-term staffing with on-call personnel of already existing beds which can vary daily.

This decision environment poses several challenges:

(i) the arbitrary topology of the bipartite ICU-downstream network (note that, except ICU connectivity, we do not impose any structural restrictions on number of units or connections);

(ii) the aggregate number of transfers out of ICUs that are sharing (and competing for) downstream resources; and

(iii) the novel aspect of studying flexible staffing decisions in conjunction with patient allocation in downstream units.

Thus, instead of an individual-based queuing theory analysis, we propose a dynamic multi-period model to determine the staffing and allocation decisions. We use the dual of its relaxed stationary counterpart to derive relative values of the downstream resources which we then embed in an operational policy. As the DSU bed values are derived from a stationary system model, they capture the *global* system dynamics, including the topology of the underlying (arbitrary) network with resource sharing, as well as the capacities at units and expected arrival rates, captured through a clinical benefit function. The proposed policy then dictates the *local* operational decisions, based on current bed and staff availability. Considering the whole system as opposed to individual waiting times ensures that resources are properly shared and utilized within the network and beds are available for the incoming patients when needed.

Our extensive computation study measures the long-run number of diverted ICU patients (i.e., arrivals that are lost to the system) and shows that our policy achieves a reduction when compared to several other policies. This metric is significant as research has shown that delayed or refused patient admission in ICUs is associated with higher length of stay and increased mortality (Metcalfe et al. 1997, Chalfin et al. 2007, Cardoso et al. 2011, Robert et al. 2012). Sensitivity analyses demonstrate that the proposed policy is most valuable in high-demand settings but remains competitive in other scenarios as well.

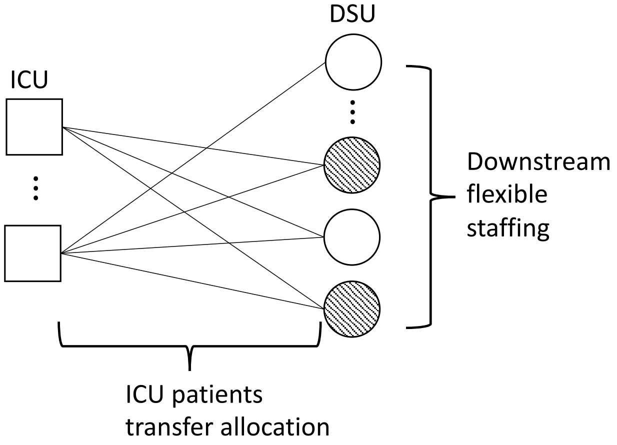

DSU
ICU
Downstream flexible staffing
ICU patients transfer allocation

Figure 1: Illustration of the decisions modeled within the ICU-DSU network including (i) the number and location of downstream flexible staffing beds and (ii) the allocation of ICU patients transferred downstream.

The main contributions of this work can be summarized as follows:

- We provide a dynamic multi-period model to determine efficient staffing and patient allocation at downstream units based on ICU requests, expected arrivals, and downstream bed availability. Our formulation allows for downstream capacity expansion through flexible staffing of beds at additional cost. *This is the first study to consider flexible staffing decisions in an optimization framework that streamlines the ICU-downstream patient transfer process.*

- Through analysis of the relaxed stationary model, we derive relative values of the downstream beds, which in turn allow us to generate an acuity-based policy for the staffing and allocation decisions. In particular, decomposition of the relaxed stationary model allows us to derive the dual bed values from the downstream network subproblem, necessary for guiding the daily downstream staffing and allocation decisions. As the bed values are obtained for the specific underlying network topology and expected unit capacities, they capture global system dynamics which are subsequently used to make local decisions (based on daily availability). The policy provides practical and easy-to-implement operational rules that do not require knowledge of the number of incoming ICU patients or frequent optimization.
- We demonstrate that the integrated decision model can have significant impact on improving efficiency in ICU-DSU management through a series of numerical experiments. In particular, we validate policy's advantages in an illustrative setting as well as a multi-week simulation using arrival and discharge rates calibrated with historic hospital data. We evaluate the performance of the proposed acuity-based policy through comparison with several other policies. The examples illustrate that our proposed acuity-based policy reduces the long-run number of diverted ICU arrivals – a crucial determinant of patient safety and outcomes.

The remainder of this paper is organized as follows: Section 2 discusses the relevant literature. Section 3 presents a dynamic multi-period model for patient transfer and flexible staffing decisions, while Section 4 derives an acuity-based policy from its relaxed stationary counterpart. We discuss the numerical experiments and results in Section 5 and conclude in Section 6.

2. Literature Review

General issues and trends in hospital bed management and patient flow are identified by Green (2005) and Hall (2012). A comprehensive literature review of operations research in ICU management is provided by Bai et al. (2018). The topics of future research identified in the survey that our work addresses include coordinating decisions between the ICU and connected wards as well as the medium and short-term bed capacity planning that involves bed allocation and patient routing. This paper draws from several different streams of research – as such, we organize our review of relevant literature based on the main research methodology.

2.1. Empirical Studies

Empirical approaches to ICU operations have considered patient admissions, discharge, and staff workload. In studying the impact of workload on service time and patient safety, Kc and Terwiesch (2009) find that an increase in workload is associated with an increase in the early patient discharge, which is in turn associated with a heightened mortality rate. Kc and Terwiesch (2012) present an empirical study focused on discharge patterns and the rationing of beds in a cardiac ICU. The authors find that a patient is likely to be discharged early when the occupancy in the ICU is high, which in turn leads to an increased likelihood of the patient having to be readmitted. Kim et al. (2014) focus on ICU admission control to "quantify the cost of denied ICU admissions" and evaluate various admission strategies. The authors use data-calibrated simulation and show that ICU congestion can significantly impact patient admission decisions and patient health outcomes. Kuntz et al. (2014) offer an empirical study on the safety tipping points in hospitals based on capacity utilization, i.e., occupancy beyond a certain level is associated with a significant escalation of in-hospital mortality. As the increased stress levels caused by high utilization forces clinical staff to ration resources and become more error prone, the authors argue that a cost-effective solution for safety improvement is flexible capacity expansion (as opposed to rigid capacity expansion) as it is only used when occupancy reaches the threshold of the tipping point. Song et al. (2020) study capacity pooling, i.e., the practice of assigning patients from a unit whose beds are currently fully occupied to an available bed in a different unit. The study finds the practice to be associated with an increase in the patients' remaining length of stay (LOS) and an increase in their readmission likelihood. The negative impacts that ICU congestion has on patients, as identified in the empirical literature, serve as motivating factors for seeking new ways to improve patient flow efficiency by allowing for flexibility through strategic staffing.

2.2. Queueing Models

There is a rich literature of queueing models in ICU management. Shmueli et al. (2003) consider different ICU admission policies and study how each impacts the expected number of saved lives. Chan et al. (2012) study discharge decisions and propose a policy for indexing (ranking) of patient criticality to use in demand-driven discharge from ICUs. Bekker et al. (2017) consider flexibility in the usage of beds in hospital wards and analyze various bed allocation policies. The study examines patient admissions in settings including both dedicated/earmarked beds (treating a specific type of patients) and flexible beds (treating multiple types of patients). Our work is also related to queueing literature with fairness routing policies in stochastic networks, see, e.g., the threshold-type of policies in Ward and Armony (2013), diffusion controls in Arapostathis and Pang (2019), and in particular, the randomized-most-idle (RMI) algorithm by Mandelbaum et al. (2012) in the network of a single emergency department (ED) to hospital wards. The proposed acuity-based policy can be potentially used to study these ED-Ward networks.

The decisions in queues are made on the individual-patient basis with detailed models of arrival and service and routing processes, and well-established theoretical results like (asymptotically) optimal policies often require strong assumptions on the topology of the underlying network (especially for parallel server networks as in our study) of servers as well as heavy-traffic asymptotic regimes. For example, the well-known $c\mu$-rule was first established for a “V” network (Buyukkoc et al. 1985) and extended to general parallel server network with “tree” topology (Mandelbaum and Stolyar 2004). Scheduling and routing problems in large-scale parallel server networks have mostly focused on tree topology networks Atar et al. (2004), Atar (2005), Harrison and Zeevi (2004), Gurvich and Whitt (2009a,b), Stolyar and Tezcan (2011), Arapostathis and Pang (2019). An exception is the “X” model in Perry and Whitt (2009, 2013).

In reality, many, if not most ICU settings entail a non-tree topology, as studied in this paper (see the ICU-DSU networks constructed from real data in the e-companion). Recognizing the specificity of DSUs within the network and their varying clinical value is especially important in the case of complex network topologies with resource sharing, as is the case in many hospital settings.

2.3. Optimization Approaches

Dobson et al. (2010) introduce a Markov chain model for a single ICU with patient bumping, i.e., the transfer of patients to other units to make room for new incoming patients. While this approach allows for exact mathematical analysis as opposed to simulation, its complexity for a single ICU makes it difficult to generalize for an entire hospital setting with multiple units. Thompson et al. (2009) study admission allocation and reallocation in anticipation of a demand surge through a finite-horizon Markov decision process (MDP). Here, the patient transfer/reallocation is not a result of medical need but rather a proactive step towards capacity reallocation. Due to the high dimensionality of MDPs, the authors resort to multiple approximations including a reduction in the policy space and time horizon. It is thus unlikely for this approach to be scalable to larger hospitals with more patient categories and unit types. Bertsimas and Pauphilet (2020) study admission patient allocation and integrate machine learning into the optimization models. Three models comprising an immediate, daily, and weekly horizon are proposed and integrated into one model that creates bed assignments based on current (fixed) bed availability and bed requests. The objective function is a sum of various cost functions and a deviation from target occupancy levels.

Methodologically, our approach is inspired by the price-directed routing and control used in manufacturing. For example, Roundy et al. (1991) develop a price-directed methodology for machine scheduling in a job shop. Other applications of price-directed approaches can be found in revenue management (Simpson 1989, Williamson 1992, Talluri and van Ryzin 1998), vehicle dispatching (Gans and van Ryzin 1999), logistics networks (Adelman 2007), restless bandits problems (Bertsimas and Niño-Mora 2000), remnant inventory control (Adelman and Nemhauser 1999, Rajgopal et al. 2009), and the economic lot scheduling problem (Adelman and Barz 2014). These approaches differ from classic price-directed methods, such as the Dantzig-Wolfe decomposition (Dantzig and Wolfe 1960), as the dual prices are not used to solve a problem instance, but instead to design control policies for the underlying dynamic system.

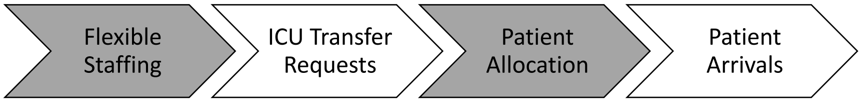


Figure 2 Timeline of evens – decisions are indicated in gray while external events are indicated in white.

In practice, highly dynamic and stochastic control problems cannot be solved to optimality due to the "curse of dimensionality" (Bellman 1957). Thus, price-directed approaches seek to (periodically) estimate the global system dynamics through a steady-state model from which they derive static dual prices of resources and then use the dual prices to dictate the local operational decisions. Hospitals fit the description of highly dynamic and stochastic systems with complex interactions between the sources of supply (staff and beds) and demand (patients from various units). As such, our approach uses the global system defined by the ICU-downstream network with expected supply and demand parameters to estimate the relative values of the downstream resources. These values are then used to guide the individual operational decisions of staffing and patient allocation through a practical acuity-based policy.

3. Model

We consider the decisions of determining the number and location of beds added through flexible staffing (i.e., beyond the regular baseline staffing), together with the transfer and allocation of ICU patients to downstream units. We first describe the environment and factors considered by the decision maker, followed by a formal presentation of the mathematical model.

3.1. Decision Environment

We consider a general setting in which a decision maker determines the transfer and allocation of ICU patients to downstream units by utilizing operational beds available through baseline staffing (matching expected demand) and flexible staffing (short-notice on-call or temporary personnel called when surges in demand occur). Specifically, the decisions in each time period can be subdivided in two stages: (i) staffing and (ii) patient allocation. Note that we model an environment in which physical beds are already available, thus the first-stage decisions are to staff the existing beds. Figure 2 illustrates a timeline of decisions. This timeline can represent a day or half day (as in our computational experiments) but the model is independent of the specific choice of time epochs. Daily or twice daily discharge/transfer decisions are common practice in hospitals (Chan et al. 2017).

The stage-one decisions represent an important feature of our decision environment – the use of flexible staffing for capacity expansion downstream. In our setting, (some) downstream units may allow for additional bed staffing to better meet ICU transfer demand when the number of baseline staffed beds becomes insufficient. Studies have shown that lack of beds at the downstream general wards is one of the most common reasons for delayed discharge of ICU patients (Levin et al. 2003, Lin et al. 2009). In particular, the additionally staffed beds at a given unit can treat incoming patients from the same set of ICUs as determined by the existing network. As staffing requires planning and lead time, we assume a "budget" measuring the maximum number of additional beds that can be staffed is available in each decision period. This value may be calculated based on a monetary flexible staffing budget that is a fraction of the whole-hospital budget. Throughout the rest of the work, we make the following assumption:

Assumption 1. *The flexible staffing cost at downstream units is linear in the number of beds staffed beyond baseline capacity.*

For our modeling purposes (and as is common practice), we assume the flexible staffing cost is proportional to the number of operational beds beyond baseline staffing in a given time period. This is a reasonable assumption given that the required staff per bed is fixed, usually state mandated and department specific. In particular, the downstream staffing cost function is mainly governed by the financial cost associated with scheduling additional personnel. Note that while nurses are generally hired to work in a given unit, floating among units may sometimes be necessary and economical but has been associated with certain negative outcomes (Bates 2013, Hendren 2011, O'Connor and Dugan 2017). Hence, we model the flexible staffing decisions with the inherent cost of on-call personnel, i.e., using the on-call nurses from the given unit rather than redeploying (floating) regularly scheduled nurses from other units.

At the stage-two decision making, each ICU provides a request for the maximum number of patients they would like to transfer downstream, i.e., the number of patients who are medically ready to be transferred. As discussed in the Introduction, patients staying at ICU while awaiting transfer is viewed as unproductive occupancy, i.e., it is not medically necessary for patients to occupy those beds (Hall 2012). Such holdups could lead to blocking of incoming ICU patients and temporary diversions – generally viewed as inefficiencies and negative factors for patient throughput, health outcomes, and quality of service (de Bruin et al. 2010, Green 2005, Hall 2012).

Remark. We assume ICU patients are ordered based on acuity and ready patients are selected for downstream transfer based on this ordering.

An important consequence of the two-stage decisions is the access to ICU beds for incoming patients. As ICU patients usually require immediate care, it is generally not possible to make the newly arriving patients queue for beds (Bai et al. 2018). In practice, those patients may be rerouted to other facilities; however, the different ICUs within a hospital tend to be specialized (e.g., cardiac, surgical, neonatal, etc.) and as such, may not be able to serve all incoming patients. Similarly, staff training at downstream units may prevent certain units from accepting unmet demand for other units. We assume that patients who cannot be admitted are routed to other facilities and are not modeled in our problem setting.

Assumption 2. *Arrivals at ICUs and downstream units are admitted as long as there are beds available. New patient arrivals who cannot be admitted are diverted (i.e., lost to the system).*

3.2. Dynamic Model Formulation

In this section we formalize a multi-period discrete-time dynamic model for the ICU-downstream patient transfer problem with flexible staffing. Table 1 presents a summary of the notation used. Let $\mathcal{I}$ represent the set of ICUs and let $\mathcal{J}$ represent the set of downstream units (DSUs). Further let $\bar{C}_i$ be the number of staffed beds at ICU i. As we consider two types of staffing downstream, let $\bar{D}_j$ be the baseline capacity and let $\bar{D}'_j$ be flexible staffing capacity in unit $j \in \mathcal{J}$. The network of feasible patient routes forms a bipartite graph $\mathcal{G} = (\mathcal{I} \cup \mathcal{J}, \mathcal{E})$ with edges between the sets $\mathcal{I}$ and $\mathcal{J}$ such that for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$, the sets $\mathcal{J}(i) \subseteq \mathcal{J}$ and $\mathcal{I}(j) \subseteq \mathcal{I}$ denote the set of downstream units to which patients from ICU i can be transferred and the set of ICUs from which patients can be admitted at downstream unit j, respectively. In particular, we have $\mathcal{J}(i) = \{j \in \mathcal{J} \mid (i,j) \in \mathcal{E}\}$ and $\mathcal{I}(j) = \{i \in \mid (i,j) \in \mathcal{E}\}$.

Parameters		Decision variables	
$\mathcal{I}$	set of ICUs	d_j^t	number of beds made available with flexible staffing at DSU j at time t
$\bar{C}_i$	staffed beds at ICU i		
$\mathcal{J}$	set of DSUs	x_{ij}^t	number of patients transferred from ICU i to DSU j at time t
$\bar{D}_j$	baseline capacity at DSU j		
$\bar{D}'_j$	flexible staffing capacity at DSU j		
$\mathcal{J}(i)$	set of feasible DSUs for ICU i		
$\mathcal{I}(j)$	set of feasible ICUs for DSU j		
$\mathcal{T}$	time horizon		
P_i^t	occupied beds at ICU i at time t		
Q_j^t	occupied baseline beds at DSU j at time t		
$Q_j^{\prime t}$	occupied flexible staffing beds at DSU j at time t		
C_i^t	transfer requests from ICU i at time t		
ν^t	number of additional beds that may be staffed given available personnel		

Table 1 Notation summary.

Given a time horizon $\mathcal{T} = \{1, 2, \dots\}$, let $\mathbf{P}^t = \{P_i^t \mid i \in \mathcal{I}\}$ be the set of patient populations at each ICU at the beginning of time period t. Let $\mathbf{Q}^t = \{Q_j^t \mid j \in \mathcal{J}\}$ and $\mathbf{Q}'^t = \{Q_j'^t \mid j \in \mathcal{J}\}$ be the sets of patients at each downstream unit occupying baseline and additional (flexible staffing) beds, respectively, at the beginning of time period t. Let $\mathbf{C}^t = \{C_i^t \mid i \in \mathcal{I}\}$ represent all the ICU transfer requests, i.e., the number of ready patients at time t as determined by their healthcare team. Let ν^t be the total number of personnel available for additional bed staffing at period t, normalized for the required bed-staff ratio. For any given period t in the time horizon $\mathcal{T}$, we define the system's state and dynamics as follows:

(i) *State of the system:* $S_t = (\mathbf{P}^t, \mathbf{C}^t, \mathbf{Q}^t, \mathbf{Q}'^t, \nu^t)$ contains the information available to the decision maker at the given time period.

(ii) *Decision variables:* $\mathbf{d}^t = \{d_j^t \in \mathbb{Z}_+ \mid j \in \mathcal{J}\}$ determines the number of additional beds available through flexible staffing at time period t (first-stage decisions) and $\mathbf{x}^t = \{x_{ij}^t \in \mathbb{Z}_+ \mid i \in \mathcal{I}, j \in \mathcal{J}(i)\}$ determines the patient transfer allocation from ICUs to DSUs at time t (second-stage decisions). Decisions are made before the exogenous information on new patient arrivals and future staff availability is revealed and must satisfy the following constraints:

$$d_j^t \le \bar{D}'_j - Q'^t_j \qquad \forall j \in \mathcal{J}, \tag{1a}$$

$$\sum_{j \in \mathcal{J}} d_j^t \le \nu^t, \tag{1b}$$

$$\sum_{j \in \mathcal{J}(i)} x_{ij}^t \le C_i^t \qquad \forall i \in \mathcal{I}, \tag{1c}$$

$$\sum_{i \in \mathcal{I}(j)} x_{ij}^t \le \bar{D}_j - Q_j^t + d_j^t \qquad \forall j \in \mathcal{J}. \tag{1d}$$

Constraints (1a) and (1b) limit the number of additional beds that can be staffed by the physical capacity at each unit ($\bar{D}'_j$ less the currently occupied beds Q'^t_j) and the total available staff (ν^t). Constraints (1c) state that no more than the number of ready patients is transferred out of each ICU. Finally, Constraints (1d) state that a DSU accepts no more patients than its available beds ($\bar{D}_j$ less the currently occupied baseline capacity beds Q_j^t) plus any additional beds available through flexible staffing.

(iii) *Exogenous information:* After decisions have been made, the following information is revealed: A_i^t for all $i \in \mathcal{I}$ representing the new patient arrivals at each ICU; $f_i^t \in [0,1]$ representing the fraction of current patients ready to be transferred; ℓ_j^t and $\ell'^t_j \in [0,1]$ representing the fraction of current patients at the regularly staffed and additionally staffed beds to be discharged; and B_j^t representing the external demand for beds at DSU j. In our case, flexible staffing beds are only used to accommodate ICU transfers and not external demand. Finally, the value $f_\nu^t \in [0,1]$ is revealed, representing the fraction of on-call personnel available for flexible staffing of additional downstream beds in the upcoming time period.

(iv) *Transition function:* Let $\mathcal{X}_t(S_t) = (\mathbf{d}^t, \mathbf{x}^t)$ represent the decisions taken at time period t given state S_t. After decisions have been made and the exogenous information described above has been observed, the system transitions to the next state $S_{t+1} = (\mathbf{P}^{t+1}, \mathbf{C}^{t+1}, \mathbf{Q}^{t+1}, \mathbf{Q}'^{t+1}, \nu^{t+1})$ as follows:

$$P_i^{t+1} = \left(P_i^t - \sum_{j \in \mathcal{J}(i)} x_{ij}^t + A_i^t\right) \wedge \bar{C}_i \qquad \forall i \in \mathcal{I}, \tag{2a}$$

$$C_i^{t+1} = f_i^t P_i^{t+1} \qquad \forall i \in \mathcal{I}, \tag{2b}$$

$$Q_j^{t+1} = Q_j^t + \left(\sum_{i \in \mathcal{I}(j)} x_{ij}^t - d_j^t \right) - \ell_j^t Q_j^t + \tilde{B}_j^t \qquad \forall j \in \mathcal{J}, \tag{2c}$$

$$\tilde{B}_j^t = \left(\bar{D}_j - \sum_{i \in \mathcal{I}(j)} x_{ij}^t + d_j^t \right) \wedge B_j^t \qquad \forall j \in \mathcal{J}, \tag{2d}$$

$$Q_j'^{t+1} = Q_j'^t + d_j^t - \ell_j'^t Q_j'^t \qquad \forall j \in \mathcal{J}, \tag{2e}$$

$$\nu^{t+1} = f_\nu^t \nu_{max}. \tag{2f}$$

Equation (2a) specifies that patient population at each ICU $i \in \mathcal{I}$ in period $t+1$ is updated by subtracting the transfers and adding the new arrivals, so as not to exceed capacity. Equation (2b) updates the transfer requests as a fraction of the number of patients at each unit. Equation (2c) updates the baseline capacity patient population at each downstream unit $j \in \mathcal{J}$ by adding the newly incoming ICU patients (less those occupying flexible staffing beds), subtracting the number of discharged patients, and adding the number of admitted external patient arrivals $\tilde{B}_j^t$. The number of external patients that can be admitted is determined by Equation (2d), i.e., not to exceed the total availability or the total demand. Next, Equation (2e) updates the patient population at additionally staffed beds by adding the newly transferred patients to the previous period's population and subtracting the discharged patients. Finally, Equation (2f) sets the available on-call staff as the revealed random fraction of the maximum.

(v) *Optimality criterion:* We define the contribution function $R(S_t, \mathcal{X}_t(S_t))$ as the reward from taking decisions $\mathcal{X}_t(S_t) = (\mathbf{d}^t, \mathbf{x}^t)$ when in state S_t. Specifically, we define the function as

$$R(S_t, \mathcal{X}_t(S_t)) = \sum_{i \in \mathcal{I}} U_i \left(\sum_{j \in \mathcal{J}(i)} x_{ij}^t \right) - \sum_{j \in \mathcal{J}} W_j(d_j^t),$$

to represent the total of a clinical benefit function, $U_i(\cdot)$, of ready ICU patients transferred out less the cost of flexible staffing. Note that we use a function of the total number of patients transferred (rather than the absolute total number of transferred patients) in order to allow for fair resource sharing among the ICUs and improved patient throughput. Details on our choice for the functions U_i follow in Section 3.3.

For a given discount factor η, Bellman's equations are:

$$V_t(S_t) = \max_{\mathcal{X}_t} \left\{ R(S_t, \mathcal{X}_t(S_t)) + \eta \sum_{s' \in \mathcal{S}} \mathbb{P}(S_{t+1} = s' \mid S_t, \mathcal{X}_t) V_{t+1}(s') \right\},$$

or, equivalently,

$$V_t(S_t) = \max_{\mathcal{X}_t} \left\{ R(S_t, \mathcal{X}_t(S_t)) + \eta \mathbb{E} V_{t+1}(S_{t+1} \mid S_t, \mathcal{X}_t) \right\}.$$

While the finite-horizon settings may be of interest for certain applications, we will focus on an infinite horizon (with possible variability due to the day of the week), seeking to capture the system dynamics on a global scale. Thus, removing the time dimension and letting $V(s) = \lim_{t\to\infty} V_t(S_t)$, the steady-state optimality equations are defined by

$$V(s) = \max_{\mathcal{X}} \left\{ R(s, \mathcal{X}(s)) + \eta \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, \mathcal{X}) V(s') \right\},$$

for each state $s \in \mathcal{S}$. For any realistic instances, the problem will have extremely large state and action spaces defined by the combinations of possible occupancies, ready patients, available staff, flexible staffing utilization, and patient allocation from ICU to downstream beds. As such, exact approaches are impractical. Hence, we focus on analyzing the stationary model counterpart from which we derive relative values of the downstream resources, capturing the global system dynamics. We then embed these values in intuitive and practical policies for the local (daily) operational decisions of flexible staffing and patient allocation. Thus, our approach seeks to capture enough of the system dynamics while maintaining computational tractability.

3.3. Contribution Function

In stating the optimality criterion in Section 3.2, we indicate the need to define a function measuring the number of ready patients allocated downstream that is “equitable” towards the ICUs, as many of the downstream resources are likely to be shared between multiple units. To fit these needs, we measure the clinical benefit of patients transferred out of a specific ICU with the function $U_i(\cdot)$. Specifically, an increasing and concave function ensures that transferring more patients from the ready list (up to the requested maximum) is always better, but the marginal increase in value is decreasing. Concavity further ensures the existence of a global maximum and is suitable in resource-sharing environments, as the system will favor distribution of resources among units as opposed to favoring a single unit with the highest benefit. Finally, concave functions are also consistent with risk-averse behavior (Kimball 1993, Clark and Oswald 1998), making benefit functions with these properties suitable for models involving patients’ health, i.e., while potentially blocking access to new arrivals, patients are safer at ICU due to the higher nurse-to-patient ratio. While we do not impose additional requirements on the functional form of $U_i(\cdot)$ in the model, distinction between the benefit functions of the various ICUs $i \in \mathcal{I}$ can be achieved by including parameters calibrated with the expected arrivals at each unit. Such parameters can further incentivize the model to prioritize ICUs according to expected patient influx and consequently reduce incoming patient blocking and diversions in the long run. This approach is illustrated in our numerical study.

The second term of the optimality criterion measures and subtracts the cost of flexible staffing of additional beds at the downstream units. Consistent with Assumption 1, we define $W(\cdot)$ to be a linear function of the number of additional beds staffed, i.e., we let $W_j(d_j) = \beta_j d_j$ for each unit $j \in \mathcal{J}$ and given cost coefficients $\beta_j > 0$. To balance the objective terms or prioritize patient transfers over staffing costs, the coefficients β_j may need to be appropriately scaled.

4. Stationary Model Analysis

While we do not explicitly solve the multi-period problem defined in Section 3.2, we analyze its stationary counterpart (relaxing the future expected value term) to derive practical insights, and easy-to-implement rules and policies. Omitting the time indexing and defining D_j and D'_j as the (expected) number of available baseline and flexible staffing beds, respectively, we summarize the decisions by the following model:

$$\max \sum_{i\in\mathcal{I}} U_i(\mathbf{x}) - \sum_{j\in\mathcal{J}} W_j(\mathbf{d}) \tag{3a}$$

$$\text{s.t. } d_j \le D'_j \qquad \forall j \in \mathcal{J}, \tag{3b}$$

$$\sum_{j\in\mathcal{J}} d_j \le \nu, \tag{3c}$$

$$\sum_{j\in\mathcal{J}(i)} x_{ij} \le C_i \qquad \forall i \in \mathcal{I}, \tag{3d}$$

$$\sum_{i\in\mathcal{I}(j)} x_{ij} \le D_j + d_j \qquad \forall j \in \mathcal{J}, \tag{3e}$$

$$\mathbf{d}, \mathbf{x} \in \mathbb{Z}_+. \tag{3f}$$

The objective (3a) maximizes the contribution function, consistent with the optimality criterion defined in Section 3.2. Note that for brevity, we write the two terms of the objective as functions of the decision variables $\mathbf{x}$ and $\mathbf{d}$. Constraints (3b)-(3c) limit the number of beds available through flexible staffing, both by the number of physical beds and the available personnel. Constraints (3d) bound the transfers out of ICUs by the transfer requests for medically ready patients. Similarly, Constraints (3e) bound the total number of patients admitted at downstream units by their respective number of available beds.

A solution to model (3) gives the steady-state rates of flexible staffing at downstream units and patient allocation from ICUs to wards. As it is a stationary model, it requires time-independent parameter estimates for the number of available personnel for flexible staffing (ν), number of available baseline and flexible staffing beds in each downstream unit (D_j and D'_j), and number of ready patients in each ICU (C_i).

Nevertheless, a solution to (3) could be used to guide daily decisions by determining the best assignment of flexible staffing beds and patient allocation using the current values of staff (ν^t), number of available beds in each downstream unit (D^t_j and D'^t_j), and number of ready patients in each ICU (C_i^t). As it is a non-linear optimization model with integer variables, it can be reformulated and solved with integer programming methods (e.g., through a commercial solver). We provide details on the reformulation in the e-companion and use the solution of (3) as a comparator in our numerical study. We call the policy of solving model (3) at each decision epoch with a commercial solver the *static policy.* A major limitation of this policy is that it requires all parameters ($\nu^t, D_j^t, D_j'^t, C_i^t$) to be known before a solution can be obtained. While it may be easy to check downstream bed availability, upstream information on the number of ready patients from all units may be more difficult to obtain due to different timing of rounds, processing of discharges, and the likely frequent change in health status of ICU patients. Thus, any delays or disruptions can easily render a solution infeasible.

4.1. Subproblem Decomposition

Instead of solving the primal model (3), we propose a dual approach that allows us to interpret the values of the downstream bed resources. Note that model (3) is a non-linear integer program, whose linear programming (LP) relaxation is a convex program. Because the dual of a non-linear integer program (or its integer reformulation or convex relaxation) is generally not computationally tractable (Geoffrion 1971), we first decompose the relaxation of model (3) into an upstream and downstream subproblems and then analyze the LP dual of the downstream subsystem only. We subsequently embed the downstream dual values in an operational decision making policy. Decomposing the problem allows us to analyze the downstream subsystem of staffing and routing decisions and obtain interpretable duals of the individual resources (referred to as the DSU bed values). The resulting policy requires less information than the static policy (the total number of ready ICU patients C_i is not needed), does not require frequent optimization, and is robust to disruptions and delays, i.e., decisions can be made on a rolling basis as beds are requested using only the number of available staff and the number of available downstream beds.

Given that the clinical benefit function $U_i(\cdot)$ associated with each ICU is a function of the total number of patients transferred out (as discussed in Section 3.3), regardless of their particular routing downstream, we reformulate model (3) by introducing variables $\mathbf{y} = \{y_i \,|\, i \in \mathcal{I}, 0 \leq y_i \leq C_i\}$ to represent the total patient flow out of unit i. Further, let $\mathbf{x} = \{x_{ij} \,|\, i \in \mathcal{I}, j \in \mathcal{J}(i)\}$ denote the flow from ICU to baseline capacity beds downstream and let $\mathbf{z} = \{z_{ij} \,|\, i \in \mathcal{I}, j \in \mathcal{J}(i)\}$ denote the flow from ICU to the additional (flexible staffing) beds. Note that this defines $d_j = \sum_{i \in \mathcal{I}(j)} z_{ij}$, i.e., the total number of incoming ICU patients allocated to the additional (flexible staffing) beds at unit j. Thus, the relaxed model (3) can be equivalently restated as

$$\max \sum_{i \in \mathcal{I}} U_i(y_i) - \sum_{j \in \mathcal{J}} \beta_j \sum_{i \in \mathcal{I}(j)} z_{ij} \tag{4a}$$

$$
\begin{aligned}
\text{s.t.}\ & \sum_{j\in\mathcal{J}(i)} (x_{ij}+z_{ij}) = y_i && \forall i \in \mathcal{I}, && \text{(4b)}\\
& y_i \le C_i && \forall i \in \mathcal{I}, && \text{(4c)}\\
& \sum_{i\in\mathcal{I}(j)} x_{ij} \le D_j && \forall j \in \mathcal{J}, && \text{(4d)}\\
& \sum_{i\in\mathcal{I}(j)} z_{ij} \le D'_j && \forall j \in \mathcal{J}, && \text{(4e)}\\
& \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}(i)} z_{ij} \le \nu, && && \text{(4f)}\\
& \mathbf{x},\mathbf{y},\mathbf{z} \ge 0. && && \text{(4g)}
\end{aligned}
$$

Seeking to decompose the model in the $\mathbf{x},\mathbf{y}$, and $\mathbf{z}$ variables, we introduce Lagrange multipliers $\boldsymbol{\lambda} = \{\lambda_i \mid i \in \mathcal{I}\}$ only for the coupling constraints (4b). We obtain

$$
\begin{aligned}
L(\mathbf{x},\mathbf{y},\mathbf{z},\boldsymbol{\lambda}) &= \sum_{i\in\mathcal{I}} U_i(y_i) - \sum_{j\in\mathcal{J}} \beta_j \sum_{i\in\mathcal{I}(j)} z_{ij} + \sum_{i\in\mathcal{I}} \lambda_i \left(\sum_{j\in\mathcal{J}(i)} (x_{ij}+z_{ij}) - y_i \right)\\
&= \sum_{i\in\mathcal{I}} (U_i(y_i) - \lambda_i y_i) + \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}(i)} (\lambda_i x_{ij} + (\lambda_i - \beta_j) z_{ij}).
\end{aligned}
$$

Note that $L(\mathbf{x},\mathbf{y},\mathbf{z},\boldsymbol{\lambda})$ is separable in the variables $\mathbf{y}$ and $(\mathbf{x},\mathbf{z})$. The dual function then is $q_1(\boldsymbol{\lambda}) + q_2(\boldsymbol{\lambda})$, where

$$
q_1(\boldsymbol{\lambda}) = \max_{0\le\mathbf{y}\le\mathbf{C}} \left\{ \sum_{i\in\mathcal{I}} (U_i(y_i) - \lambda_i y_i) \right\}, q_2(\boldsymbol{\lambda}) = \max_{\mathbf{x},\mathbf{z}\ge 0} \left\{ \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}(i)} (\lambda_i x_{ij} + (\lambda_i - \beta_j) z_{ij}) \;\middle|\; \text{(4d), (4e), (4f)} \right\}.
$$

The dual function returns the optimal multipliers $\boldsymbol{\lambda}$ which we later use in deriving staffing and routing policies. This representation allows us to decompose model (4) into the following subproblems:

$$
\begin{aligned}
\max\ & \sum_{i\in\mathcal{I}} (U_i(y_i) - \lambda_i y_i) && \text{(5a)}\\
\text{s.t.}\ & 0 \le y_i \le C_i \quad \forall i \in \mathcal{I}, && \text{(5b)}
\end{aligned}
$$

and

$$
\begin{aligned}
\max\ & \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}(i)} (\lambda_i x_{ij} + (\lambda_i - \beta_j) z_{ij}) && && \text{(6a)}\\
\text{s.t.}\ & \sum_{i\in\mathcal{I}(j)} x_{ij} \le D_j && \forall j \in \mathcal{J}, && \text{(6b)}\\
& \sum_{i\in\mathcal{I}(j)} z_{ij} \le D'_j && \forall j \in \mathcal{J}, && \text{(6c)}\\
& \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}(i)} z_{ij} \le \nu, && && \text{(6d)}\\
& \mathbf{x},\mathbf{z} \ge 0. && && \text{(6e)}
\end{aligned}
$$

Note that the coefficients $\boldsymbol{\lambda}$ can be obtained by solving the resulting dual optimization problem

$$
\min_{\boldsymbol{\lambda}} q_1(\boldsymbol{\lambda}) + q_2(\boldsymbol{\lambda}). \tag{7}
$$

Here, the ICU subproblems (5) determines how many patients to transfer out, regardless of their routing, and the downstream subproblem (6) determines how to staff additional beds and route patients. This decomposition results in $|\mathcal{I}|$ individual ICU subproblems, each of which consists of maximizing a concave function over a closed interval, and a single network subproblem, routing the flows of ICU patients to the downstream resources consisting of baseline capacity beds and additional (flexible staffing) beds. Note that the Lagrange multipliers $\boldsymbol{\lambda}$ can be found numerically using a subgradient algorithm (see the e-companion).

Proposition 4.1 *There exists a vector* $\boldsymbol{\lambda}$ *such that the solutions* $\mathbf{y}$ *to the subproblems* (5) *with* $y_i = \sum_{j\in\mathcal{J}(i)}(x_{ij}+z_{ij})$ *are optimal for the network subproblem* (6)*. Furthermore,* $(\mathbf{x},\mathbf{y},\mathbf{z})$ *are optimal for the system model* (4)*.*

4.2. Downstream Capacity Planning and Policy Derivation

Seeking to derive relative values of the downstream resources in order to guide the decisions in the two stages, we next focus on analyzing the network subproblem (6). Note that (6) is an LP whose dual is given by

$$\min \sum_{j\in\mathcal{J}}(D_j u_j + D'_j v_j) + \nu w \tag{8a}$$

$$\text{s.t. } u_j \geq \lambda_i \qquad \forall i \in \mathcal{I}, j \in \mathcal{J}(i), \tag{8b}$$

$$v_j + w \geq \lambda_i - \beta_j \qquad \forall i \in \mathcal{I}, j \in \mathcal{J}(i), \tag{8c}$$

$$\mathbf{u}, \mathbf{v}, w, \geq 0, \tag{8d}$$

where $\mathbf{u} = \{u_j \mid j \in \mathcal{J}\}$ and $\mathbf{v} = \{v_j \mid j \in \mathcal{J}\}$ represent the shadow prices on the baseline and flexible staffing beds, respectively, and w represents the shadow price of additional personnel. Thus, by knowing the vector of Lagrange multipliers $\boldsymbol{\lambda}$ and solving the dual model (8), the bed manager can obtain the relative values of beds at the various downstream units and how they contribute to the system objective of maximizing clinical benefit less staffing costs. The values of $\mathbf{u}$ and $\mathbf{v}$ can be obtained in polynomial time without explicitly solving the LP (see details in the e-companion). Thus, using the information on the values of beds at the various downstream units, we propose a weighted randomized acuity-based policy for guiding the first- and second-stage decisions (Algorithm 1).

The acuity-based policy uses the values of the DSU beds together with the current occupancies at units in determining the flexible staffing and routing of patients. The motivation is to choose from the available beds in a strategic way, based on values derived from the stationary model which uses both the network topology and the expected arrivals encoded in the clinical benefit function. In the first stage, the policy prioritizes beds with high value v_j^* at DSUs with more

```
Algorithm 1: Acuity-based policy
  Input: ICU-DSU network current state and values of $\boldsymbol{\lambda}$, $\mathbf{u}^*$, and $\mathbf{v}^*$ in the steady state
  Output: Staffing and allocation assignments
  While on-call personnel is available or a budget threshold is not reached, choose a unit $j$ with
    probability $\frac{v_j^* \bar{D}'_j}{\sum_{k \in \mathcal{J}} v_k^* \bar{D}'_k}$ and add a flexible staffing bed.
  For a given ICU $i \in \mathcal{I}$, choose a unit $j \in \mathcal{J}(i)$ with probability $\frac{D_j / u_j^*}{\sum_{k \in \mathcal{J}(i)} D_k / u_k^*}$ while
    $\sum_{k \in \mathcal{J}(i)} D_k > 0$. If $\sum_{k \in \mathcal{J}(i)} D_k = 0$, choose a unit $j \in \mathcal{J}(i)$ with probability $\frac{D'_j / u_j^*}{\sum_{k \in \mathcal{J}(i)} D'_k / u_k^*}$ while
    $\sum_{k \in \mathcal{J}(i)} D'_k > 0$. Allocate a patient from $i$ to the chosen unit $j$.
```

available unstaffed beds $\bar{D}'_j$. In the second stage, the policy prioritizes beds with low relative value (by taking the inverse of u_j^*), saving the higher-value beds for future transfers. The goal is to (i) ensure "high-value" beds (as measured by the optimal dual variable values) are staffed and available and (ii) route patients in a "fair" manner by considering current occupancies at units. The global dynamics captured refer to the network topology with resource sharing and the expected arrivals at ICUs (represented in the stationary model), while the local dynamics captured are the current bed occupancies and staff availability (represented in the daily decision-making policy). Thus, the policy aims to improve efficiency by ensuring bed availability to patients in need, either incoming to ICUs or seeking transfer downstream.

Our acuity-based approach is inspired by Mandelbaum et al. (2012), who use a similar randomized-most-idle (RMI) policy in determining transfers from an emergency department to wards, corresponding to the second-stage decisions in our case. However, their policy considers equal values of beds, i.e., $u_j = 1$ for all $j \in \mathcal{J}$. Their allocation policy measures fairness by "idleness ratios", while ours can be regarded as a "fair" policy with "weighted idleness ratios". Moreover, while they do not consider flexible staffing, a generalized version of the RMI policy incorporating the first-stage decisions is obtained by setting $v_j = 1$ and staffing beds at unit j with probability $\frac{\bar{D}'_j}{\sum_{k \in \mathcal{J}} \bar{D}'_k}$. This will be used in the numerical experiments to compare with our "weighted" policy.

We note that calculating the DSU bed values requires all the parameters in the stationary model, i.e., expected number of ICU transfer requests, expected DSU bed availabilities, and expected staff availability. The bed values may be periodically recalibrated, if changes in expected demand are observed. Once the DSU bed values are known, the policy requires only the current number of available beds in the downstream units. Hence, it can be applied either with full knowledge of the number of ready patients from all ICUs or with partial knowledge as transfer requests arrive (i.e., on a rolling basis).

While a deterministic policy is more intuitive, the probabilistic policy is advantageous in a stochastic environment where transfer decisions need to be made before information on incoming patients is available. A probabilistic policy is also preferable when seeking to achieve fairness in the number of patients routed to the various downstream units. The proposed acuity-based policy can be implemented in a simple decision support tool that chooses units with the appropriate probabilities, i.e., through a software using random sampling. In particular, in the first stage, a nurse is assigned to a unit chosen from the available pool based on sampling with assigned probabilities. Similarly, in the second stage, a patient is assigned to a unit chosen from the available feasible units based on random sampling with assigned probabilities as above.

5. Numerical Study

We conduct several numerical studies that demonstrate the performance of the proposed acuity-based policy. We first offer a basic illustrative example and later analyze a large-scale simulation calibrated with historic hospital data. In running the subgradient algorithm (see the e-companion for details) throughout all instances, we set the initial $\boldsymbol{\lambda}^{(0)} = 0.1$, the step size $\alpha^{(k)} = \frac{0.01}{\sqrt{k}}$ and maximum number of iterations $K = 1000$. Classical convergence results show that for the diminishing step size and step length rules, the algorithm is guaranteed to converge to the optimal value (Bertsekas 1999). Furthermore, while we do not use the optimal Polyak step size, as it requires knowledge of the optimal function value, we experimentally calibrate the step size and number of iterations for our instances. Specifically, our computational experiments showed no significant improvement beyond 1000 iterations of the algorithm.

5.1. Illustrative Example

We first consider a simple network and a single decision period. The network consists of two ICUs and three downstream units, so that ICU A is connected to downstream units 2 and 3 and ICU B is connected to all downstream units (see Figure 3). Note that even this simple network is not amenable to queuing analyses in the literature. We assume that the number of ready to transfer patients is 7 for ICU A and 5 for ICU B. We further assume that $U_A(y) = 1 - \exp(-0.1y)$ and $U_B(y) = 1 - \exp(-0.01y)$. Note that functions of the form $U_i(y) = 1 - \exp(-a_i y)$ are concave and range between 0 and 1. We assume that the coefficient a_i is proportional to the expected arrivals at unit i. In this example, ICU A has a higher expected patient influx and thus $U_A(y) \geq U_B(y)$. We let the downstream capacities be $\bar{D}_1 = 15, \bar{D}_2 = 25$, and $\bar{D}_3 = 10$ for baseline (regularly staffed) beds, and $\bar{D}'_1 = 2, \bar{D}'_2 = 3$, and $\bar{D}'_3 = 2$ for additional beds (available for contingent staffing). We set the flexible staffing cost of $\beta_j = 0.001$ for all $j = 1, 2, 3$. Note that while we choose identical cost of flexible staffing for simplicity, the model does not preclude us from having different β_j for the different units j. Finally, there are sufficient on-call personnel to staff five additional beds throughout the downstream units, i.e., $\nu = 5$.

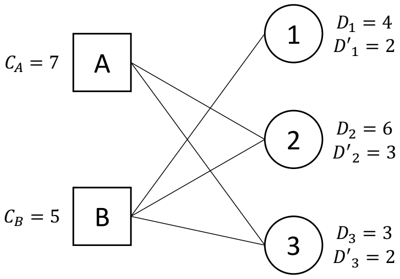


Figure 3 **Illustrative example network with two ICUs (A and B on the left) and three downstream units (1, 2, and 3 on the right).**

For the stationary model, we consider downstream occupancy of $\approx 75\%$ so that the baseline capacity beds available to accept patients are $D_1 = 4, D_2 = 6$, and $D_3 = 3$. We assume that all beds beyond baseline capacity are unoccupied and available for flexible staffing (i.e., $D'_j = \bar{D}'_j$ for all $j = 1, 2, 3$). The subgradient algorithm with the above input returns coefficients $\lambda_A = 0.030$ and $\lambda_B = 0.009$. The optimal dual values of the downstream beds are $u_2 = u_3 > u_1$ and $v_2 = v_3 > v_1$ (see Table 2). Those can be obtained by using Algorithm 4 in the e-companion.

	A	B
λ_i	0.030	0.009
w^*	0.008	

(a)

	1	2	3
u_j^*	0.009	0.030	0.030
v_j^*	0	0.021	0.021

(b)

Table 2 **Summary of Lagrange multipliers and dual values for the illustrative example.**

According to the generalized RMI policy, additional beds will be staffed at downstream units in order $[2, 1, 3]$ or $[2, 3, 1]$ (see probabilities in Table 3). Here, unit 2 is prioritized due to its highest number of available beds (3/7) followed by units 1 and 3 which each have 2/7 beds available. The acuity-based policy would instead select downstream units in order $[2, 3]$ for flexible staffing. The policy thus captures the intuition that DSU 1 can only accommodate patients from ICU B, so extra staffed beds there are likely to be needed less. In the allocation stage, the generalized RMI policy would transfer patients out of ICU A to downstream units by priority order of $[2, 3]$, while patients out of ICU B would be routed to units in order of $[2, 1, 3]$. The acuity-based policy would similarly route patients out of ICU A to downstream units by priority order $[2, 3]$, however, patients out of ICU B would be transferred to units by priority order $[1, 2, 3]$. Note that DSU 1 can only be utilized by patients from ICU B, so giving it higher priority means that beds at DSUs 2 and 3 can remain available for patients out of ICU A, which in this example also has higher expected demand.

5.2. Large-Scale Instances

Next, we turn our attention to testing the proposed policy in a more realistic setting using a simulation calibrated with historic hospital transfer data.

	1	2	3
(i) staffing	0.286	0.429	0.286
(ii) allocation			
A	0	0.667	0.333
B	0.308	0.462	0.231

(a) Generalized RMI

	1	2	3
(i) staffing	0	0.600	0.400
(ii) allocation			
A	0	0.667	0.333
B	0.597	0.269	0.134

(b) Acuity-based

Table 3 Probabilities of flexible staffing per unit and patient transfer allocation from ICU do downstream units.

5.2.1. Data and Network Topology We use hospital data containing anonymized patient admissions and discharge timestamps during an 80-months period from a major academic hospital. The dataset contains 19,882 patient IDs and 91,878 transfer records. We consider the units labeled *ICU* and aggregate the units labeled *Ward*, *Clinic*, or *Stepdown* into our layer of downstream units. Based on the patient ID number, we calculate the number of transfers between each pair of units and remove the one ICU which has no historic transfers to any of the downstream units and the 13 downstream units with no historic transfers from ICU. We infer the capacities of units by calculating the maximum concurrent occupancy and remove the six downstream units with inferred capacity of less than 10 beds. Thus, we consider 17 ICUs and 28 downstream units with capacities ranging between 20–39 for ICUs and 10–59 for DSUs.

We infer four different network topologies, ensuring that an ICU is connected to at least one downstream unit. A summary of the number of links and average node degree of the networks is provided in Table 4 (also see the e-companion for graphic illustrations). Specifically, network N28 builds links based on the frequency of transfers between units. For a given ICU i, we order the downstream units in descending order of number of transfers they accepted from i. We then build links going down the list until we have the DSUs that accepted 30% of ICU i's transfer volume. Network N37 builds links in a similar manner, adding links from the most frequently utilized units until 40% of the total volume is reached. Network N60 builds a link (i, j) if there were at least 100 historic transfers from ICU i to DSU j. Finally, network N28* was not created using historic transfers, but rather with an integer programming model to build the minimum number of links that connect all ICU to at least one DSU and establish a feasible flow for all patients while seeking

Network	Number of links	Average ICU node degree	Average DSU node degree
N28	28	1.647	1.000
N28*	28	1.647	1.000
N37	37	2.176	1.321
N60	28	3.529	2.143

Table 4 Summary of network topologies used for the data-driven instances.

a target occupancy of $\leq 80\%$ downstream. Note that none of the networks are trees as they are not connected. Furthermore, N28, N37, and N60 all contain cycles. Network N28* is acyclic and hence represents a forest of trees.

5.2.2. Implementation Similar to the illustrative example, we use the exponential functions $U_i(y) = 1 - \exp(-a_i y)$ and set a_i proportional to the mean arrival rate per day calculated from the data. Specifically, if μ_k^{at} represents the arrival rate at unit k on day t, we set $a_i^t := 0.1\mu_k^{at}$ for each day of the week so that $t = 1$ represents Monday and $t = 7$ represents Sunday. The scaling ensures adequate differentiation in the range of the U_i functions. We further set the cost of flexible staffing to $\beta_j = 0.001$ for the units j with additional beds available. Note that the approach will use additional beds if the value λ from admitting an ICU patient at a downstream bed is greater than the cost of the bed β. Given that we start with small initial $\boldsymbol{\lambda}$ and use diminishing step size, we keep β close to zero as well. If additional beds are available, we assume their number is 30% (rounded down to the nearest integer) of the inferred base capacity at each unit. Staffing is assumed to be at $\nu = 100$ (note that $\sum_{j\in\mathcal{J}} D'_j = 190$), so that not all extra beds can be staffed, and location of the flexible staffing beds is important.

We first calculate the values λ for each ICU and the bed values u, v for each DSU separately for each day of the week (Monday–Sunday) based on expected values of the input parameters. We then simulate patient arrivals and discharges during a 26-week time horizon. We set the initial occupancy (beginning of Monday) for each ICU to be a random value between 10–100% of capacity and for each downstream unit a random value between 50–100% of capacity. Additionally, we set the initial bed occupancy at the additionally staffed beds at downstream units to be a random value between 0–100% of the number of beds. To aggregate the patient arrival distributions from the historic data, we estimate the arrivals in each particular day of the week and calculate the average, μ_k^{at} for each day of the week t and each unit $k \in \mathcal{I} \cup \mathcal{J}$. We sample arrivals from $\texttt{Poisson}(3\mu_k^{at})$ for each ICU and DSU. Here, we augment the mean parameter to simulate a high demand setting. We sample ICU transfer requests and DSU discharges from $\texttt{Poisson}(3\mu_k^{dt})$ based on historic discharge rates, where the average discharge rates μ_k^{dt} are calculated similar to the average arrival rates μ_k^{at}.

Note that calculating the DSU bed values requires all the parameters in the stationary model, i.e., expected number of ICU transfer requests, expected DSU bed availabilities, and expected staff availability. Once the DSU bed values are known, the policy requires only the current number of available beds in the downstream units. While we use full knowledge of all transfer requests in the simulation study, in order to have a meaningful comparison with the stationary policy which does require full information, the acuity-based policy can also be applied on a rolling basis as transfer requests arrive, i.e., only requiring current DSU bed availabilities. We note that bed values can be periodically recalibrated, if needed. For the simulation, we calculate and use DSU bed values for each day of the week (based on expected demand Monday–Sunday). While we use day of the week to distinguish bed values, finer discretizations may also be used, e.g., by hour of the day, simply by solving for the u_j and v_j values at the appropriate time intervals. For instance, Shi et al. (2016) show time-varying waiting times in emergency departments sensitive to time of the day discharges. Although more computationally intensive, we expect that a finer discretization would show similar or better performance to the daily discretization we use (see the last section of the Numerical Study for results). Due to the random aspect of the proposed policy (staff and patients are allocated to units with a given probability), we run the 26-week period simulation 100 times and report averaged results. At each decision epoch, flexible staffing is first determined based on the first-stage rule of the acuity-based policy. Next, in making the second-stage transfer decisions, the decision maker orders the ICUs in descending order by the expected arrival rate and transfers one patient from each unit in the list using the DSU probability according to the policy. Ready patients are updated and she iterates over the list until no more patients or no more available beds remain. Patients are transferred to baseline capacity beds first and to additionally opened beds second, after no more of the former are available.

We compare the performance of the acuity-based policy to three other policies: a deterministic policy, a static policy, and a generalized RMI policy. The deterministic heuristic policy seeks to staff (in the first stage) and "guard" (in the second stage) the most in-demand beds. In particular, for a given network, we define $\text{Priority}(j) := \dfrac{\deg(j)}{D_j}$ to determine how in-demand a given DSU is. The numerator represents the degree of the DSU node in the graph, i.e., the number of ICUs that are connected to the particular ward, while the denominator represents the number of available beds. In the first stage, we make staffing decision moving from highest priority to lowest. In the second stage, we make allocation decision moving from lowest priority to highest. For instance, if two units j and k have the same node degree but unit k has more available beds than unit j ($D_j \leq D_k$), then $\text{Priority}(j) \geq \text{Priority}(k)$, so unit j will be prioritized for flexible staffing while unit k will be prioritized for patient allocation. Conversely, if two units j and k have the same number of available beds but unit k is connected to more ICUs ($\deg(j) \leq \deg(k)$), then $\text{Priority}(j) \leq \text{Priority}(k)$, so unit k is prioritized for flexible staffing while unit j is prioritized for patient allocation.

The static policy, as described in Section 4, uses an optimization model at each stage and as such is a more computationally intensive option. Moreover, it requires information on all ready patients from all ICUs in order to obtain a feasible solution. As a result, it is more sensitive to disruptions and delays as patients cannot be trarnsferred on a rolling basis.

Finally, the generalized RMI policy, as introduced in Section 4.2, uses equal weights for all downstream beds and is analogous to the policy proposed by Mandelbaum et al. (2012) for patient transfers from a single emergency department to multiple general wards. Similar to the acuity-based policy, we run the 26-week period simulation for the generalized RMI policy 100 times and report averaged results. We furthermore follow the same ordering of ICUs (based on expected demand) and downstream beds (allocating to baseline before flexible staffing beds) as in running the acuity-based policy.

5.2.3. Results and Discussion Figure 4 offers a comparison in ICU lost arrivals (i.e., patients diverted to other facilities due to lack of available beds) for the four network topologies between the acuity-based and deterministic policies. The values shown are averages for each day of the week, measuring lost arrivals as percentage of total arrivals (recall that we simulate a high-demand scenario, hence the high percentages of patient diversions). We can see that clearly the acuity-based policy outperforms the deterministic policy by reducing the number of lost arrivals among all network topologies and days of the week. The acuity-based policy offers statistically significant reduction in the number of diverted patients in all cases.

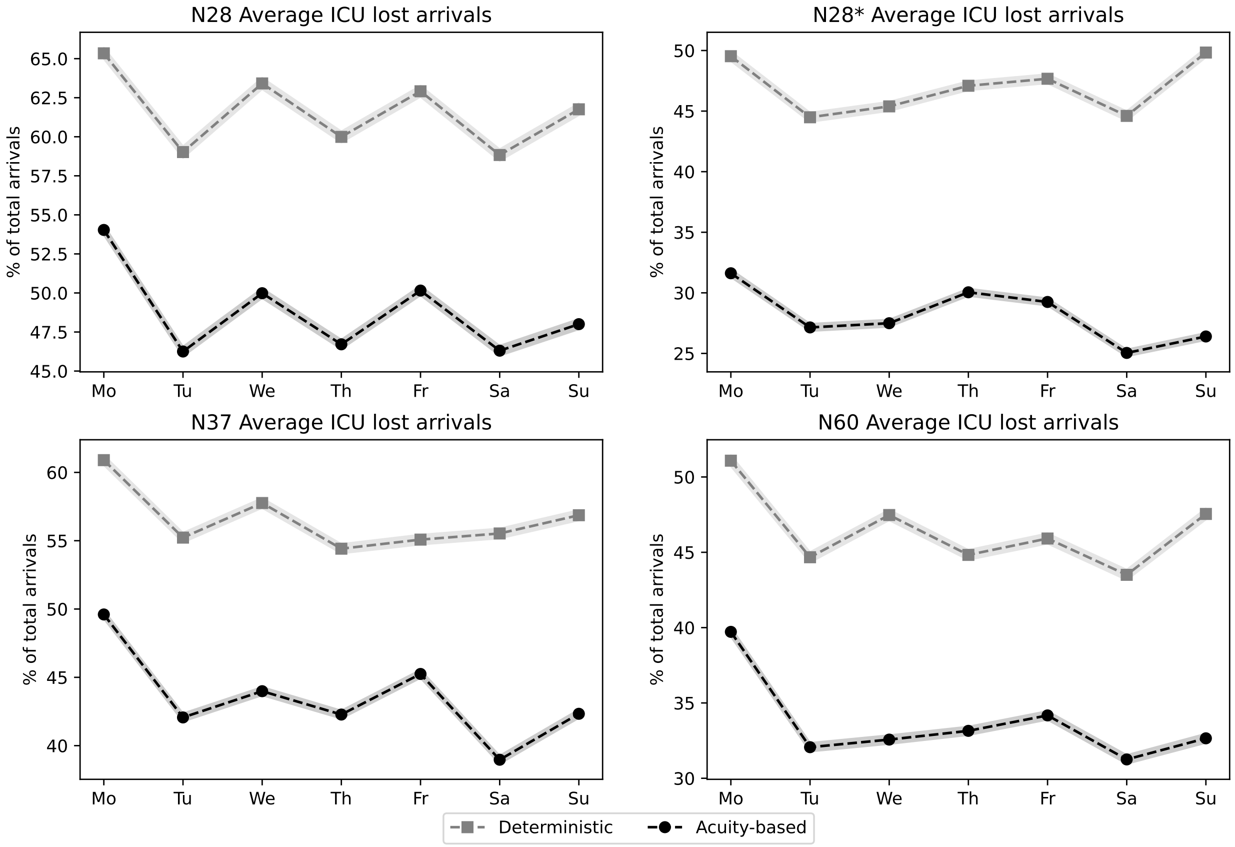


Figure 4 **Seven day patient flow simulation for four network topologies depicting average (among all ICU for each day) lost ICU arrivals (diverted patients) for the deterministic and acuity-based policies with 95% confidence intervals.**

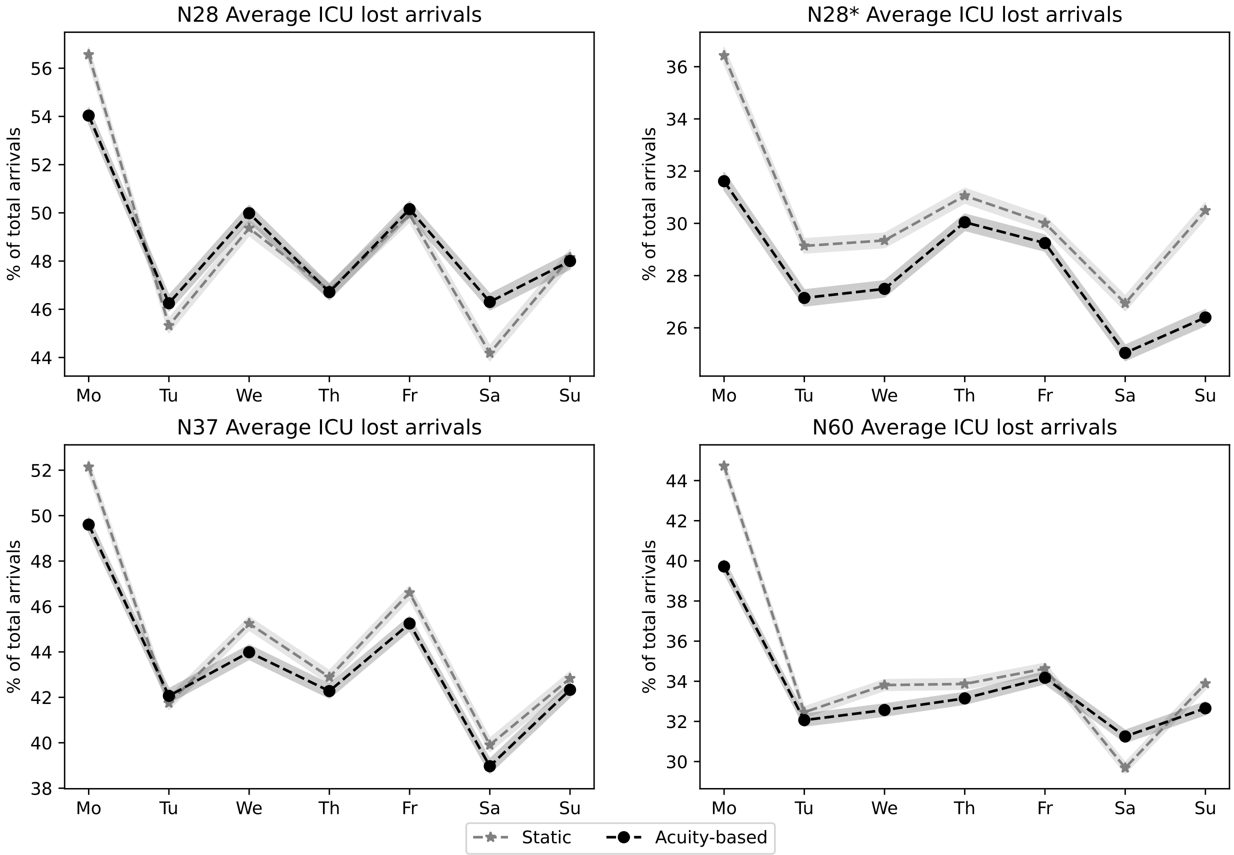


Figure 5 **Seven day patient flow simulation for four network topologies depicting average (among all ICU for each day) lost ICU arrivals (diverted patients) for the static and acuity-based policies with 95% confidence intervals.**

Figure 5 offers a similar comparison in ICU lost arrivals between the acuity-based and static policies. While this policy ensures that as many as possible of the ready patients are transferred at each period and takes into account the expected arrivals at ICUs as measured by the objective function, it makes decisions that are optimal in the current time period and do not consider the overall value of beds at different units, as captured by the DSU dual values. The acuity-based policy offers statistically significant reduction in the number of diverted patients in all days in N28*, all days except Tuesday in N37, and all days except Saturday in N60. The reduction for N28 is only significant for Monday. Thus, our proposed acuity-based policy offers an easy-to-implement and competitive approach to making flexible staffing and allocation decisions in complex ICU-downstream networks without requiting optimization. We emphasize that as a major practical advantage, the acuity-based policy does not require solving an integer program in each decision epoch as the static policy does (see the e-companion for details). Furthermore, the acuity-based policy can be applied with partial information, i.e., as ICU patients become ready, they can be transferred one by one using the allocation rules of Stage II. On the other hand, the static policy requires full information on all ready patients from all ICUs, making it more difficult to implement in large-scale hospital networks and sensitive to disruptions and/or changes in number of ready patients, available beds, or available personnel.

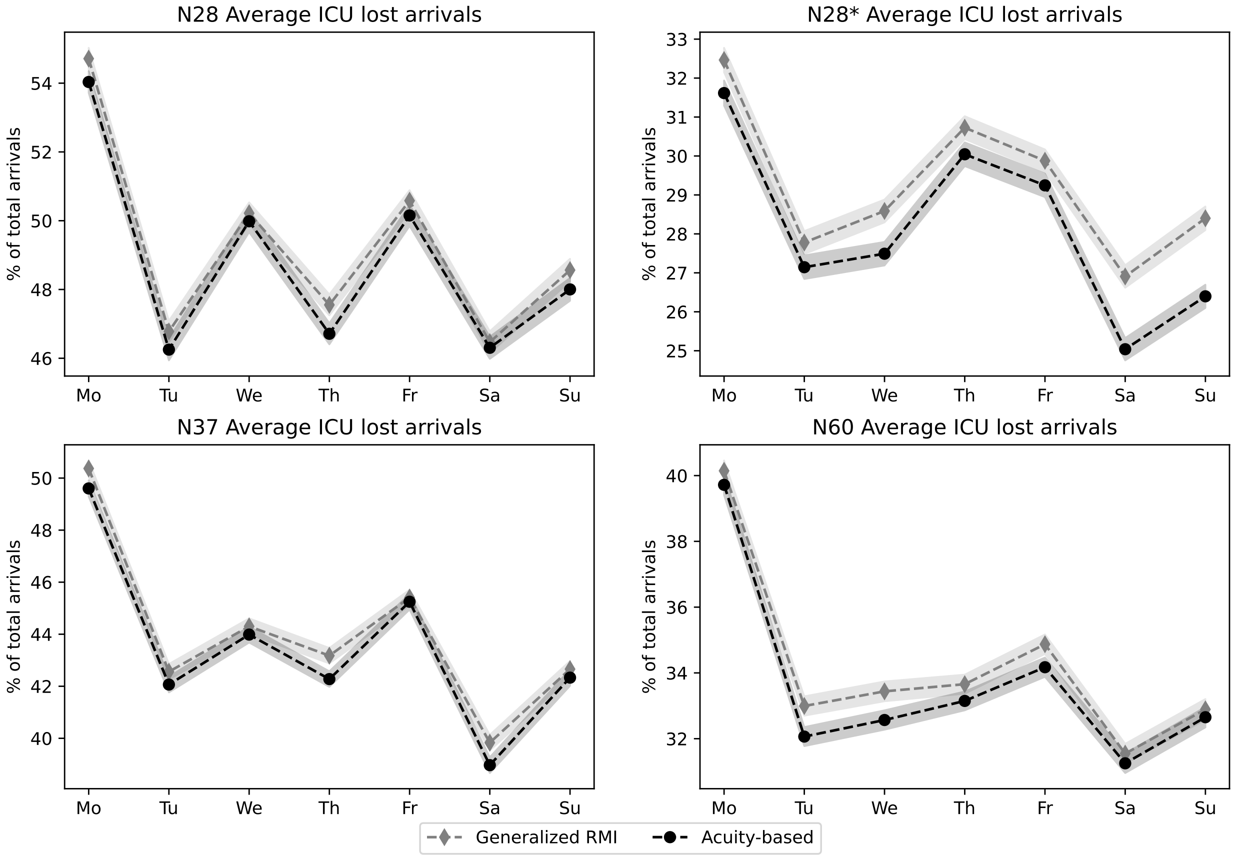


Figure 6 **Seven day patient flow simulation for four network topologies depicting average (among all ICU for each day) lost ICU arrivals (diverted patients) for the generalized RMI and acuity-based policies with 95% confidence intervals.**

Figure 6 similarly illustrates the ICU lost arrivals comparing the generalized RMI policy and the proposed acuity-based policy. The acuity-based policy offers statistically significant reduction in the number of diverted patients in four out of seven days in N28, N37, and N60, and in all seven days in N28*. Noticeably, the reduction is higher in networks N28* and N60 compared to networks N28 and N37. A rationale behind this observation is that N28 and N37 use a smaller number of DSUs, meaning that more downstream units are shared and thus, DSUs are less distinguishable from each other. In particular, the dual values of downstream beds are likely to be similar to each other, rendering the weighting in the policy less significant. Network N60, on the other hand, offers more differentiation among units due to the higher number of links and the higher number of DSUs available to accept incoming ICU patients. Finally, network N28* offers distinction between units

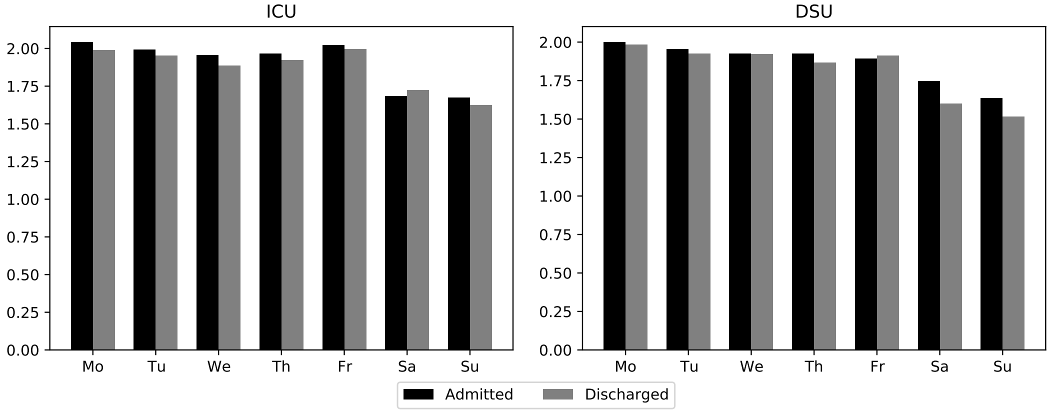


Figure 7 **Average number of admitted and discharged patients by unit type and day of the week.**

due to the algorithm used to create it, i.e., seeking equatable distribution of patient demand among units, resulting in more varied dual values of the downstream beds. As the generalized RMI policy is the most competitive to the acuity-based policy, a constructive future research direction is to further characterize network types for which one policy is preferable over the other.

While the proposed acuity-based policy is most valuable for high-demand settings, as illustrated by the computational experiments, sensitivity analyses for lower demand settings confirm that it outperforms the deterministic policy in all cases and remains competitive to the static and generalized RMI policies. Note that in cases where there is sufficient capacity to meet all demand, flexible staffing and strategic patient allocation have little to no benefit. Details on the results of the sensitivity analyses with lower demand settings are provided in the e-companion.

An interesting observation from the computational experiments is that the acuity-based policy has a more pronounced advantage on weekends. Turning our attention to the original data used to calibrate our simulation, we summarize the average number of patients admitted and discharged by unit type and day of the week in Figure 7. Overall, both admissions and discharges are lower on weekends relative to weekdays. We observe, however, that downstream units have notably more admissions than discharges on weekends. This pattern is likely to cause downstream congestions on weekends and, as shown in the sensitivity analyses, the acuity-based policy is especially valuable in congested systems by strategically staffing and utilizing downstream beds. While a more in-depth analysis is needed to fully understand why the acuity-based policy has a particular advantage on weekends, the observations from this work could be useful in future studies of the "weekend effect" in hospitals, i.e., the higher mortality in patients admitted on weekend (Pauls et al. 2017).

Finally, as a probabilistic policy is generally less interpretable and might face resistance to implementation, the probabilities calculated in the acuity-based policy can be used as priority values to assign bed staffing and patient allocation in a deterministic fashion. This approach is formalized in Algorithm 2. Figure 8 illustrates the absolute difference between means in ICU lost arrivals between the (probabilistic) acuity-based policy and the dual-based deterministic policy. The negative values for three out of the four network topologies indicate that the probabilistic acuity-based policy offers a statistically significant reduction in the percentage of ICU lost arrivals when compared to the dual-based deterministic one. While we recommend the probabilistic acuity-based policy as it outperforms the determinsitic one in most cases, the reductions are small, and thus a dual-based deterministic version using the previously derived dual values u_j^* and v_j^* is a viable alternative.

```
Algorithm 2: Dual-based policy
  Input: ICU-DSU network current state and values of $\boldsymbol{\lambda}$, $\mathbf{u}^*$, and $\mathbf{v}^*$ in the steady state
  Output: Staffing and allocation assignments
  While on-call personnel is available or a budget threshold is not reached, select a unit
    $j = \arg\max\left\{\frac{v_j^* \bar{D}'_j}{\sum_{k\in\mathcal{J}} v_k^* \bar{D}'_k}\right\}$ and add a flexible staffing bed.
  For a given ICU $i \in \mathcal{I}$, if $\sum_{k\in\mathcal{J}(i)} D_k > 0$, select a unit $j = \arg\max\left\{\frac{D_j/u_j^*}{\sum_{k\in\mathcal{J}(i)} D_k/u_k^*}\right\}$. If
    $\sum_{k\in\mathcal{J}(i)} D_k = 0$, select a unit $j = \arg\max\left\{\frac{D'_j/u_j^*}{\sum_{k\in\mathcal{J}(i)} D'_k/u_k^*}\right\}$ while $\sum_{k\in\mathcal{J}(i)} D'_k > 0$. Allocate a
    patient from $i$ to the selected unit $j$.
```

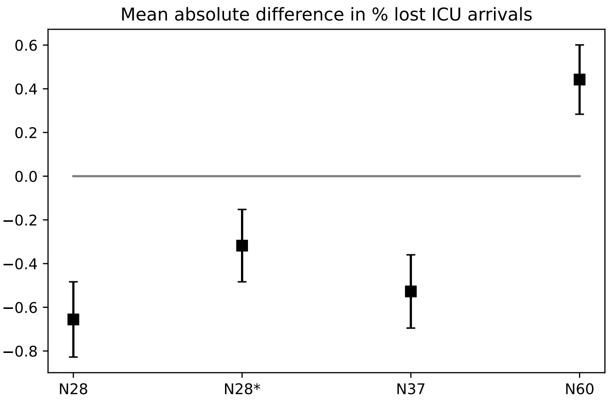


Figure 8 Difference between means of % lost ICU arrivals (relative to all ICU arrivals) in four network topologies for the (probabilistic) acuity-based policy and the dual-based deterministic policy with 95% confidence intervals.

5.3. Parameter Calibration

Seeking further performance improvement, we conduct several additional experiments in which we modify the frequency of decision making. In particular, the above experiments assume decisions are made once a day. We devise a scenario in which decisions are made twice a day, morning and afternoon. Here, we sample new arrival rates and find optimal λ and dual bed values for each half day and apply the acuity-based policy twice a day, i.e., patients are admitted and discharged twice a day. For the sake of comparison with the once-a-day decision making, we sum the total number of diverted patients per day (see Figure 9). We can see the the more frequent decision making offers advantages in all networks by reducing the number of diverted patients on average by 1.9%-4.2%. The reduction is statistically significant in all illustrated cases except for one instance.

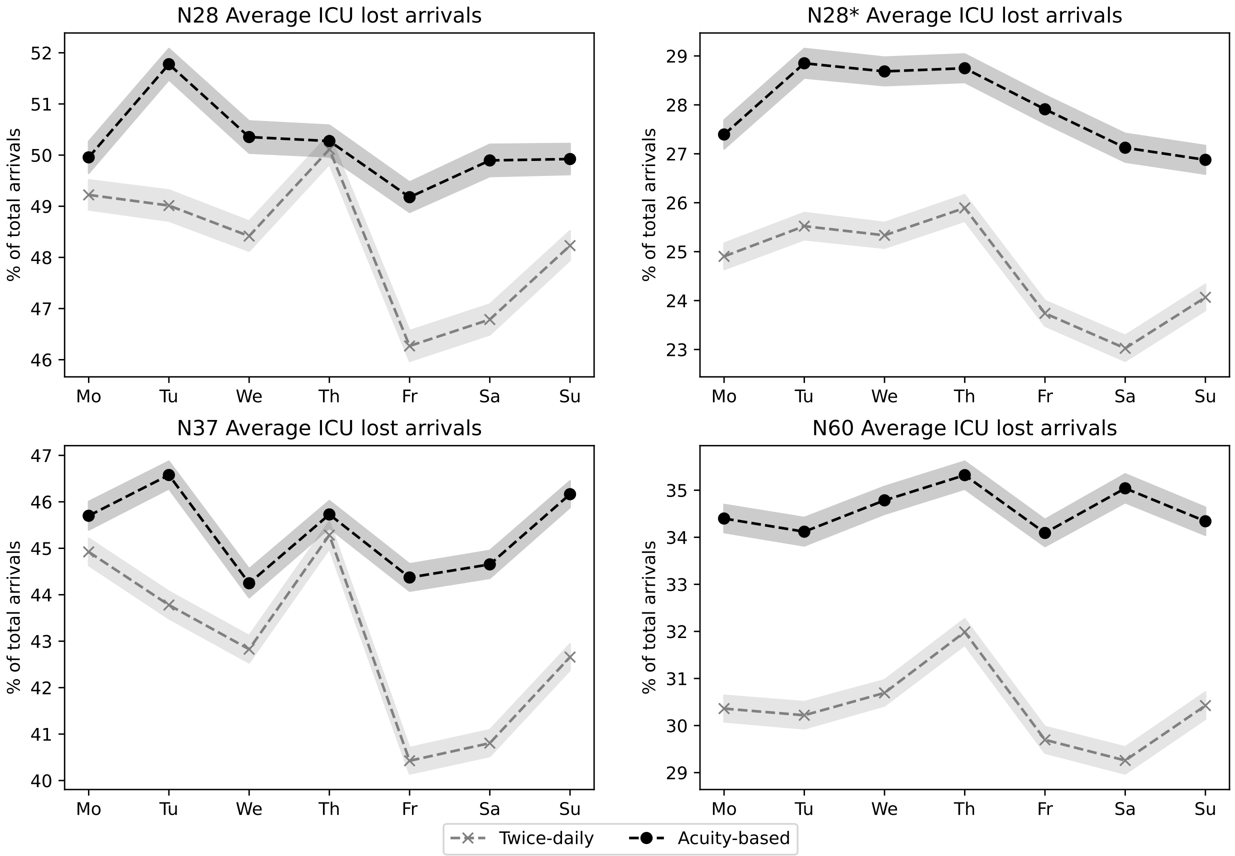


Figure 9 **Seven day patient flow simulation for four network topologies depicting average (among all ICU for each day) lost ICU arrivals (diverted patients) for the acuity-based policy applied twice daily and daily with 95% confidence intervals.**

Another practical consideration is the planning for on-call personnel available for flexible staffing. While the presented model assumes the budget is a known parameter (ν in constraint (3c)), determining an appropriate budget likely requires a whole-hospital model that considers multiple sub-systems competing for resources (e.g., funds allocation). The presented model focuses on the ICU-downstream network only, hence it does not capture the entire hospital's budgeting considerations. Nevertheless, the model and the resulting acuity-based policy can be used for conducting sensitivity analyses to budget fluctuations. Specifically, adding more flexible staffing beds will increase total utility as long as it is the bounding parameter. This will be beneficial up to the point when either physical capacity is reached or demand is less than the supply of downstream beds.

6. Conclusions

This paper focuses on the management of patient transfers from ICUs to downstream units with flexible staffing. The goal in studying this problem is to derive practical policies that can improve efficiency in the process and reduce unproductive occupancy and blocking of incoming ICU patients by ensuring ready patients are transferred on time and downstream resources are utilized strategically. We propose a dynamic multi-period model utilizing both baseline and flexible staffing at downstream units. Using analysis of its stationary counterpart, we propose a method to derive downstream bed values that we embed in a practical policy for determining both the staffing and allocation decisions. Our numerical studies, calibrated with historical hospital data, indicate that the acuity-based policy can reduce the number of ICU patient diversions when compared to a deterministic policy, a static period-by-period optimization policy, and a generalized randomized-most-idle policy in networks with various topologies. The proposed methodology is significant as it allows us to model complex and diverse networks (not limited to tree structures), which often arise in healthcare settings. Moreover, the demonstrated reduction in the number of ICU patient diversions is significant for both patient safety and outcomes as well as quality of service. Sensitivity analyses demonstrate that the proposed policy is most valuable in high-demand scenarios where strategic use of existing resources and flexible staffing is needed the most. In the extreme surge circumstances caused by the COVID-19 onslaught, there has been extensive discussion on ethical and fair assignment of resources, in particular, ICU beds (Robert et al. 2020, Vergano et al. 2020). Difficult decisions are unavoidable consequences of exceptional system strains. Approaches such as the one we propose may contribute to offset the delay of these very real decisions.

While we focus on seeking to improve efficiency in the transfer of ICU patients by measuring ICU clinical benefit of the total number of ready patients allocated to downstream units, other considerations of interest to hospital management can be embedded in future work extensions. For example, demand-driven discharge to units of intermediate care is sometimes indicated when incoming critical patients are in need of ICU beds (Chan et al. 2012, Hosseinifard et al. 2014). Thus, an alternative ICU objective may be to not only transfer ready patients before demand is realized, but to also determine threshold values of demand-driven discharge in the case of bed shortage. Furthermore, patient readmission that often results from premature discharge is a highly undesirable occurrence, with negative consequences both for the patient's safety and the hospital's perceived quality of service (Rosenberg and Watts 2000, Rosenberg et al. 2001). Discharge decisions with readmissions were studied by Chan et al. (2012) who incorporate a predictive model of the readmission risk. Thus, risk models of readmission and patient ranking can be embedded in future extensions of our decision model and policy derivation, seeking to minimize the number of discharged patients requiring ICU readmission. Further theoretical contributions can be made in exploring asymptotic optimality in specific network structures or parameter conditions.

Acknowledgments

This work is partially supported by National Science Foundation grants CMMI-1635301/1635410/1635642. We thank Professor M. Yasin Ulukus of Istanbul Technical University for his assistance.

References

D. Adelman. Price-directed control of a closed logistics queueing network. *Operations Research*, 55(6): 1022–1038, 2007.

D. Adelman and C. Barz. A price-directed heuristic for the economic lot scheduling problem. *IIE Transactions*, 46(12):1343–1356, 2014.

D. Adelman and G. L. Nemhauser. Price-directed control of remnant inventory systems. *Operations Research*, 47(6):889–898, 1999.

A. Arapostathis and G. Pang. Infinite horizon asymptotic average optimality for large-scale parallel server networks. *Stochastic Processes and their Applications*, 129(1):283–322, 2019.

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

R. Atar. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, 15(4):2606–2650, 2005.

R. Atar, A. Mandelbaum, and M. I. Reiman. Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, 14(3):1084–1134, 2004.

J. Bai, A. Fügener, J. Schoenfelder, and J. O. Brunner. Operations research in intensive care unit management: A literature review. *Health Care Management Science*, 21(1):1–24, 2018.

M. Barrett, M. Smith, A. Elixhauser, L. Honigman, and J. Pines. Utilization of intensive care services, 2011: HCUP Statistical Brief #185. *Agency for Healthcare Research and Quality, Rockville, MD*, 2014. URL `https://hcup-us.ahrq.gov/reports/statbriefs/sb185-Hospital-Intensive-Care-Units-2011.pdf`.

K. J. Bates. Floating as a reality: Helping nursing staff keep their heads above water. *Medsurg Nursing*, 22 (3):197–200, 2013.

R. Bekker, G. Koole, and D. Roubos. Flexible bed allocations for hospital wards. *Health Care Management Science*, 20(4):453–466, 2017.

R. E. Bellman. *Dynamic Programming.* Princeton University Press, 1957.

D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

D. Bertsimas and J. Niño-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

D. Bertsimas and J. Pauphilet. Hospital-wide inpatient flow optimization. *Optimization Online*, 2020.

C. Buyukkoc, P. Varaiya, and J. Walrand. The cμ rule revisited. *Advances in Applied Probability*, 17(1): 237–238, 1985.

L. T. Cardoso, C. M. Grion, T. Matsuo, E. H. Anami, I. A. Kauss, L. Seko, and A. M. Bonametti. Impact of delayed admission to intensive care units on mortality of critically ill patients: A cohort study. *Critical Care*, 15(1):1–8, 2011.

W. Chaboyer, L. Thalib, M. Foster, D. Elliott, R. Endacott, and B. Richards. The impact of an ICU liaison nurse on discharge delay in patients after prolonged ICU stay. *Anaesthesia and Intensive Care*, 34(1): 55–60, 2006.

D. B. Chalfin, S. Trzeciak, A. Likourezos, B. M. Baumann, and R. P. Dellinger. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine*, 35(6):1477–1483, 2007.

C. W. Chan, V. F. Farias, N. Bambos, and G. J. Escobar. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research*, 60(6):1323–1341, 2012.

C. W. Chan, J. Dong, and L. V. Green. Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research*, 65(2):469–495, 2017.

C. A. Chrusch, K. P. Olafson, P. M. McMillan, D. E. Roberts, and P. R. Gray. High occupancy increases the risk of early death or readmission after transfer from intensive care. *Critical Care Medicine*, 37(10): 2753–2758, 2009.

A. E. Clark and A. J. Oswald. Comparison-concave utility and following behaviour in social and economic settings. *Journal of Public Economics*, 70(1):133–155, 1998.

G. B. Dantzig and P. Wolfe. Decomposition principle for linear programs. *Operations Research*, 8(1):101–111, 1960.

A. M. de Bruin, R. Bekker, L. Van Zanten, and G. Koole. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178(1):23–43, 2010.

G. Dobson, H.-H. Lee, and E. Pinker. A model of ICU bumping. *Operations Research*, 58(6):1564–1576, 2010.

N. Gans and G. van Ryzin. Dynamic vehicle dispatching: Optimal heavy traffic performance and practical insights. *Operations Research*, 47(5):675–692, 1999.

A. M. Geoffrion. Duality in nonlinear programming: A simplified applications-oriented development. *SIAM Review*, 13(1):1–37, 1971.

L. V. Green. Capacity planning and management in hospitals. In *Operations Research and Health Care*, pages 15–41. Springer, 2005.

I. Gurvich and W. Whitt. Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research*, 34(2):363–396, 2009a.

I. Gurvich and W. Whitt. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing & Service Operations Management*, 11(2):237–253, 2009b.

R. Hall. Bed assignment and bed management. In *Handbook of Healthcare System Scheduling*, pages 177–200. Springer, 2012.

J. M. Harrison and A. Zeevi. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Operations Research*, 52(2):243–257, 2004.

R. Hendren. Five reasons nurses want to leave your hospital, 2011. URL `https://www.healthleadersmedia.com/nursing/5-reasons-nurses-want-leave-your-hospital`.

S. Z. Hosseinifard, B. Abbasi, and J. P. Minas. Intensive care unit discharge policies prior to treatment completion. *Operations Research for Health Care*, 3(3):168–175, 2014.

D. S. Kc and C. Terwiesch. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498, 2009.

D. S. Kc and C. Terwiesch. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*, 14(1):50–65, 2012.

S.-H. Kim, C. W. Chan, M. Olivares, and G. Escobar. ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science*, 61(1):19–38, 2014.

M. S. Kimball. Standard risk aversion. *Econometrica: Journal of the Econometric Society*, 61(3):589–611, 1993.

L. Kuntz, R. Mennicken, and S. Scholtes. Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science*, 61(4):754–771, 2014.

L. Landro. Rethinking the hospital for the next pandemic. *The Wall Street Journal*, 2020. URL `https://www.wsj.com/articles/rethinking-the-hospital-for-the-next-pandemic-11591652504`.

L. Leatherby, J. Keefe, L. Tompkins, C. Smart, and M. Conlen. 'There's no place for them to go': I.C.U. beds near capacity across U.S. *The New York Times*, December 2020. URL `https://www.nytimes.com/interactive/2020/12/09/us/covid-hospitals-icu-capacity.html`.

P. D. Levin, T. M. Worner, S. Sviri, S. V. Goodman, Y. G. Weiss, S. Einav, C. Weissman, and C. L. Sprung. Intensive care outflow limitation – frequency, etiology, and impact. *Journal of Critical Care*, 18(4): 206–211, 2003.

F. Lin, W. Chaboyer, and M. Wallis. A literature review of organisational, individual and teamwork factors contributing to the ICU discharge process. *Australian Critical Care*, 22(1):29–43, 2009.

A. Mandelbaum and A. L. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research*, 52(6):836–855, 2004.

A. Mandelbaum, P. Momčilović, and Y. Tseytlin. On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science*, 58(7):1273–1291, 2012.

M. A. Metcalfe, A. Sloggett, and K. McPherson. Mortality among appropriately referred patients refused admission to intensive-care units. *The Lancet*, 350(9070):7–11, 1997.

K. O'Connor and J. L. Dugan. Addressing floating and patient safety. *Nursing2017*, 47(2):57–58, 2017.

L. A. Pauls, R. Johnson-Paben, J. McGready, J. D. Murphy, P. J. Pronovost, and C. L. Wu. The weekend effect in hospitalized patients: A meta-analysis. *Journal of Hospital Medicine*, 12(9):760–766, 2017.

D. F. Perlmutter, C. Suico, A. N. Krauss, and P. Auld. A program to reduce discharge delays in a neonatal intensive care unit. *The American Journal of Managed Care*, 4(4):548–552, 1998.

O. Perry and W. Whitt. Responding to unexpected overloads in large-scale service systems. *Management Science*, 55(8):1353–1367, 2009.

O. Perry and W. Whitt. A fluid limit for an overloaded X model via a stochastic averaging principle. *Mathematics of Operations Research*, 38(2):294–349, 2013.

J. Rajgopal, Z. Wang, A. Schaefer, and O. Prokopyev. Effective management policies for remnant inventory supply chains. *IIE Transactions*, 41(5):437–447, 2009.

R. Robert, J. Reignier, C. Tournoux-Facon, T. Boulain, O. Lesieur, V. Gissot, V. Souday, M. Hamrouni, C. Chapon, and J.-P. Gouello. Refusal of intensive care unit admission due to a full unit: Impact on mortality. *American Journal of Respiratory and Critical Care Medicine*, 185(10):1081–1087, 2012.

R. Robert, N. Kentish-Barnes, A. Boyer, A. Laurent, E. Azoulay, and J. Reignier. Ethical dilemmas due to the Covid-19 pandemic. *Annals of Intensive Care*, 10(1):1–9, 2020.

A. L. Rosenberg and C. Watts. Patients readmitted to ICUs: A systematic review of risk factors and outcomes. *Chest*, 118(2):492–502, 2000.

A. L. Rosenberg, T. P. Hofer, R. A. Hayward, C. Strachan, and C. M. Watts. Who bounces back? Physiologic and other predictors of intensive care unit readmission. *Critical Care Medicine*, 29(3):511–518, 2001.

R. O. Roundy, W. L. Maxwell, Y. T. Herer, S. R. Tayur, and A. W. Getzler. A price-directed approach to real-time scheduling of production operations. *IIE Transactions*, 23(2):149–160, 1991.

C. Saville, T. Monks, P. Griffiths, and J. E. Ball. Costs and consequences of using average demand to plan baseline nurse staffing levels: A computer simulation study. *BMJ Quality & Safety*, 2020.

P. Shi, M. C. Chou, J. G. Dai, D. Ding, and J. Sim. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science*, 62(1):1–28, 2016.

A. Shmueli, C. L. Sprung, and E. H. Kaplan. Optimizing admissions to an intensive care unit. *Health Care Management Science*, 6(3):131–136, 2003.

R. W. Simpson. *Using Network Flow Techniques to Find Shadow Prices for Market Demands and Seat Inventory Control*. MIT, Department of Aeronautics and Astronautics, Flight Transportation Laboratory, 1989.

H. Song, A. L. Tucker, R. Graue, S. Moravick, and J. J. Yang. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science*, 66(9):3825–3842, 2020.

A. L. Stolyar and T. Tezcan. Shadow-routing based control of flexible multiserver pools in overload. *Operations Research*, 59(6):1427–1444, 2011.

K. Talluri and G. van Ryzin. An analysis of bid-price controls for network revenue management. *Management Science*, 44(11-part-1):1577–1593, 1998.

S. Thompson, M. Nunez, R. Garfinkel, and M. D. Dean. OR practice–efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research*, 57(2): 261–273, 2009.

E. Toner and R. Waldhorn. What US hospitals should do now to prepare for a COVID-19 pandemic. *Clinicians' Biosecurity News*, 2020. URL `http://www.centerforhealthsecurity.org/cbn/2020/cbnreport-02272020.html`.

M. Vergano, G. Bertolini, A. Giannini, G. R. Gristina, S. Livigni, G. Mistraletti, L. Riccioni, and F. Petrini. Clinical ethics recommendations for the allocation of intensive care treatments in exceptional, resource-limited circumstances: The Italian perspective during the COVID-19 epidemic. *Critical Care*, 24:165, 2020.

D. A. Walker. Walras's theories of tatonnement. *Journal of Political Economy*, 95(4):758–774, 1987.

A. R. Ward and M. Armony. Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research*, 61(1):228–243, 2013.

T. Williams and G. Leslie. Delayed discharges from an adult intensive care unit. *Australian Health Review*, 28(1):87–96, 2004.

E. L. Williamson. *Airline Network Seat Inventory Control: Methodologies and Revenue Impacts*. PhD thesis, Massachusetts Institute of Technology, 1992.

N. Zychlinski, A. Mandelbaum, P. Momčilović, and I. Cohen. Bed blocking in hospitals due to scarce capacity in geriatric institutionscost minimization via fluid models. *Manufacturing & Service Operations Management*, 22(2):396–411, 2020.

Electronic Companion

EC.1. Algorithmic Details

In our implementation of the subgradient algorithm, the dual optimization problem (7) is the master problem in dual decomposition when seeking to determine the optimal $\boldsymbol{\lambda}$. Given $\boldsymbol{\lambda}$, we can evaluate the dual function by solving the subproblems (5) and (6) in Section 4.1. Here, the Lagrange multiplier λ_i represents the per-patient priority coefficient for an ICU transfer out of unit i. Specifically, higher λ_i means patients from ICU i are prioritized in the allocation subproblem. Note that the subproblem maximizing ICU clinical benefit can be solved analytically and independently for each ICU i. In solving the dual problem, the master problem sets the coefficients $\boldsymbol{\lambda}$, then each ICU determines the total flow or rate of patient transfers. At the same time, the network determines the allocation of flow to downstream units. In the network subproblem, the value from directing flow to units within baseline capacity is λ_i, while the value from directing flow to units using additionally staffed beds is $\lambda_i - \beta_j$, depending on both the patients' origin i and destination j. Finally, the dual decomposition master problem adjusts the priority coefficients in order to bring the supply of downstream beds into consistency with demand from ICU requests. Algorithm 3 formalizes this subgradient approach.

Algorithm 3: Subgradient algorithm for minimizing the dual function

```
Input: Initial $\boldsymbol{\lambda}^{(0)}$, step size $\alpha^{(k)}$, maximum number of iterations $K$
Output: Estimates of $\boldsymbol{\lambda}^*$
for $i \in \mathcal{I}$ do
    Calculate initial subgradient $g_i^{(0)} \leftarrow \sum_{j \in \mathcal{J}(i)} (x_{ij}^{(0)} + z_{ij}^{(0)}) - y_i^{(0)}$
$q_{best} \leftarrow q(\boldsymbol{\lambda}^{(0)})$
for $k = 1, 2, \ldots, K$ do
    $\boldsymbol{\lambda}^{(k)} \leftarrow \boldsymbol{\lambda}^{(k-1)} - \alpha^{(k)} \mathbf{g}^{(k-1)}$
    Solve the subproblems with $\boldsymbol{\lambda}^{(k)}$ to obtain $\mathbf{x}^{(k)}, \mathbf{y}^{(k)}, \mathbf{z}^{(k)}, \mathbf{d}^{(k)}$
    for $i \in \mathcal{I}$ do
        Calculate the subgradient $g_i^{(k)} \leftarrow \sum_{j \in \mathcal{J}(i)} (x_{ij}^{(k)} + z_{ij}^{(k)}) - y_i^{(k)}$
    if $q(\boldsymbol{\lambda}^{(k)}) < q_{best}$ then
        $q_{best} \leftarrow q(\boldsymbol{\lambda}^{(k)})$
```

Here, g_i represents the violation of the coupling flow balance constraints. In particular, if $g_i < 0$, then ICU i seeks to transfer out more patients than can be routed, so the coefficient λ_i is increased. If, on the other hand, $g_i > 0$, then ICU i sends less patients than what the network can route, so the coefficient λ_i is decreased. With this priority adjustment, the algorithm lets the ICU subproblems decrease/increase the number of patients they seek to transfer, bringing the values closer to what is feasible for the network to allocate. This process is analogous to the tatonnement process in economics, which is the natural tendency of free competition markets to balance supply and demand through price adjustment (Walker 1987). We use a diminishing step size $\alpha^{(k)} \to 0$ which makes progress towards the minimum more likely at each iteration. To counteract the slow progress at iterations associated with diminishing step size rules, we also require $\sum_{k=0}^{\infty} \alpha^{(k)} = \infty$. This guarantees that $\lambda^{(k)}$ does not converge to a non-stationary point. The presented algorithm uses a fixed number of iterations K; however, alternative stopping criteria may be used (see Bertsekas (1999) for more details on subgradient algorithms).

Algorithm 4 finds the optimal value of the dual variables $\mathbf{u}, \mathbf{v}, w$ in model (8). In particular, the shadow prices of base capacity beds $\mathbf{u}$ only appear in constraints (8b) which are binding, as we are minimizing the objective with non-negative coefficients and variables. To find the shadow price of additional personnel w^*, we consider the downstream units in descending order of $\max_{i \in \mathcal{I}(j)}\{\lambda_i - \beta_j\}$. We create a list $\mathcal{L}$, where for each j in the ordered list of downstream units, we add D'_j items of value $\max\{0, \max_{i \in \mathcal{I}(j)}\{\lambda_i - \beta_j\}\}$. Then, w^* is the value at position $\nu + 1$ in $\mathcal{L}$. If $\nu + 1 > |\mathcal{L}|$, then $w^* = 0$.

```
Algorithm 4: Bed values (shadow prices) at downstream units
  Input: Network $\mathcal{G}$, capacities $\mathbf{D}$ and $\mathbf{D}'$, coefficients $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$
  Output: Optimal values $\mathbf{u}^*, \mathbf{v}^*, \mathbf{w}^*$
  for $j \in \mathcal{J}$ do
      $u_j^* \leftarrow \max\{0, \max_{i \in \mathcal{I}(j)}\{\lambda_i\}\}$
  $\hat{\mathcal{J}} \leftarrow$ list of units $j \in \mathcal{J}$ in descending order of $\max_{i \in \mathcal{I}(j)}\{\lambda_i - \beta_j\}$
  $\mathcal{L} \leftarrow [\,]$ empty list
  for $j \in \hat{\mathcal{J}}$ do
      for $l = 1, \ldots, D'_j$ do
          $\mathcal{L} \leftarrow \mathcal{L} + \max\{0, \max_{i \in \mathcal{I}(j)}\{\lambda_i - \beta_j\}\}$
  if $\nu + 1 \leq |\mathcal{L}|$ then
      $w^* \leftarrow \mathcal{L}[\nu + 1]$
  else if $\nu + 1 > |\mathcal{L}|$ then
      $w^* \leftarrow 0$
  for $j \in \mathcal{J}$ do
      $v_j^* \leftarrow \max\{0, \max_{i \in \mathcal{I}(j)}\{\lambda_i - \beta_j\} - w^*\}$
```

Proposition EC.1.1 *Algorithm 4 gives the optimal values of the variables in model* (8)*, i.e., the dual shadow prices of the downstream resources in model* (6)*.*

EC.2. Proofs of Propositions

We restate the propositions previously introduced and follow each with a proof.

Proposition 4.1 *There exists a vector* $\boldsymbol{\lambda}$ *such that the solutions* $\mathbf{y}$ *to the subproblems* (5) *with* $y_i = \sum_{j\in\mathcal{J}(i)}(x_{ij}+z_{ij})$ *are optimal for the network subproblem* (6)*. Furthermore,* $(\mathbf{x},\mathbf{y},\mathbf{z})$ *are optimal for the system model* (4)*.*

Proof. Model (4) is a concave maximization problem over a convex set, meaning that the KKT conditions are necessary and sufficient. Hence, there exists a tuple $(\mathbf{x}^*,\mathbf{y}^*,\mathbf{z}^*,\mathbf{p},\mathbf{q},\mathbf{r},\mathbf{s},\mathbf{t})$ satisfying the following conditions:

$$
\begin{array}{llll}
\text{[Stationarity]} & U'(y_i^*) - p_i - q_i = 0 & \text{if } y_i^* > 0, & \\
 & \le 0 & \text{if } y_i^* = 0 & \forall i \in \mathcal{I}, \\
 & p_i - r_j = 0 & \text{if } x_{ij}^* > 0, & \\
 & \le 0 & \text{if } x_{ij}^* = 0, & \forall i \in \mathcal{I}, j \in \mathcal{J}, \\
 & -\beta_j + p_i - s_j - t = 0 & \text{if } z_{ij}^* > 0, & \\
 & \le 0 & \text{if } z_{ij}^* = 0 & \forall j \in \mathcal{J}, \\
\text{[Primal Feasibility]} & \sum_{j\in\mathcal{I}(j)} (x_{ij}^* + z_{ij}^*) = y_i^* & & \forall i \in \mathcal{I}, \\
 & y_i^* \le C_i & & \forall i \in \mathcal{I}, \\
 & \sum_{i\in\mathcal{I}(j)} x_{ij}^* \le D_j & & \forall j \in \mathcal{J}, \\
 & \sum_{i\in\mathcal{I}(j)} z_{ij}^* \le D_j' & & \forall j \in \mathcal{J}, \\
 & \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}(i)} z_{ij}^* \le \nu, & & \\
\text{[Dual Feasibility]} & \mathbf{q},\mathbf{r},\mathbf{s},\mathbf{t} \ge 0, & & \\
\text{[Complementary Slackness]} & q_i(y_i^* - C_i) = 0 & & \forall i \in \mathcal{I} \\
 & r_j \left(\sum_{i\in\mathcal{I}(j)} x_{ij}^* - D_j \right) = 0 & & \forall j \in \mathcal{J}, \\
 & s_j \left(\sum_{i\in\mathcal{I}(j)} z_{ij}^* - D_j' \right) = 0 & & \forall j \in \mathcal{J}, \\
 & t \left(\sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}(i)} z_{ij}^* - \nu \right) = 0. & &
\end{array}
$$

Moreover, $(\mathbf{x}^*,\mathbf{y}^*,\mathbf{z}^*)$ solve model (4).

The ICU model maximizes a concave function over a convex set, hence there exist multipliers $\boldsymbol{\gamma}$ that satisfy the conditions:

$$\begin{array}{llll}
\text{[Stationarity]} & U'(y_i^*) - \lambda_i - \gamma_i = 0 & \text{if } y_i^* > 0, & \\
 & \leq 0 & \text{if } y_i^* = 0 & \forall i \in \mathcal{I}, \\
\text{[Primal Feasibility]} & y_i^* \leq C_i & & \forall i \in \mathcal{I}, \\
\text{[Dual Feasibility]} & \boldsymbol{\gamma} \geq 0, & & \\
\text{[Complementary Slackness]} & \gamma_i(y_i^* - C_i) = 0 & & \forall i \in \mathcal{I}
\end{array}$$

and $\mathbf{y}^*$ solve the ICU value-maximizing models.

Finally, the necessary and sufficient conditions for optimality of the network subproblem that allocates ICU patient flows to downstream units state that exists a tuple $(\mathbf{x}^*, \mathbf{z}^*, \mathbf{u}, \mathbf{v}, \mathbf{w})$ satisfying:

$$\begin{array}{llll}
\text{[Stationarity]} & \lambda_i - u_j = 0 & \text{if } x_{ij}^* > 0, & \\
 & \leq 0 & \text{if } x_{ij}^* = 0, & \forall i \in \mathcal{I}, j \in \mathcal{J}, \\
 & \lambda_i - \beta_j - v_j - w = 0 & \text{if } z_{ij}^* > 0, & \\
 & \leq 0 & \text{if } z_{ij}^* = 0 & \forall j \in \mathcal{J}, \\
\text{[Primal Feasibility]} & \sum_{i \in \mathcal{I}(j)} x_{ij}^* \leq D_j & & \forall j \in \mathcal{J}, \\
 & \sum_{i \in \mathcal{I}(j)} z_{ij}^* \leq D'_j & & \forall j \in \mathcal{J}, \\
 & \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}(i)} z_{ij}^* \leq \nu, & & \\
\text{[Dual Feasibility]} & \mathbf{u}, \mathbf{v}, \mathbf{w} \geq 0, & & \\
\text{[Complementary Slackness]} & u_j \left(\sum_{i \in \mathcal{I}(j)} x_{ij}^* - D_j \right) = 0 & & \forall j \in \mathcal{J}, \\
 & v_j \left(\sum_{i \in \mathcal{I}(j)} z_{ij}^* - D'_j \right) = 0 & & \forall j \in \mathcal{J}, \\
 & w \left(\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}(i)} z_{ij}^* - \nu \right) = 0, & &
\end{array}$$

so that $(\mathbf{x}^*, \mathbf{z}^*)$ solve the network allocation model.

Therefore, if $\boldsymbol{\lambda} = \mathbf{p}$ and $y_i = \sum_{j \in \mathcal{J}(i)} (x_{ij} + z_{ij})$, then the solutions to the ICU subproblems $\mathbf{y}^*$ and the solution to the network subproblem $(\mathbf{x}^*, \mathbf{z}^*)$ comprise a solution to the system model (4). The conditions identify $\boldsymbol{\gamma} = \mathbf{q}$, $\mathbf{u} = \mathbf{r}$, $\mathbf{v} = \mathbf{s}$, and $\mathbf{w} = \mathbf{t}$. □

Proposition EC.1.1 *Algorithm 4 gives the optimal values of the variables in model* (8)*, i.e., the dual shadow prices of the downstream resources in model* (6)*.*

Proof. Note that the shadow prices of regularly staffed beds only appear in constraints (8b) which can be restated as $u_j \geq \max_{i\in\mathcal{I}(j)}\{\lambda_i\}$ for all $j \in \mathcal{J}$. Furthermore, as we are minimizing the objective with non-negative coefficients and variables, it follows that (8b) are binding. Hence, line 2 of Algorithm 4 assigns the optimal shadow prices of regular beds u_j^*.

If $\nu \geq \sum_{j\in\mathcal{J}} D'_j$, there is no value in adding more personnel as there are not enough beds to be staffed. Hence, lines 10-11 of Algorithm 4 correctly assign $w^* \leftarrow 0$.

If $\nu < \sum_{j\in\mathcal{J}} D'_j$, there is at least one extra downstream bed that may be staffed. According to the objective function of model (6), the contribution of an additional operational bed at unit j is $\max\{0, \max_{i\in\mathcal{I}(j)}\{\lambda_i - \beta_j\}\}$. The maximizing objective ensures that the ν beds with highest contribution are already staffed. Hence, the value of additional staff is equal to the highest value of $\max\{0, \max_{i\in\mathcal{I}(j)}\{\lambda_i - \beta_j\}\}$ after removing the first ν values at the corresponding units. Therefore, the ordering of the contributions in $\mathcal{L}$ of Algorithm 4 guarantees the value of w^* is assigned correctly.

Given w^* and the fact that v_j are non-negative, it follows that constraints (8c) can be restated as $v_j \geq \max\{0, \max_{i\in\mathcal{I}(j)}\{\lambda_i - \beta_j\} - w^*\}$. As the coefficients of v_j are non-negative in a minimization objective, it follows that (8c) are binding and lines 12-13 make the correct assignment of v_j^*. □

EC.3. Network Illustrations

Figure EC.1 illustrates the topology of the networks used in the large-scale numerical instances.

EC.4. Static Policy

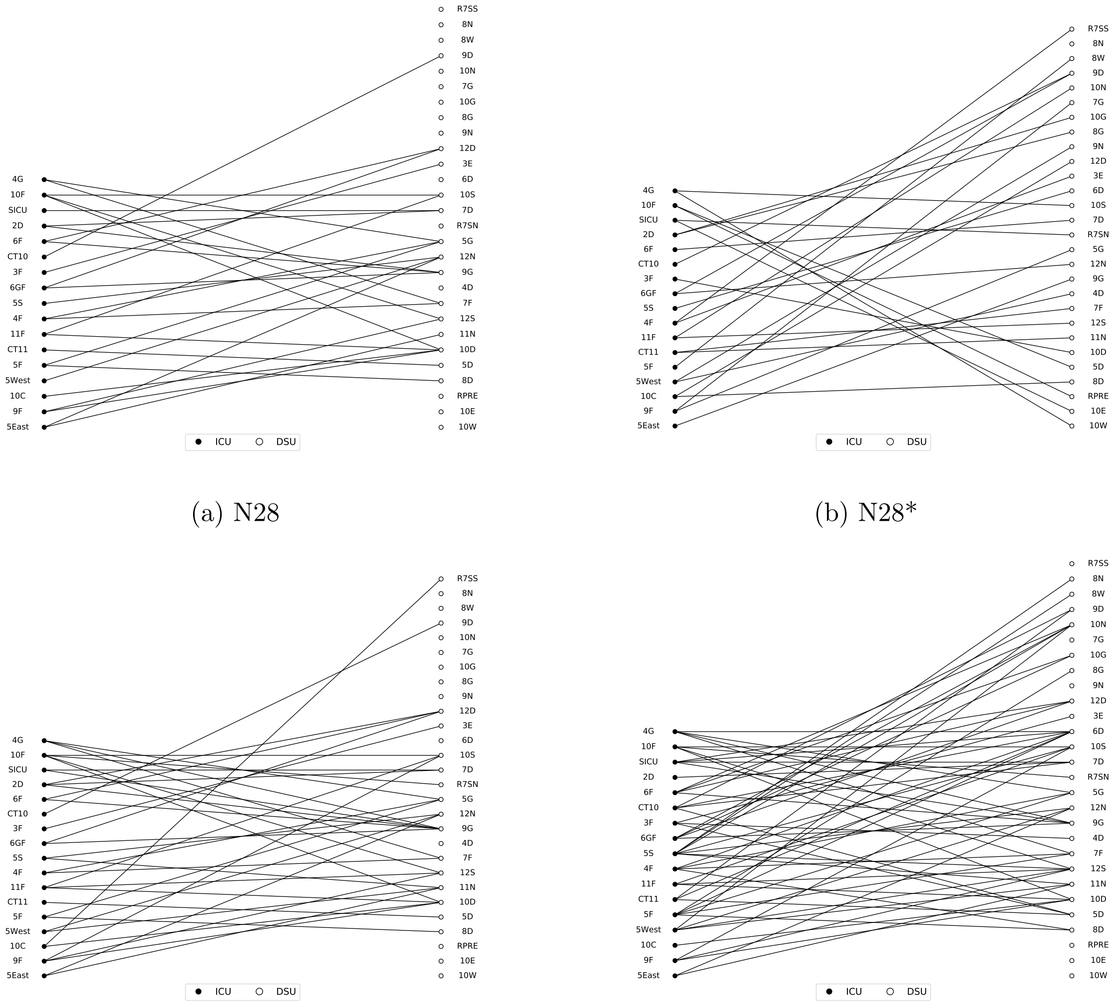
We provide details on the linearizing reformulation of model (3) in order to obtain the static policy used for comparison with the proposed acuity-based policies in the numerical studies. For each $i \in \mathcal{I}$, let $u_i^k := U_i(k)$ for $k = 1, 2, \ldots, C_i$. Let $x_{ij}, z_{ij} \in \mathbb{Z}_+$ represent the patient allocations from i to j to baseline capacity and flexible staffing beds, respectively (as before). We redefine the variables $y_{ik} \in \{0,1\}$ to indicate if the total number of patients transferred out of unit i is equal to k.

$$\max \sum_{i\in\mathcal{I}}\sum_{k=0}^{C_i} u_i^k y_{ik} - \sum_{j\in\mathcal{J}} \beta_j \sum_{i\in\mathcal{I}(j)} z_{ij} \tag{EC.1a}$$

$$\text{s.t.} \sum_{j\in\mathcal{J}(i)} x_{ij} + z_{ij} = \sum_{k=0}^{C_i} k y_{ik} \qquad \forall i \in \mathcal{I}, \tag{EC.1b}$$

$$\sum_{k=0}^{C_i} y_{ik} = 1 \qquad \forall i \in \mathcal{I}, \tag{EC.1c}$$

$$\sum_{i\in\mathcal{I}(j)} x_{ij} \leq D_j \qquad \forall j \in \mathcal{J}, \tag{EC.1d}$$



(a) N28

(b) N28*

(c) N37

(d) N60

Figure EC.1 Illustrations of the network topologies used in the numerical studies.

$$\sum_{i\in\mathcal{I}(j)} z_{ij} \le D'_j \qquad \forall j\in\mathcal{J}, \tag{EC.1e}$$

$$\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}(j)} z_{ij} \le \nu \tag{EC.1f}$$

$$x_{ij}, z_{ij} \in \mathbb{Z}_+ \qquad \forall i\in\mathcal{I}, j\in\mathcal{J}, \tag{EC.1g}$$

$$y_{ik} \in \{0,1\} \qquad \forall i\in\mathcal{I}, k\in\{1,2,\ldots,C_i\}. \tag{EC.1h}$$

The objective function (EC.1a) calculates the total clinical benefit of ICU transfers less cost of staffing (as in (4a)). Constraints (EC.1b) ensure the number of patients allocated downstream equals the number of patients transferred out of ICU. Constraints (EC.1c) ensure exactly one number of total patients to transfer is chosen for each ICU. Constraints (EC.1d)-(EC.1f) enforce the bed capacity and staffing limitations at downstream units.

EC.5. Sensitivity Analyses

In order to provide insights on the proposed acuity-based policy's performance in other demand scenarios, we conduct several additional experiments. We test the policies in a medium demand scenario, where we sample arrivals from $\texttt{Poisson}(2\mu_k^{at})$ and ICU transfer requests and DSU discharges from $\texttt{Poisson}(2\mu_k^{dt})$ based on historic discharge rates. Similarly, we simulate a low (or baseline) demand scenario where we sample arrivals from $\texttt{Poisson}(\mu_k^{at})$ and ICU transfer requests and DSU discharges from $\texttt{Poisson}(\mu_k^{dt})$. All remaining parameters are kept as described in Section 5.2.2.

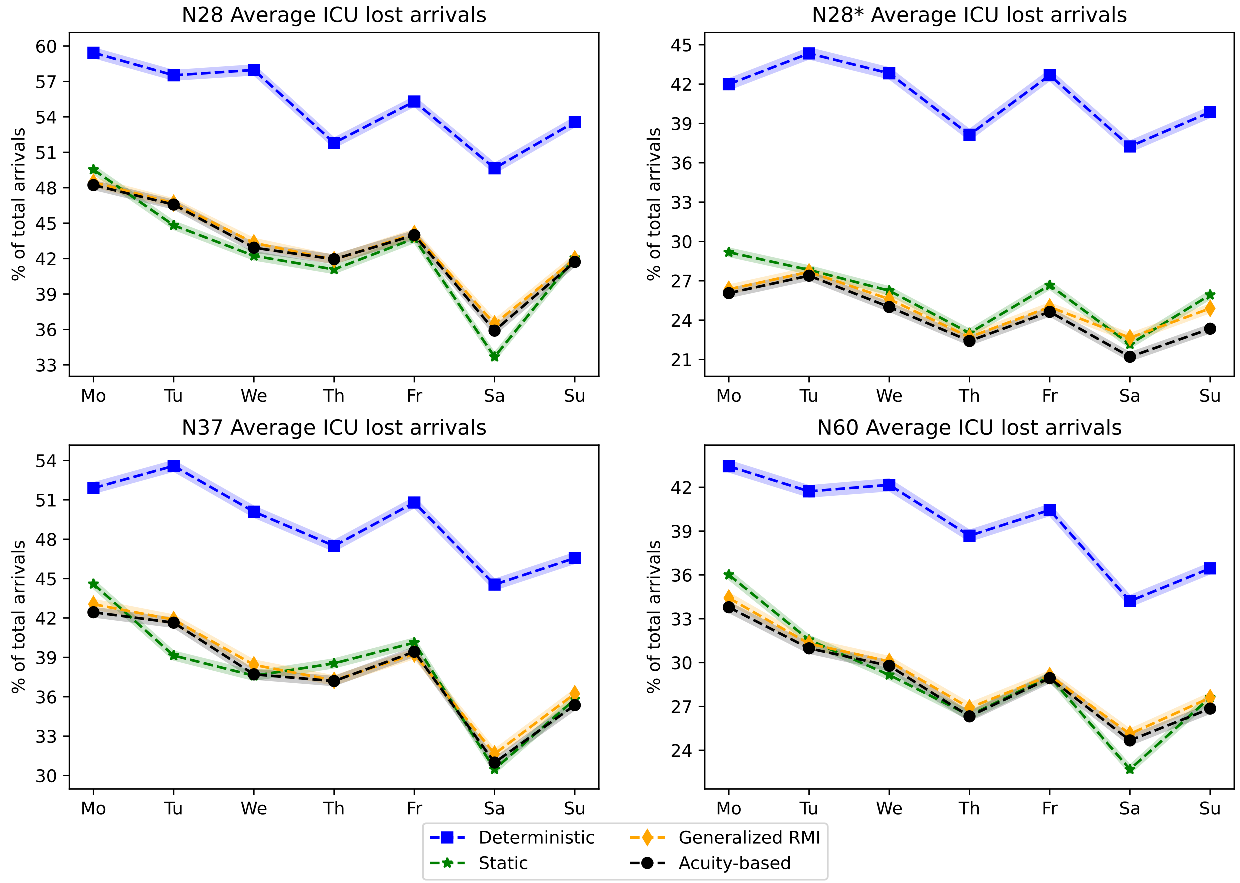


Figure EC.2 **Medium Demand Setting: Seven day patient flow simulation for four network topologies depicting average (among all ICU for each day) lost ICU arrivals (diverted patients) with 95% confidence intervals.**

Figure EC.2 shows a comparison in the number of diverted patients for all tested policies in the medium demand setting. Notably, the deterministic policy performs the worst, as before. The acuity-based policy performs similarly to the generalized RMI and the static policies. Moreover, the acuity-based policy appears to have a slight advantage over the generalized RMI policy.

Next, Figure EC.3 shows a comparison in the number of diverted patients for all tested policies in the low demand setting. We observe overall lower average numbers of diverted patients for all

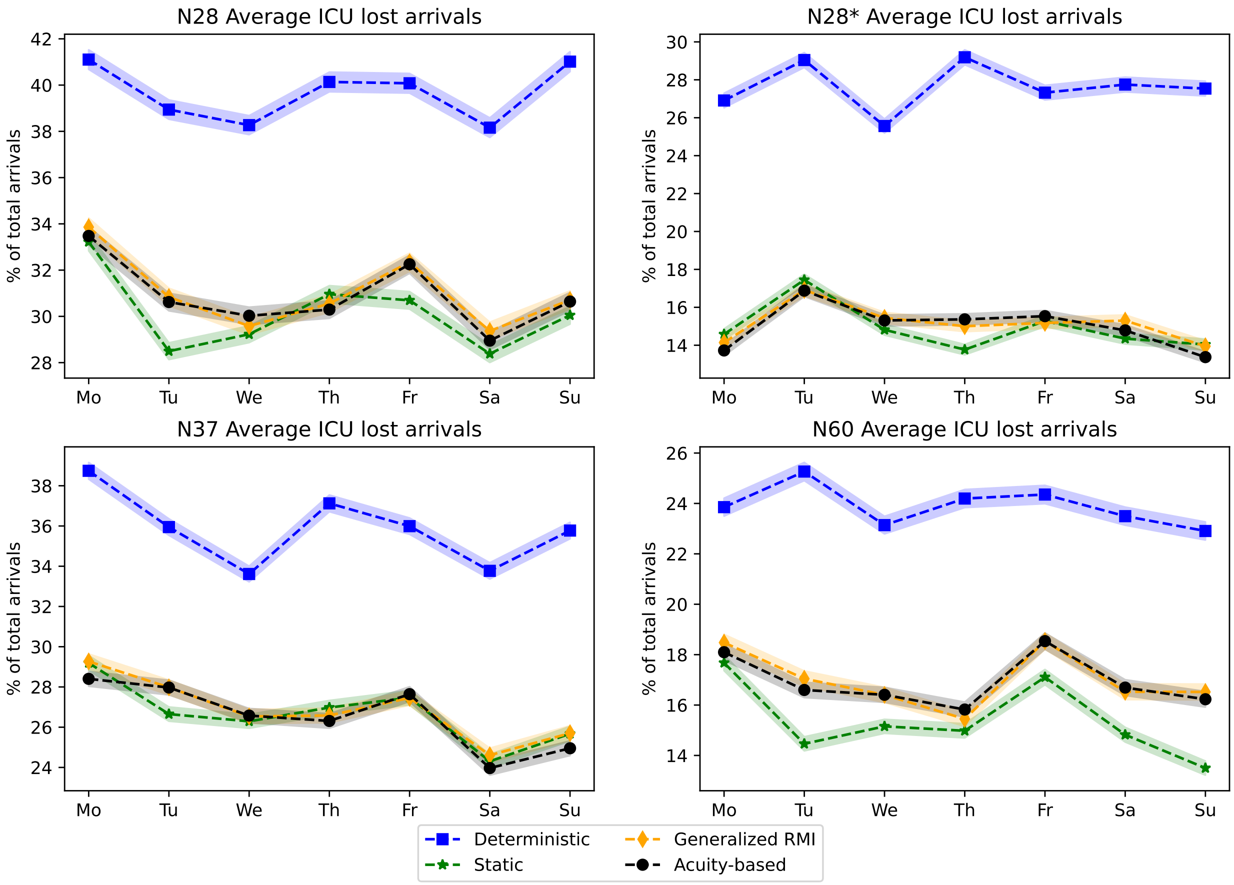


Figure EC.3 **Low Demand Setting: Seven day patient flow simulation for four network topologies depicting average (among all ICU for each day) lost ICU arrivals (diverted patients) with 95% confidence intervals.**

network topologies. Once again, the deterministic policy performs the worst. Here, the acuity-based policy performs similarly to the generalized RMI policy. The static policy appears to have an advantage in many of the cases. The lower variation in demand makes it more likely that a myopic solution like the one provided by the static policy is optimal for the system. Nevertheless, the acuity-based policy remains competitive in many cases. More importantly, it is robust to disruptions and delays in relaying information which can easily render a solution obtained from the static policy infeasible.

These experiments confirm the intuition that the acuity-based policy is most valuable in high-demand settings. The high variation in expected arrivals and discharges makes flexible staffing and strategic patient allocation significantly more important for ensuring increased patient throughput. While the deterministic policy is the easiest to implement, its overall performance leads us to recommend the implementation of at least the generalized RMI when expecting medium or baseline-level demand.