

For more information on this journal please visit the journal website at www.tandfonline.com/loi/uhse21
For more information on Taylor & Francis please visit the Taylor & Francis website at www.tandfonline.com
For more information on Taylor & Francis Group please visit the Taylor & Francis Group website at www.tandfonline.com
For more information on Taylor & Francis Open Access please visit the Taylor & Francis Open Access website at www.tandfonline.com
For more information on Taylor & Francis Digital Rights Management please visit the Taylor & Francis Digital Rights Management website at www.tandfonline.com
For more information on Taylor & Francis Copyright Clearance Center please visit the Taylor & Francis Copyright Clearance Center website at www.copyright.com
For more information on Taylor & Francis CrossMark please visit the Taylor & Francis CrossMark website at www.crossmark.com
For more information on Taylor & Francis Scopus please visit the Taylor & Francis Scopus website at www.scopus.com
For more information on Taylor & Francis Web of Science please visit the Taylor & Francis Web of Science website at www.webofscience.com
For more information on Taylor & Francis Research Alert please visit the Taylor & Francis Research Alert website at www.researchalert.com
For more information on Taylor & Francis Alerting Services please visit the Taylor & Francis Alerting Services website at www.alertingservices.com
For more information on Taylor & Francis Taylor & Francis Online please visit the Taylor & Francis Online website at www.tandfonline.com
For more information on Taylor & Francis Taylor & Francis Group please visit the Taylor & Francis Group website at www.tandfonline.com



ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/uhse21>

Routing and staffing in emergency departments: A multiclass queueing model with workload dependent service times

Siddhartha Nambiar, Maria E. Mayorga & Yunan Liu

To cite this article: Siddhartha Nambiar, Maria E. Mayorga & Yunan Liu (2022): Routing and staffing in emergency departments: A multiclass queueing model with workload dependent service times, IISE Transactions on Healthcare Systems Engineering, DOI: [10.1080/24725579.2022.2100522](https://doi.org/10.1080/24725579.2022.2100522)

To link to this article: <https://doi.org/10.1080/24725579.2022.2100522>



Published online: 26 Jul 2022.



Submit your article to this journal [↗](#)



Article views: 19






View related articles [↗](#)



View Crossmark data [↗](#)

Routing and staffing in emergency departments: A multiclass queueing model with workload dependent service times

Siddhartha Nambiar , Maria E. Mayorga , and Yunan Liu 

^aEdward P. Fitts Industrial & Systems Engineering, North Carolina State University, Raleigh, NC, USA

ABSTRACT

Efficient patient flow through an emergency department is a critical factor that contributes to a hospital's performance, which influences overall patient health outcomes. In this work, we model a multiclass multiserver queueing system where patients of varying acuity receive care from one of several wards, each ward is attended by several nurses who work as a team. Supported by empirical evidence that a patient's time-in-ward is a function of the nurse-patient ratio in that ward, we incorporate state-dependent service times into our model. Our objective is to reduce patient time in system and to control nurse workload by jointly optimizing patient routing and nurse allocation decisions. Due to the computational challenges in formulating and solving the queueing model representation, we study a corresponding deterministic fluid model which serves as a first-order approximation of the multiclass queueing model. Next, we formulate and solve an optimization model using the first-order control equations and input the results into a discrete-event simulation to estimate performance measures, such as patient length-of-stay and ward workload. Finally, we present a case study using retrospective data from a real hospital which highlights the importance of accounting for nurse workload and service behavior in developing routing and staffing policies.

KEYWORDS

Simulation; fluid approximation; queueing model; patient flow

1. Introduction

The emergency department (ED) is arguably the most operationally complex clinical setting of the modern hospital. EDs in most hospitals around the world suffer from common issues such as long waits, inefficient processes and poor patient satisfaction (Derlet & Richards, 2000). The issue of long waits, in particular, is a result of increasing ED volumes and a sign of ED overcrowding. According to the Agency for Healthcare Research and Quality (AHRQ, 2018), 90% of EDs in the country reported that they were "holding" admitted patients in the ED while awaiting inpatient beds. This backlog of patients having to wait within the ED disrupts the efficient flow of patients throughout the hospital system. Efficient patient flow has been shown to be an important factor contributing to patient safety (Carayon & Wood, 2009). Some indicators of effective patient flow include high patient throughput, and low patient waiting times while maintaining adequate staff utilization rates and low physician idle times (Jun et al., 1999).

Improving patient flow is challenging because the rate of patient arrivals to a hospital is uncertain both in timing and volume (Denton, 2013). Despite this uncertainty, EDs have it in their power to manage the flow of patients once they arrive in order to provide effective care. Emergency

departments typically stratify incoming patients into groups based on their severity. Examples of triage systems being used by hospital systems today to assess the severity of incoming patients' conditions include the Australasian Triage Scale (ATS) (Considine et al., 2004), the Canadian Triage and Acuity Scale (CTAS) (Murray, 2003), the Manchester Triage System (MTS) (Parenti et al., 2014), and the Emergency Severity Index (ESI) (Tanabe et al., 2004). Hospitals use such groupings of patients to route them to appropriate units (or wards) within the ED for treatment. This routing (also known as "streaming") of patients plays a vital role in improving the efficiency of an ED's operations.

There exists extensive literature on the operational and monetary benefits of efficient patient flow and routing (Armony et al., 2015; Carnes et al., 2015; Haraden & Resar, 2004). Furthermore, there is a recent focus on better understanding the impact of workload experienced by nurses and providers resulting from flow redesign (Nicosia et al., 2018). The workload experienced by clinicians and nurses is a critical factor in the evaluation of operational metrics (e.g., clinician performance and staffing decisions) in healthcare systems (Mazur et al., 2016; Upenieks et al., 2007). High workload is associated with nurse turnover and shortages, clinician burnout, and undesired patient outcomes. Some examples of negative patient outcomes as a result of high

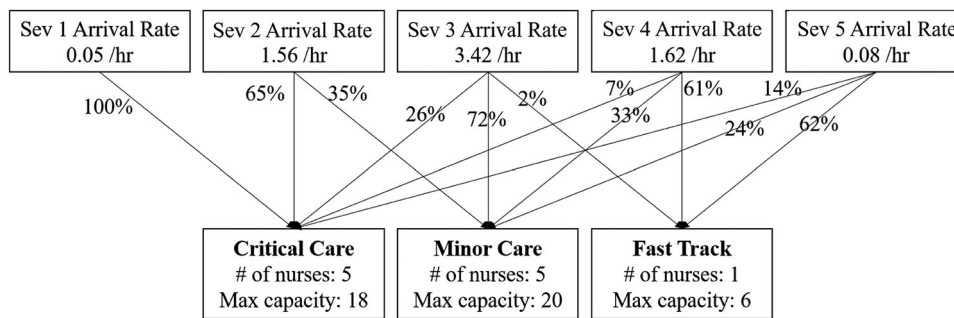


Figure 1. Pictorial representation for the status quo values of patient arrival rates, routing proportions, ward staffing, and ward capacity at the hospital's ED.

workload include increased mortality in the intensive care unit (ICU) and during post-operative recovery, prolonged length-of-stay (LOS) and higher rates for procedure related infections (Ball et al., 2018; Holden et al., 2011; Lamy Filho et al., 2011; Magalhães et al., 2017).

Because workload plays a significant role in affecting the efficiency and quality of care, there is a need to redesign routing protocols and restructure resource allocation policies while considering workload.

1.1. Motivating example - How nurse workload affects patient service time

We motivate our work using findings from preliminary analyses of how workload affects patient time-in-ward (service time) based on data from a regional hospital in North Carolina, USA. The dataset contains information on over 88,000 unique ED visits from November 2017 to April 2019, each with over 150 variables including timestamps, visit attributes and patient outcomes. We also have information on physician and nurse schedules and daily bed assignments. Patients arriving to this ED were triaged into one of five different severity types (with a severity level of 1 being the highest and a severity level of 5 being the lowest) and were assigned to one of three different wards. The wards were named critical care (CC), minor care (MC), and fast track (FT). The CC ward was typically occupied by patients of severity levels 1 and 2, while the FT ward was usually visited by less severe patients (levels 4 and 5). We inferred patient arrival rates by calculating the inter-arrival times for each patient severity type. Figure 1 gives the values for arrival rates, staffing levels, maximum capacity and routing proportions that we obtained from the data and used in our numerical analyses. We refer the reader to work by Swan et al. (2019) for a more detailed description of the data. Our goal, during this preliminary review of the data, was to investigate the relationship between workload experienced by nurses in a ward and the average time patients spend in a ward. As universal measures for workload do not exist (Fishbein et al., 2019), we used the ratio of patients to nurses in a ward as a proxy measure for workload.

In Figure 2, we demonstrate this relationship for patients of severity level 3 assigned to the MC ward. For each patient we calculate the average patient-nurse ratio.

For example, suppose a patient is in the ward from 8 am to 11 am with one nurse, he/she is the only patient from 8 am to 9 am, and another patient joins from 9 am to 11 am and a third patient is there from 10 am to 11 am, then the average patient-nurse ratio for that patient is 2 and the time-in-ward is 180 minutes. Next, we aggregated average patient-nurse ratio (x-axis) into buckets and calculated the average value of patient time-in-ward (y-axis) within each of those buckets. (We note here that we excluded data points with time-in-ward values greater than 24 hours to remove outliers.) We see that the curve follows a distinct polynomial form. We fit the curve in Figure 2 using a second order polynomial function (dashed red line) as well as a LOESS (Locally Estimated Scatterplot Smoothing) regression (black line). We see that the LOESS fit matches quite well with the polynomial fit in this case. A similar analysis for all combinations of patient severity types and wards shows similar trends, as is later shown in Section 4. Considering the form of Figure 2, we see that a patient's time-in-ward first increases on increasing patient-nurse ratio as each nurse in the ward is required to care for more patients on average. However, at higher values of patient-nurse ratio, the patient time-in-ward begins to decrease, leading to an inverted U-shaped curve. Though workload is measured in different ways, this observation is well-supported by prior literature. For example Batt and Terwiesch (2012) find that the service time in an ED is a U-shaped function of the number of patients in the waiting room. Based on inpatient data from over 200 California hospitals, Berry Jaeger and Tucker (2017) find that patient LOS (which includes service time and wait time) increases as occupancy increases, until a tipping point, resulting in an inverted U-shaped relationship between utilization and throughput time. Both studies find similar reasons for this, where initially a slow-down occurs due to the strain on a complex system with shared resources, then a speed-up due to change in service such as early discharge to alleviate congestion.

The key takeaway from this exploratory analysis is that patient time-in-ward is clearly dependent on the workload experienced by nurses in that ward. We argue that a model that attempts to assign resources to reduce LOS should take into account the functional relationship between workload and service time. This observation sets the stage for the work we undertake in the remainder of this article.

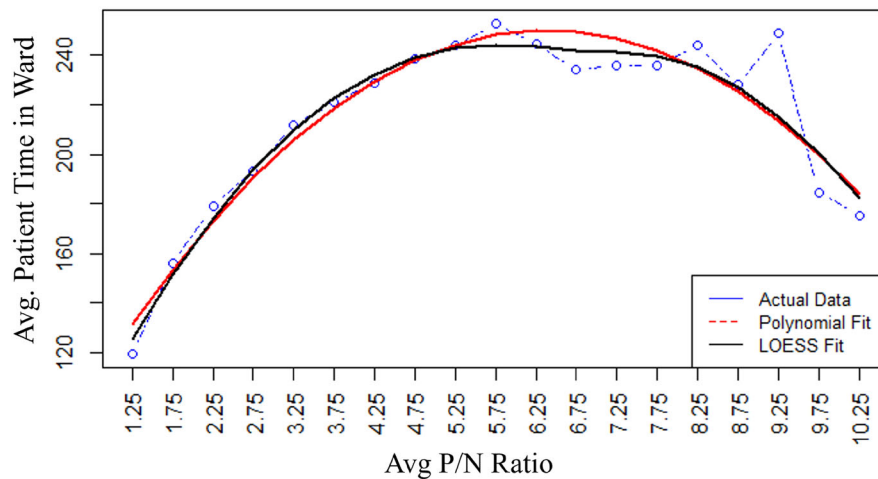


Figure 2. Average patient time-in-ward (for patients of severity level 3 in MC ward) fit against bucketed values of average patient-nurse ratio showing actual data, a LOESS fit, and a polynomial fit that we used in our experimental analyses.

1.2. Setting and related literature

In the US patients arriving to an ED are usually routed to wards based on their severity. These areas are sometimes referred to as bays, zones, or units-within-units. Furthermore, team-based care (or pods), in which physicians and nurses are assigned to work in teams in specific areas, has gained popularity as it has shown to be efficient and improve patient outcomes (Dinh et al., 2015; Mitchell & Golden, 2012). In this article we model a multiclass multi-server queueing system where patients of different acuity levels are assigned to one of several ED wards and receive care from the medical team in their ward. While a team may consist of physicians, technicians, clerks and nurses, here we focus on the assignment of nurses to units and assume the other staff is held constant¹. We assume that a patient's service time is a function of the workload of nurses represented as the ratio of patients-to-nurses in their ward (as shown in Figure 2). Our model reduces patient LOS (which includes time-in-ward and wait time) and controls nurse workload by optimizing routing and nurse allocation decisions between the units.

In this work we consider joint staffing and routing in an emergency department, modeled as a multiclass multiserver queueing control problem. While there is much literature related to staffing and/or patient assignment, some ignore time dynamics and queueing effects. Thus we focus our discussion on multiclass multiserver queueing models with healthcare applications.

1.2.1. Multiclass multiserver queueing control

The patient assignment or admission control problem has been studied by several, including (Helm et al., 2011; Saghaian et al., 2012, 2014), who use dynamic programming to improve flow in the ED. In recent work, Dai and Shi

(2019) model hospital inpatient flow as a multiclass, multi-pool parallel-server queueing system and formulate the overflow decision problem, where overflow is used to re-route patients to different server pools if their primary unit assignment is at capacity, the problem is solved using approximate dynamic programming. We refer the reader to work by Saghaian et al. (2015) for a more comprehensive review of articles that discuss patient flow optimization in emergency departments. Staffing decisions have also been studied. Liu and Whitt (2012c) determine staffing levels in a queueing model with non-exponential service times and time-varying arrivals; they later extend the model to feedforward (Liu & Whitt, 2014c) and feedback (Liu & Whitt, 2017) queueing networks. Cohen et al. (2014) allocate surgeons to an ED in a mass casualty context. In recent work, Chan et al. (2021) consider staffing decisions, where the ability to reassign servers happens at discrete-time intervals, or under partial flexibility and this work is extended to consider two types of nurses (Chan et al., 2020). These articles use fluid approximations to solve the problem, due to their complexity. While staffing decisions are closely related to routing, as optimal staffing depends on the choice of routing rule used and vice versa (Gans et al., 2003), the articles above treat these decisions separately. In fact, Harrison and Zeevi (2005) note that nurse staffing and patient routing are often treated in a separate but hierarchical manner due to the computational complexity involved. We do find one closely related stream of work, which is motivated by emergency departments and the Canadian triage and acuity scale, develops a dynamic staffing-and-routing rule for a multiclass V model subject to chance constraints on customer delays (Liu, 2018; Liu et al., 2022). In addition to tackling the joint staffing and routing problem, our work considers pooled service (team-based care) and state-dependent service.

1.2.2. Pooled service

In our work we consider team-based care; few analytical models for resource allocation in healthcare consider the fact that resources within units are partially shared, central resources. In general service systems, the use of pooled

¹The following articles discuss how team-based care was implemented in ED pods at Sharp Memorial Hospital in San Diego CA <https://healthmanagement.org/c/hospital/news/pod-and-huddle-ed-model-speeds-processes> and at NYU Lutheran Medical Center in Brooklyn NY <https://www.reliamedia.com/articles/140681-team-based-pod-system-reduces-lengths-of-stay-for-treat-and-release-patients>

resources is related to the concept of “processor sharing” (Kleinrock, 1967). Processor sharing is a service policy where customers are all served simultaneously in a queueing system. Under processor sharing, each customer receives an equal fraction of the service capacity available. Sharing resources within a unit is an idea that is relatively new in healthcare analytics literature. Agor et al. (2017) developed a simulation model in which incoming patients are assigned to teams of providers of different skill levels. Mandelbaum et al. (2012) showed that based on empirical hospital data the Inverted-V queueing model best models patients spending time in units within a hospital. The Inverted-V model assumes that upon entering a queueing system, an agent (patient) is assigned to a “pool” of servers instead of being assigned to a single server. Several authors continued to build on this by proposing a variety of patient/customer routing algorithms in an Inverted-V queueing context (Almehdawe et al., 2013; Armony & Ward, 2010; Ward & Armony, 2013). In addition to considering pooled service, we model state-dependent service.

1.2.3. Workload-dependent service

We next review both empirical and theoretical works on queues with workload-dependent service times. Kc and Terwiesch (2009), Kc and Terwiesch (2012), and Anderson et al. (2011) study how high workload impacts ICU LOS; these works reveal that high occupancy rates can lead to shorter LOS due to the need for accommodating new and more critical patients. Also see Kim et al. (2021) for an empirical study on how an ICU’s capacity strain affects the patients’ LOS. Chan et al. (2017) study a new hospital queueing model in which excessive patient delay leads to adverse health conditions, which in turn result in a longer LOS. Batt and Terwiesch (2012); Berry Jaeker and Tucker (2017) find that the service time in an ED exhibits a U-shaped function of occupancy; these works serve as the primary motivation for the workload-dependent assumption in the present work (see Figure 2). The present work is also related to the more general literature on queues having a state-dependent service rate. See Ata and Shneerson (2006); George and Harrison (2001); Powell and Schultz (2004) for settings of queueing systems where service rates may be dynamically controlled based on the system’s congestion level; these theoretical works show that service rates should increase with congestion in order to achieve the optimal system-level performance. The present work draws distinctions from the above extant literature by studying the multiclass multiserver setting, and in addition, it investigates the joint staffing-and-routing decisions in response to the workload-dependent aspect of the model.

To summarize, our main contributions are as follows:

- **Queueing model with joint staffing and routing assignments.** We propose a multiclass and multiserver queueing system that allows for dynamic server assignment and routing in a joint optimization framework.
- **Team-based care and ward-level workload considerations.** We model team-based care, in which servers are

shared within a pool (or ward). Staffing and routing decisions can often lead to imbalanced workload between units. We specifically model workload, show the impact that staffing and routing have on workload, and optimize subject to workload constraints.

- **State-dependent service.** We are the first to consider the impact of workload on patient service time in a decision-modeling framework. While this phenomenon has been observed empirically, this form of state-dependent service time has not been previously incorporated into a multiclass and multiserver queueing system. The control problem under this assumption introduces new complexities, and thus cannot be modeled using the typical Markov decision process (MDP) approach. Thus we conduct a fluid analysis and find asymptotically optimal policies.

An outline of the remainder of this article is as follows. Section 2 describes the multiclass multiserver queueing model with pooled service and workload dependent service. To solve the problem, we develop a fluid model approximation and optimization model, which we validate with simulation, this is described in Section 3. In Section 4, we conduct several experiments by optimizing patient routing and nurse staffing for a case study under different constraints. Lastly, we conclude with Section 5.

2. Model description & formulation

To describe our model formulation we first describe the abstraction of the process of patient flow through an ED as a multiclass multiserver queueing model. Then we specify two key features of our model (pooled service and consideration for nurse workload) and define them mathematically.

2.1 Queueing model

We define a multiclass queueing model to represent the arrival and service process within a hospital emergency department. Let us consider patients of I different severity types and J wards, with ward j , $\forall 1 \leq j \leq J$ containing s^j nurses. Each ward j can house a maximum of M^j patients which would presumably be greater than the number of nurses s^j , though this is not a requirement for our model. Unlike a traditional queueing model, where a single patient is served by a single nurse, we assume that the patients within a ward receive team-based care, or pooled service from a team of nurses in the ward.

A pictorial representation of the patient flow process is provided in Figure 3. Patients of severity type i arrive to the system with average inter-arrival time $\frac{1}{\lambda_i}$. These patients may be assigned to ward j according to a routing proportion r_i^j with $\sum_{j=1}^J r_i^j = 1$ for $1 \leq i \leq I$. In other words, a proportion r_i^j of patients of severity i are served by nurses in ward j . These proportions may be thought of as probabilities and are treated as decision variables in our model. We assume that each patient severity type is associated with a queue where they wait if they are unable to enter service immediately upon arrival. We assume an infinite buffer for this queue. After arrival and before joining service, patients may abandon (leave after

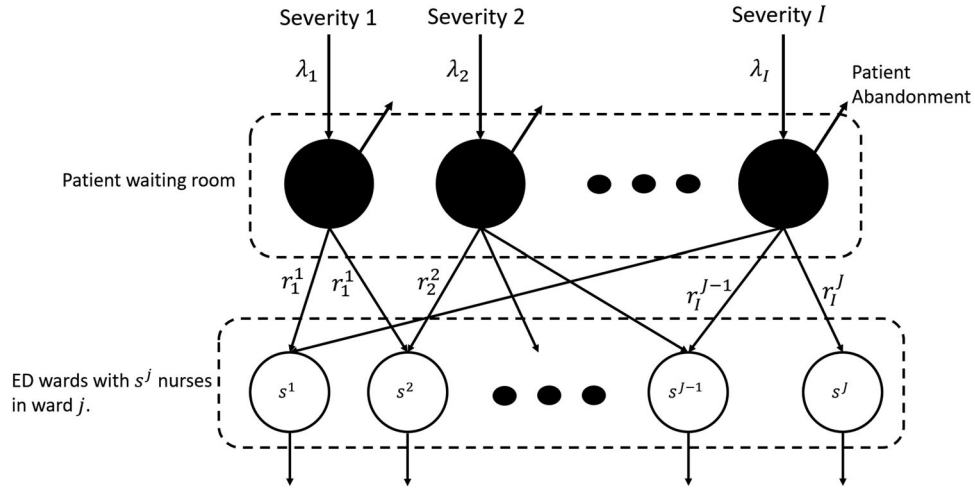


Figure 3. An overview of the patient flow process being considered in this article. Patients, following triage into one of several different severity levels, arrive to an ED and are assigned to a ward if space is available and depending on the given routing policy. If there is no space, patients wait until they are able to join a ward. Patients waiting for too long may abandon the system before entering service in a ward.

joining the queue but before starting service). We assume that the successive times to abandon for patients of severity type i are i.i.d random variables with CDF F_i . We note here that F_i is dependent on the workload present within a ward (as later described). Following ward assignment, the time spent by a patient in the ward before departure is assumed to be a random variable drawn from a distribution such that this time is a function of the number of nurses and the number of patients of all severity types in the ward.

Our model has two decisions to be made: staffing and routing. Staffing is the choice of numbers s^j for $1 \leq j \leq J$ that specifies how many nurses must be assigned to each ward while routing is the choice of numbers r_i^j for $1 \leq i \leq I$ and $1 \leq j \leq J$ that specify the proportion of incoming patients of severity type i that must be routed to ward j . Characterizing the queueing model described so far via closed-form expressions is difficult due to non-exponential distributions, pooled service, and workload-dependent service rates. In the next section, we will define an approximating fluid model that allows us to characterize the system via a set of equations.

2.2. Pooled service and workload dependent service

We assume that the time spent in service by a patient is a function of the number of nurses in the ward and the number of patients in the ward. In other words, the random variable S_i^j corresponding to the amount of time a patient of severity type i spends in service in ward j depends on the number of nurses s^j in ward j and the number of patients $\mathbf{n}^j = (n_1^j, \dots, n_I^j)$ of all severity types $i \in \{1, \dots, I\}$ in the ward. We characterize this dependency by assuming the mean value of this random variable $\mathbb{E}[S_i^j]$ to be the function m_i^j as

$$\mathbb{E}[S_i^j] \equiv m_i^j(s^j, \mathbf{n}^j) \rightarrow \mathbb{R}_{>0}.$$

Here, we do not specify the form of the function m_i^j , though a particular form is provided for the case study in Section 4.1.2.

The operations research literature that operationalizes workload metrics via mathematical modeling to balance/minimize nurse workload is sparse. Most of the existing work provides models within the context of an inpatient (Agor et al., 2017; Milburn, 2012) or home health care setting (Punnakitikashem et al., 2006; Sir et al., 2015), or attempt to balance and minimize workload by redesigning existing staffing methods (Wright et al., 2006). A recent article by Fishbein et al. (2019) takes the important step of reviewing objective measures of workload that can be obtained from electronic records to inform operationalization of workload measurement.

An important feature of our model is the ability to optimize staffing and routing while ensuring that the workload experienced by nurses in wards is maintained below pre-defined thresholds. We assume that the workload experienced by nurses in ward j depends both on the number of nurses in the ward s^j and the number of patients of each different severity type $\mathbf{n}^j = (n_1^j, n_2^j, \dots, n_I^j)$ according to the function γ_j as

$$\gamma_j(s^j, \mathbf{n}^j) \rightarrow \mathbb{R}_{\geq 0} \quad (1)$$

We will use this workload function γ_j to define constraints within our optimization model stating that the workload of all the wards be within some desired range determined by the decision maker. When performing experiments, we will consider two types of workload constraints. The first attempts to keep the workload of each ward under a pre-defined threshold while the second attempts to keep the absolute difference in workload across all pairs of wards under a pre-defined balance threshold.

3. The fluid model

The multiclass queueing model described in Section 2 is complex and difficult to express in a closed-form. As a result, it becomes difficult to formulate a model to optimize staffing and routing. Toward this, we approximate our stochastic queueing model by its fluid limit, which requires

scaling up the arrival rates of patients, ward capacity, and nurse staffing, while fixing abandonment-time and service time distributions. In what follows, in [Section 3.1](#) we briefly review relevant literature on fluid models and discuss why it is an effective approach for our problem. In [Section 3.2](#) we describe how to obtain the fluid limits using its stochastic pre-limit processes. In [Section 3.3](#), we develop a performance optimization problem based on fluid functions. In [Section 3.4](#), we explain how the performance of fluid solutions can be verified via computer simulations.

3.1. Literature review of fluid models

Our article is related to the vast literature on fluid approximations for queues. In what follows, we introduce different queueing models following the standard Kendall's notation;² see [Ross \(2019\)](#) for example. We hereby only review Many-Server Heavy-Traffic (MSHT) fluid queues that are most closely related to the present work. Heavy-traffic fluid and diffusion limits were developed by [Mandelbaum et al. \(1998\)](#) for time-varying Markovian queueing networks with Poisson arrivals and exponential service times. Adopting a two-parameter queue length descriptor, the pioneering work by [Whitt \(2006a\)](#) studied the $G/GI/s + GI$ fluid model having non-exponential service and abandonment times. [Whitt \(2006b\)](#) confirms that the discrete-time setting can be used as an approximation for the continuous-time setting (of the $G_t(n)/GI/s + GI$ model) as time increments of the discrete-time setting can be arbitrarily short. Scaling a stochastic system to its fluid limits has been shown to be asymptotically correct in the scaled regime for the Markovian $M/M/s + M$ model ([Mandelbaum & Pats, 1995](#); [Whitt, 2004](#)) and for a discrete-time analog of the general $G_t(n)/GI/s + GI$ model ([Whitt, 2006a](#)). Extending the work in [Whitt \(2006a\)](#), [Liu and Whitt \(2012a\)](#) developed a fluid approximation for the $G_t/GI/s_t + GI$ queue with time-varying arrivals and non-exponential distributions; they later extended it to the framework of fluid networks ([Liu and Whitt \(2011\)](#), [Liu and Whitt \(2014a\)](#)). A functional weak law of large numbers (FWLLN) ([Liu & Whitt, 2012b](#)) was established to substantiate the fluid approximation in [Liu and Whitt \(2012a\)](#) and functional central limit theorems (FCLTs) were developed for the $G_t/M/s_t + GI$ model by [Liu and Whitt \(2014b\)](#) and for the overloaded $G/GI/s + GI$ model by [Aras et al. \(2018\)](#).

3.1.1. Advantages of fluid models

The complexity of a stochastic system is often due to challenges in two separate dimensions: (i) time variability (non-stationary variability in time relative to its “steady-state” level) and (ii) stochastic variability (sample-path fluctuation relative to its sample average trajectory). As a limiting model

²Following Kendall's notation, a queueing model is specified by notation $Arr/Ser/s + Ab$, where “Arr” means the arrival process, “Ser” means the service distribution, s is the number of servers, and $+Ab$ means the abandonment-time distribution. For example, “M” means exponential distribution, G or GI means general (nonexponential) distribution, and a subscript t (e.g., M_t and G_t) indicates the time nonstationarity of the arrival process.

driven by FWLLN, a fluid model is useful for capturing the time variability while ignoring the stochastic variability. The biggest advantage of fluid analysis is its tractability. For queueing systems, exact analysis is often extremely challenging due to the sophistication of sample stochasticity and state-space discreteness. Fluid limits of queueing systems nicely address the above two issues by working with deterministic and continuous fluid processes (which are in general specified by a set of differential equations). We next give more in-depth discussions on the benefit of fluid models relative to two commonly adopted methods.

- **Fluid model vs. MDP.** Unlike the standard *Markov decision process* (MDP) analysis which often suffers from the curse of dimensionality, the large-scale assumption in fact becomes an advantage for the fluid model rather than a disadvantage. This is because the asymptotic optimality of fluid models requires that the system size grows large. Besides, the system's scale has no bearing on the analysis complexity and solution efficiency of the fluid dynamics because its performance functions arise from the asymptotic setting in which all “entities” (e.g., customers and servers) are shrunk down to infinitely divisible “atoms” of fluid.
- **Fluid model vs. simulation.** Comparing to simulation-based methods (e.g., sample average approximations and scenario generation), fluid models do not require building complex discrete-event simulation models, of which the accuracy relies on sufficiently large simulation budgets. In addition, fluid solutions are often in closed forms which can be used to generate useful structural insights. For example, the seminal work by [Whitt \(2006a\)](#) gives a clear-cut description of how the service time and abandonment-time distributions play a role in system performance functions and capacity sizing decisions. The analytic clarity of this result has opened a new research line and sparked many subsequent works. See for example [Liu and Whitt \(2012c\)](#) for an application to optimal staffing in queues with time-varying demand.

Past researchers have used fluid models to solve OR problems in service systems and healthcare; see [Anderson \(2014\)](#); [Dotoli et al. \(2009\)](#); [Yom-Tov and Mandelbaum \(2014\)](#); [Yousefi et al. \(2019\)](#).

3.2. From stochastic model to fluid limit

In our work, we follow the procedure outlined by [Whitt \(2006b\)](#), to perform the fluid scaling. Accordingly, we introduce a sequence of models indexed by a scaling parameter η , and then let $\eta \rightarrow \infty$. The arrival rates, maximum patient capacity in a ward, and number of servers are then set to be functions of η as

$$\frac{\lambda_i(\eta)}{\eta} \rightarrow \lambda_i, \quad \frac{M^j(\eta)}{\eta} \rightarrow M^j \quad \text{and} \quad \frac{s^j(\eta)}{\eta} \rightarrow s^j$$

as $\eta \rightarrow \infty$.

Thus, $\lambda_i(\eta) \approx \eta \lambda_i$ is the arrival rate of patients into the queueing model indexed by η but λ_i is the arrival rate of class- i fluid after scaling. Similar interpretations hold for $M^j(\eta)$ and $s^j(\eta)$.

Our fluid model is characterized by the parameter sextuple $(\lambda, \mathbf{x}, \mathbf{F}, \mathbf{r}, \mathbf{S}, \mathbf{s})$ where $\lambda \equiv (\lambda_1, \lambda_2, \dots, \lambda_I)$ is an I -tuple of numbers corresponding to arrivals, $\mathbf{F} = (F_1, \dots, F_I)$ is an I -tuple of CDFs corresponding to abandonment, $\mathbf{S} \equiv (S_i^j : 1 \leq j \leq J, 1 \leq i \leq I)$ is an $I \times J$ matrix of service time CDFs, $\mathbf{x} \equiv (x_i^j : 1 \leq j \leq J, 1 \leq i \leq I)$ is an $I \times J$ matrix of numbers corresponding to number of patients of each severity type in a ward, $\mathbf{r} \equiv (r_i^j : 1 \leq j \leq J, 1 \leq i \leq I)$ is an $I \times J$ matrix of numbers corresponding to patient routing proportions, and $\mathbf{s} \equiv (s_1, \dots, s^j)$ is a J -tuple of numbers corresponding to ward staffing.

To describe how the fluid model evolves over time, we define w_i as a deterministic time a fluid of class- i waits before entering service. This measure is relevant as the proportion of customers who do not abandon while waiting for service equals $F_i^c(w_i)$ (the CCDF of the abandonment distribution after class- i fluid has waited for time w_i).

One aspect of our model that is different from the framework outlined by Whitt (2006b) is in our relationship between offered load and service capacity. We begin by recognizing that fluids of two different classes within a ward do not interact. This is because the fluids of two different classes are able to share the same pool of nurses at the same time. This is unlike in a traditional queueing system where if one of the servers was occupied due to serving a particular class of fluid, that server is unavailable to other fluid classes. As a result, we define service capacity and offered load for any given fluid class independently from other fluid classes present in the same ward.

Before we define the system control equations, we note that since service time in our original queueing model is dependent on the number of patients, we require a scaling of the number of patients and need to represent it by a certain amount of fluid. Thus, if n_i^j is the number of patients of type i in ward j in the queueing model, we define x_i^j as its scaled counterpart in the fluid model as

$$\frac{n_i^j(\eta)}{\eta} \rightarrow x_i^j \quad \text{as} \quad \eta \rightarrow \infty.$$

The service time function $m_i^j(s^j, \mathbf{x}^j)$ can thus be defined as a fixed deterministic quantity since both s^j and \mathbf{x}^j are fixed deterministic numbers.

We can now express the system control equations for each ward in terms of the system control equations for each fluid type within the ward. In other words, we can express the system control equations via the expression “rate-in = rate-out” for each class- i fluid in ward j . Now, the arrival rate of class- i fluid entering service at ward j (which is also the “rate-in”) equals $\lambda_i r_i^j F_i^c(w_i)$. The first term (λ_i) is the overall arrival rate of fluid i . The second term (r_i^j) is the proportion of fluid i that is routed to ward j while the last term ($F_i^c(w_i)$) is the proportion of class- i fluid that does not

abandon after having waited for w_i units of time. The mean service time for class- i fluid entering ward j equals $m_i^j(s^j, \mathbf{x}^j)$. The rate-out thus equals the inverse of the mean service time ($m_i^j(s^j, \mathbf{x}^j)^{-1}$) multiplied by the service capacity (x_i^j) giving us the following control equation:

$$\lambda_i \times r_i^j \times F_i^c(w_i) = x_i^j \times m_i^j(s^j, \mathbf{x}^j)^{-1}, \quad \forall i, \forall j.$$

In addition we have the following sets of constraints to prevent fluid loss during routing

$$\sum_i r_i^j = 1, \quad \forall 1 \leq j \leq J.$$

Finally, we have a set of constraints to ensure that the total amount of fluid (of all classes) is capped in each ward j according to maximum capacity M^j as

$$\sum_i x_i^j \leq M^j \quad \forall 1 \leq j \leq J.$$

3.3. A fluid optimization problem

Before defining the objective of our optimization model, we first define the associated cost and reward coefficients. We break the objective function into three parts -

- A reward v_i^j is earned per ward for serving a unit of class- i fluid in ward j . Within the context of our ED setting, a higher reward is earned on serving more patients. The reward for class- i fluid is obtained for all fluid which has not abandoned after waiting w_i time units ($F_i^c(w_i)$), under routing proportions \mathbf{r} . The total reward for class- i fluid is given by $\lambda_i F_i^c(w_i) \sum_j r_i^j v_i^j$.
- A cost c_i^a is incurred for a unit of class- i fluid that abandons after waiting for time t . Within the context of our ED setting, a higher cost is incurred as more fluid ($\lambda_i \int_0^{w_i} dF_i(t)$) abandons the system after having to wait for w_i units of time. We assume a linear function for abandonment cost.
- A holding cost of $c_i^h(y)$ is incurred for having y units of class- i fluid waiting in queue. Here, y is the amount of class- i fluid waiting in queue in the fluid limit and is calculated using the expression $y = \lambda_i \int_0^{w_i} F_i(t) dt$. Within the context of our ED setting, a higher holding cost $c_i^h(\cdot)$ is incurred as more patients are forced to wait before being assigned to a bed in a ward.

We note here that our reward functions and coefficients are adapted from (Whitt, 2006b). We thus have the following expression for total reward.

$$R \equiv R(\mathbf{s}, \mathbf{r}, \mathbf{w}) = \sum_i \left(\lambda_i F_i^c(w_i) \sum_j r_i^j v_i^j - \lambda_i \int_0^{w_i} c_i^a dF_i(t) - c_i^h(\lambda_i \int_0^{w_i} F_i(t) dt) \right). \quad (2)$$

We note here that the above expression does not explicitly minimize a patient’s LOS. However, by attempting to reduce wait times with abandonment penalties, the model

incentivizes the system to establish smart routing and staffing policies that lead to faster patient service. This in turn ensures that patient wait time is reduced further downstream in the wait queues. The complete *fluid optimization problem* (FOP) may now be written as follows

$$\begin{aligned}
& \underset{\mathbf{w}, \mathbf{r}, \mathbf{s}}{\text{maximize}} && \sum_i (\lambda_i F_i^c(w_i) \sum_j r_{i,j} v_i^j - \lambda_i \int_0^{w_i} c_i^a dF_i(t)) \\
& \text{subject to} && -c_i^h(\lambda_i \int_0^{w_i} F_i^c(t) dt) \\
& && \lambda_i r_i^j F_i^c(w_i) m_i^j(s^j, \mathbf{x}^j) = x_i^j, \\
& && 1 \leq j \leq J, \quad 1 \leq i \leq I \\
& && \sum_j r_i^j = 1, \quad 1 \leq i \leq I \\
& && \sum_j x_i^j \leq M^j, \quad 1 \leq j \leq J \\
& && \Psi(\gamma^j(s^j, \mathbf{x}^j), \quad \forall j) \in \psi \\
& && 0 \leq \mathbf{r} \leq 1 \\
& && \sum_j s^j = \Theta \\
& && 0 \leq \mathbf{w}
\end{aligned} \tag{3}$$

Here, Θ is the maximum number of servers available for assignment. We have not placed any restriction on the nature of the functions $\gamma^j(s^j, \mathbf{x}^j)$ and $m_i^j(s^j, \mathbf{x}^j)$. Presumably, $\gamma^j(s^j, \mathbf{x}^j)$ would increase with an addition to the amount of fluid (\mathbf{x}^j) in the ward and would decrease with the addition of servers (s^j). However, we do not place any restrictions on the functional forms. Similarly, we do not place any restrictions on the form of $m_i^j(s^j, \mathbf{x}^j)$.

The workload constraint ($\Psi(\gamma^j(s^j, \mathbf{x}^j), \quad \forall j) \in \psi$) is a key component within our model. When performing numerical analyses, we consider two types of workload constraints, 1) workload threshold constraint, and 2) workload balance constraint. The workload threshold constraint guarantees that the maximum allowable workload for each ward to be under a certain pre-defined limit. The constraint thus takes the form $\gamma^j(s^j, \mathbf{x}^j) \leq \Gamma_j, \quad \forall j$. The workload balance constraint on the other hand aims to keep the absolute difference in workload between any two pairs of wards below a pre-determined value. In this scenario, the constraint takes the form $|\gamma^j(s^j, \mathbf{x}^j) - \gamma^k(s^k, \mathbf{x}^k)| \leq \Gamma^b, \quad \forall j, k$. We note here that both of these constraints as we define them are linear (or easily linearized) and do not pose additional modeling complexity offered by the first flow balance constraint.

We note here that additional constraints may be included depending on the decision maker's requirements. An example of such constraints would be to specify a minimum number of nurses required in any given ward. Another example would be to specify that some patient severity types be routed only to a certain ward.

To test the performance of the optimal strategy obtained via the fluid optimization model described earlier, we developed a computer simulation that functions as a virtual abstraction of the real ED. Details about the simulation model are provided in the following subsection.

Remark 1. The deterministic FOP (3) largely reduces the complexity of the original stochastic optimization problem by

omitting the random fluctuation. Nevertheless, the solution to the FOP is still not straightforward. Indeed, the problem is clearly non-convex (because neither the constraints or the objective are convex), thus there is no guarantee that the optimal solution ought be unique. Also see Lee et al. (2021) and Whitt (2006a) for discussions on why uniqueness of fluid-based optimization problems can be challenging. Although the technical investigation of FOP's existence and uniqueness is not the focus of the article, we hope to provide some guideline on how to select an initial feasible solution to the FOP. This will be useful because, as soon will become clear in Section 3.4, our FOP will be solved by commercial solvers which require an available feasible policy as an initial candidate solution. Determining an initial feasible solution can be fairly straightforward if the existing routing and staffing policies being used within the emergency department being considered are available. Given these existing staffing and routing values for r_i^j and s^j (which are inherently feasible on account of being the existing policy values we wish to optimize), solving the first equation within the optimization model gives us an initial value for w_i , which gives us a full set of initial feasible values for our decision variables. In the event that information about the existing policy is unavailable, it is sufficient to identify values for \mathbf{r} and \mathbf{s} that satisfy the constraints $\sum_j r_i^j = 1$, $\sum_i x_i^j \leq M^j$, and $\sum_j s^j = \Theta$. This is due to the fact that the only remaining variable w_i within the flow balance constraint is flexible and can be adjusted to ensure constraint feasibility as $F_i^c(w_i)$ always lies between 0 and 1.

3.4. Performance validation via simulations

We test the performance of our fluid approximation by analyzing the approximate model against the original queueing model. As we discussed earlier, analyzing the queueing model in its closed form is difficult; thus we developed a simulation to represent the dynamics of the queueing model. We developed the simulation using AnyLogic software's personal learning edition. Each new agent within the simulation is generated from one of $I=5$ different source modules (one for each severity type) with an inter-arrival time distributed exponentially with a rate value as shown in Figure 1. If all delay modules (representing wards) are at capacity, the patient enters a queue module (representing the wait room). On entry to the queue module, a random variable is drawn from the CDF for the patient's abandonment distribution. Once a patient has waited in the queue module for an amount of time equal to the drawn random variable, the patient is pushed out of the module and the counter for abandonment of the patient's severity type is incremented by one.

Patients are routed to wards according to pre-defined routing proportions (read in from an external file). Once a patient enters a ward, the time that they will spend in the ward is determined by drawing a random variable from an exponential distribution with a rate function (m_i^j) that depends on the patient's severity type, the ward that the patient is in, and the patient-nurse ratio of the ward. It must be noted here that each time a patient enters or leaves

a ward, the simulation draws a new random variable for each patient's remaining time in service. This allows us to effectively capture the memoryless property of the state-dependent exponential distribution that we assume for a patient's time in service.

Output statistics include average patient LOS and average ward workload. The average patient LOS includes the time spent by a patient waiting in queue and the time in service. The average ward workload is obtained by averaging the workload of the ward calculated from Eq. (1) over the model's time horizon. We use a time horizon of one year, which begins after a warm-up period (set as two weeks in our simulation). To collect summary statistics, we run 20 replications leading to a total run-time of 5 minutes for each simulation experiment on a Windows 10-based personal device consisting of an 8-core, 16-thread, 3.6 GHz CPU and 32GB of RAM.

3.4.1. Solution procedure to optimize and analyze routing and staffing

The full solution procedure to optimize and analyze patient routing and ward staffing policies is outlined as follows.

Step 1. *Solving the fluid optimization problem in Eq. (3).* We note here that the fourth constraint corresponding to the workload constraint is modified and chosen according to the experiment being considered. To solve the optimization model, we use a nonlinear programming solver `fmincon` provided in MATLAB's Optimization Toolbox. Specifically, we use the Sequential Quadratic Programming (SQP) method provided within the solver. In this method, a Quadratic Programming (QP) subproblem is solved at each iteration with an estimate of the Hessian of the Lagrangian being updated at each iteration.

Step 2. *Constructing operational policies for the queueing model based on fluid solutions.* The solution of the fluid optimization problem, specifically the routing proportions and staffing levels r_i^j and s^j , are then fed into the stochastic simulation model in AnyLogic software.

Step 3. *Validating effectiveness of the fluid-based results using simulations.* We run multiple replications of the simulation model and store the value of average patient LOS and ward workload for each replication. We record the result for optimal patient LOS and ward workloads as the average across all the replications.

We note here that the optimization model presented in Eq. (3) does not constrain staffing decisions to be integer valued. However, staffing is often discussed in terms of the number of personnel, which is an integer value. In the case study presented later, we ensure integer staffing values from the output of the fluid model by solving for the optimal routing policy for all possible staffing combinations and selecting the best objective over all staffing combinations. While more sophisticated algorithms may be employed, our method is efficient for our case study. One can also incorporate additional staffing constraints by restricting the

generated staffing combinations as desired, as we do in the case study.

Before proceeding to describe the data for our case study, we wish to remind the reader about the importance of obtaining engineering confirmation that the control equations representing the fluid model match up with the real system in the scaled regime. We refer the reader to previous work by Nambiar (2020) for an empirical discussion about how values of $\eta > 10$ lead to a practically significantly accurate match between the control equations representing the fluid model and the real system in the scaled regime.

4. Numerical analysis and case study

We demonstrate an application of the approach developed in Section 3 by considering data from a hospital in North Carolina as an experimental case study. We thus outline the data available to us and how we inferred the various input parameters from the data to inform the fluid optimization and simulation. Consider the hospital described in Section 1.1, with patients triaged into one of five severity types (with 1 being the highest) and assigned to one of three wards with a maximum of 11 nurses. The arrival rates and routing proportions were shown in Figure 1 along with the current assignment of nurses. For the optimization problem, we assume that there are always at least 3 nurses assigned to the ward seeing the most severe patients.

4.1. Data Analysis and experiment settings

To fully characterize our fluid and queueing/simulation models, we require the following six sets of parameter estimates related to patient flow: (1) arrival rate by patient severity type (λ_i), (2) current nurse staffing levels for each ward during the status quo (s^j), (3) maximum ward capacity (M^j), (4) CDF for patient abandonment for each patient severity type (F_i), (5) routing proportion of patient severity type to each ward (r_i^j), and (6) rate function for the time spent by a patient in service (μ_i^j). We note here that any mention of status quo henceforth in this article refers to the set of operational parameters being used in the hospital during our observation period. Our goal is to modify and optimize the parameters corresponding to nurse staffing (s^j) and patient routing (r_i^j). As mentioned earlier, (1)-(3) above are shown in Figure 1. What remains is to describe patient abandonment estimates and to provide functional form for mean patient service time categorized by patient severity type and ward.

4.1.1. Patient abandonment estimates

Estimating the CDF for patient abandonment from data is not trivial, due to the hospital being unable to keep records of when a patient abandons. Though the data had a small percentage of patient departures from the system coded as LWBS (left without being seen), this number refers to those patients who, after triage, had been assigned to a bed but departed before being seen by a nurse or physician. The lack of

Table 1. Mean patient service time categorized by patient severity type and ward. pn in the above expressions refers to patient-nurse ratio of the ward.

Severity Type (i)	Ward (j)	Mean Service Time (m_i^j)
1	Critical Care	$185.1 - 6.54pn + 0.86pn^2$
2	Critical Care	$140.37 + 22.52pn - 0.87pn^2$
3	Critical Care	$84.63 + 28.06pn - 1.05pn^2$
4	Critical Care	$113.07 + 3.78pn - 0.09pn^2$
5	Critical Care	$21.52 + 14.97pn - 0.29pn^2$
1	Minor Care	N/A
2	Minor Care	$224.05 + 39.05pn - 2.25pn^2$
3	Minor Care	$107.81 + 25.39pn - 1.12pn^2$
4	Minor Care	$43.75 + 19.70pn - 0.77pn^2$
5	Minor Care	$56.46 + 10.48pn - 0.05pn^2$
1	Fast Track	N/A
2	Fast Track	N/A
3	Fast Track	$149.97 + 1.5pn - 0.05pn^2$
4	Fast Track	$76.55 + 7.55pn - 0.28pn^2$
5	Fast Track	$172.33 + 25.46pn - 1.5pn^2$

sufficient data meant that we assumed patients were unlikely to abandon in our model unless they waited for an extremely long period of time. We thus assumed in our model an exponential function for patient abandonment distribution with mean values assumed to be 15, 12, 10, 8, 6 hrs for patient severity types 5, 4, 3, 2, and 1, respectively. Such a distribution ensured that the probability of patient abandonment remained low unless they waited for an unreasonable amount of time. For instance, the probability of patients of severity type 1 abandoning becomes greater than 10% only after waiting at least about 1.5 hours before being assigned to a ward. We note here that average patient wait times were never that high (e.g., 15 hours for Severity type 1) in the fluid model or simulation and therefore patients were not likely to abandon as a result of using this distribution. However, the CDF of the abandonment distribution plays a critical role in determining the final steady-state performance of the system. We refer the reader to the work by Whitt (2006b) for details.

4.1.2. Patient time in service estimates

In Section 1.1 we described the methodology we used to estimate how patient time in service varies based on the workload experienced by nurses within the wards. We noted that the mean patient time in service can be represented using polynomial functions per ward, number of other patients in the ward, and number of nurses in the ward. The resulting equations for each patient severity type within each ward are provided in Table 1. We note here that some combinations for patient type and ward are listed as N/A since there was not enough data to infer any sort of a functional form. These included patients of severity type 1 in minor care and fast track wards and patients of severity type 2 in the fast track ward. Accordingly, we restrict these routing combinations in our optimization model while performing experimental analyses. In other words, we enforce constraints that force the model to prevent patients of type 1 from going to minor care and fast track wards, and patients of type 2 from going to fast track wards.

We note from Table 1 that the functional form of the mean patient service time for patients of severity 1 within the critical care ward is different compared to all of the other combinations for severity type and assigned ward. The functional form

Table 2. Baseline parameter values used in case study.

Parameter	Description	Values for Severity 1-5
u_i	Workload	(10, 8, 7, 3, 2)
v_i	Throughput reward	(200, 150, 100, 50, 10)
c_a	Abandonment cost	(10, 15, 12, 8, 4)
c_i^h	Waiting cost	(10, 7, 4, 2, 1)

for m_1^1 indicates that the mean service time for severity type 1 patients decreases in the patient-nurse ratio until the ratio reaches about 3.5, and starts to increase beyond that point. A ratio this high is unusual for this system as Type 1 patients are the most severe, require critical care, and have the lowest arrival rate of any type. If there were many critically ill patients this would strain resources beyond any efficiencies that could be gained by changes in service, such as early discharge, as these strategies may not be possible with patients requiring critical care. While a detailed analysis of the relationship between functional form and optimal policy is outside the scope of this work, we note the solution to the fluid model is independent of functional form.

4.1.3. Ward workload function

We assume a linear function for workload by separating the patient-nurse ratio terms by each patient severity type. The workload function thus takes the form

$$\gamma^j(s^j, \mathbf{n}^j) = \sum_i u_i^j \frac{n_i^j}{s_j^j}, \quad (4)$$

where u_i^j is a measure of workload experienced by a single nurse caring for a single patient of severity i in ward j . For instance, the value of u_i^j is higher for patients of higher severity (say, severity 1 compared to severity 3) within the same ward. Similarly, the value of u_i^j is higher for patients of similar severity receiving care in the critical care ward as opposed to the minor care ward. While performing numerical experiments we assumed similar workload coefficients u_i^j across wards and only assumed differences across patient severity types. We did this to allow for easier representation of the workload measure when attempting to keep it under desired thresholds or to balance it across wards. However, we note that differences in patient health outcomes across wards are captured by variation in coefficient values for patient service time for patients of the same severity type across different wards, as seen in Table 1. The parameter values chosen for the case study are listed in Table 2. As can be seen they indicate that more severe patients lead to higher level of workload, reward for completion, and cost for abandoning and holding.

Having defined the operational parameters for our case study, we next proceed to conduct numerical analyses and optimize existing patient routing and ward staffing strategies.

4.2. Numerical results and discussions

Before conducting numerical analyses, we first calibrate patient arrival rates within the simulation and fluid model

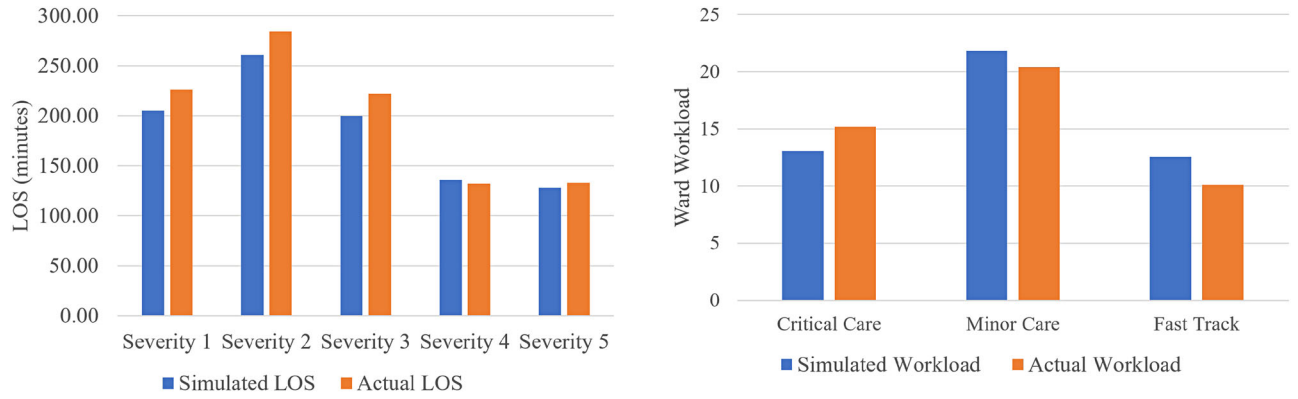


Figure 4. A comparison of patient LOS and ward workloads using the calibrated arrival rates within the simulation vs actual values obtained from data.

in order to better match the patient LOS computed from data from the real system. We note here that we chose to calibrate the arrival rates (instead of other parameters such as service times or abandonment probability) because our model inherently ignores structural characteristics of the system that lead to the actual arrival rates varying over time. For instance, the fast track ward at the hospital is only open from 7 am to 10 pm and all patients arriving outside of those times are sent to the critical care ward. Our analysis of the data to estimate patient arrival rates did not include this structural characteristic and therefore calibrating patient arrival rates helps in better estimating actual arrival rates. **Figure 4** compares the patient LOS for each of the five patient severity types obtained from simulating the system using calibrated patient arrival rates against the patient LOS as computed directly from the data. We note that real patient LOS post-calibration is a close match to the simulated LOS. The workload of the real system and the simulated workload are also shown in **Figure 4** on the right (note that we do not calibrate to workload directly).

We test the fluid optimization model by comparing the performance of the resulting optimal staffing and routing policies against the status quo. Specifically, we look at patient LOS and nurse workload, separated by patient severity type and ward, respectively. Recall that the fluid optimization model provides us with estimates in the scaled (“fluid”) regime. As a result, we need to use the results from the fluid optimization model to obtain patient LOS and nurse workload estimates from the real system via simulation.

In the remainder of the numerical analysis section, we describe multiple experimental scenarios to showcase the flexibility offered by our methodology before providing results for each experimental scenario.

4.2.1. Numerical analyses & optimal strategies

In this section we wish to optimize the routing and staffing values for the hospital’s ED being considered in our case study under different experimental settings that aim to control for workload by either balancing it between wards or by minimizing it across wards. Under each experimental scenario, we follow the solution procedure from **Section 3.4.1** to

obtain optimal values for patient LOS and ward workload then simulate the LOS and workload resulting from these routing and staffing policies. In addition to LOS for each patient severity type, we compute the weighted average for LOS across all severity types using arrival rates as the weights for each patient severity type. We compare our results to status quo by simulating the current routing and staffing rules.

Recall here the addition of a few constraints based on observations from data such as 1) patients of severity type 1 only being admitted to critical care and not minor care or fast track, and 2) patients of severity type 2 only being admitted to critical care and minor care and not to fast track. These constraints allow us to demonstrate special problem structure that may be necessary from an implementation standpoint, such as ensuring that patients of higher severity are not sent to a ward with nurses who are not equipped to handle them.

The experimental studies show the flexibility offered by the model to a decision maker in setting their desired operations goals for the system and are outlined below.

- **Optimization w/o workload constraints:** Here, we solve the fluid optimization model in **Eq. (3)** without the workload constraint.
- **Optimization w/workload threshold constraints:** Here, we solve the fluid optimization model in **Eq. (3)** with the workload constraint formulated to keep the long-run average workload of each ward under a pre-defined threshold (γ^*). The objective here is to minimize patient LOS while ensuring that ward workload does not exceed the specified thresholds. Specifically, the constraint set is formulated as

$$\gamma^j(s^j, \mathbf{x}^j) \leq \gamma^*, \quad \forall j = 1, 2, 3.$$

While performing the experiment, we test multiple values of γ_j^* which is kept the same for all three wards. Specifically, we vary the value of γ_j^* between 20 and 14 for all j . Intuitively, the workload of a ward having one patient of each severity type being cared for by two nurses equals 15. Additionally, we note from **Figure 4** that one of the three wards experiences a workload value

Table 3. Simulation results from experiments restricting ward workloads. The LOS for each severity type and the average weighted LOS is given for each value of γ^* , as well as the percent change in LOS when compared to status quo.

	LOS					Weighted Average
	Sev 1	Sev 2	Sev 3	Sev 4	Sev 5	
Status Quo	204.95	260.55	199.82	135.83	127.91	222.47
∞	199.80	229.67	213.47	172.41	147.15	216.96
	-2.5%	-11.9%	6.8%	26.9%	15.0%	-2.5%
20	207.50	220.02	277.52	127.11	39.57	234.46
	1.2%	-15.6%	38.9%	-6.4%	-69.1%	5.4%
18	204.58	238.40	187.73	148.74	129.02	208.18
	-0.2%	-8.5%	-6.0%	9.5%	0.9%	-6.4%
16	228.27	231.83	208.50	158.12	111.05	214.82
	11.4%	-11.0%	4.3%	16.4%	-13.2%	-3.4%
14	196.25	211.20	177.13	129.72	33.67	188.79
	-4.2%	-18.9%	-11.4%	-4.5%	-73.7%	-15.1%

of over 20 and our goal is to manage routing and staffing such that all wards experience workload values under our pre-defined thresholds.

- **Optimization w/workload balance constraints:** Here, we solve the fluid optimization model in Eq. (3) with the workload constraint formulated to keep the difference in workload between any two pairs of wards (j, k) under a pre-defined threshold ($\hat{\gamma}$). The objective here is to balance the workload across wards while also reducing patient LOS. Specifically, the constraint set is formulated as

$$|\gamma^j(s^j, \mathbf{x}^j) - \gamma^k(s^k, \mathbf{x}^k)| \leq \Gamma^b, \quad \forall, (j, k) \in \{(1, 2), (2, 3), (1, 3)\}$$

While performing the experiment, we test three values for Γ^b including 7.5, 5, and 2.5. To provide some intuition behind this number, let us consider two wards with one patient of each severity type in each ward. A difference in workload value of 2.5 would mean that one ward has 4 nurses while the other has 3. A difference in ward workload of 2.5 is highly restrictive as the number of nurses is limited and in order to satisfy the workload balance constraint, the model would need to increase the number of nurses in all the wards. We note here that a unique value of Γ^b could be chosen for each pair of wards being considered for workload balance.

Under each experimental setup, we compare a variety of performance measures, including LOS, a weighted average measure of patients LOS (\overline{LOS}), workload, and workload differences. To compute the weighted average LOS, we use a combination of arrival rate and severity weights, represented by objective function reward coefficient v_i . These weights were chosen to appropriately weight patient traffic and patient severity. Thus, the weighted average LOS (\overline{LOS}) is calculated as

$$\overline{LOS} = \frac{\sum_i (LOS_i \times \lambda_i \times v_i)}{\sum_i (\lambda_i \times v_i)}$$

Finally, since we assume a fixed value for the total number of nurses in our model, we discuss the possible

Table 4. Simulation results from experiments restricting ward workloads. Workload in each ward and maximum difference in workload between wards.

γ^*	Critical Care	Minor Care	Fast Track	Max Diff.
Status Quo	13.08	21.84	12.58	9.26
∞	11.80	19.70	30.38	18.58
20	20.34	12.19	9.42	10.92
18	13.75	18.11	17.73	4.36
16	14.46	16.48	14.76	2.02
14	13.64	16.02	14.21	2.38

improvements in system performance as a result of increasing the total number of nurses available for staffing.

4.2.1.1. On restricting workload threshold. Simulated results from our experiments varying the workload threshold γ^* are shown in Table 3 and Table 4. Table 3 shows LOS for each severity type as well as \overline{LOS} , and the percent difference in LOS compared to status quo as a percent shown below each LOS value. Table 4 presents the workload in each ward, as well as the maximum difference in workload between wards. First, from Table 3 we see that the unconstrained problem provides lower average weighted LOS by decreasing the LOS of severity 1 and severity 2 patients; however this value is close to the results obtained from simulating the routing and staffing values currently being used at the hospital (status quo), indicating that the current policy does a good job at maintaining low LOS. In addition, initially, constraining the workload threshold leads to increased LOS, but as the workload threshold is decreased we see even further reductions in average weighted LOS. When $\gamma^* = 14$, we see reduced LOS for all severity types and over a 15% reduction in average weighted LOS.

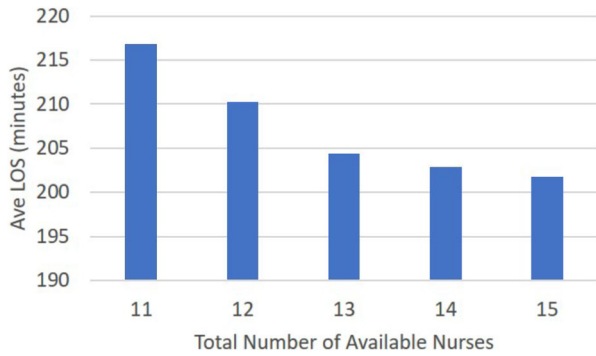
From Table 4 we see that status quo has a large workload in minor care, and the unconstrained problem also results in high workload and high workload imbalance. As we tighten the workload constraint, not only does the workload decrease, but so does the difference in workload. When we look at both the LOS and workload results together we see that tightening the workload constraint not only has a positive effect on workload, but also indirectly improves LOS. Even though the optimization model does not directly optimize LOS, this result makes sense since LOS depends on workload.

We note here that the simulated workloads may exceed the theoretical workload constraints (γ^*). This is because while the solution from the fluid model is able to satisfy the fluid constraints, the estimate for workload obtained from inputting the solution of the fluid optimization model into the simulation does not necessarily need to satisfy the threshold constraint as the fluid model result is a deterministic approximation of the stochastic system.

4.2.1.2. On balancing ward workload. Results for changes in the workload balance constraint are shown in Table 5. Here we show both the LOS and workload results in one table. The status quo and unconstrained results are shown again for comparison. Decreasing the value of Γ^b not only improves workload balance as expected, but also results in lower average weighted LOS. While not shown here, we find

Table 5. Simulation results from experiments balancing workload. LOS for each type as well as workload for each ward are shown.

Workload Balance (Γ^b)	LOS					Weighted Average	Workload			
	Sev 1	Sev 2	Sev 3	Sev 4	Sev 5		Critical Care	Minor Care	Fast Track	Max Diff.
Status Quo	204.9	260.5	199.8	135.8	127.9	222.5	13.08	21.84	12.58	9.26
∞	199.8	229.7	213.5	172.4	147.1	217.0	11.80	19.70	30.38	18.58
7.5	191.8	212.0	239.3	135.4	38.9	215.5	17.78	18.23	9.19	9.03
5	220.5	241.1	191.8	150.4	129.6	211.5	13.78	19.46	18.83	5.69
2.5	222.0	244.0	192.6	149.8	100.6	213.1	14.08	17.72	16.41	3.64

**Figure 5.** Results for increasing resources.

that the results are not very sensitive to the cost and reward parameters, as long as the cost/reward for higher severity patients is higher than that of other patients. The solution is sensitive to the workload constraints, as shown above.

4.2.1.3. On increasing available resources. In all of our experiments we assumed a fixed number of nurses and attempted to redistribute the nurses across wards. However, we note from Sections 4.2.1.1 and 4.2.2.1 that attempting to optimize for patient LOS and ward workload leads to relatively small improvements. This indicates to us that the current system is already operating at capacity and to see significant improvements in performance, we may require additional nurses/resources. Thus, we test the impact of adding nurses (with no workload constraints). As is evident from Figure 5, increasing the number of resources available leads to a reduction in average patient LOS. The reduction in LOS is around 3% on adding one additional nurse, and increases to about 7% on adding 4 additional nurses. However, the marginal benefit decreases in the number of nurses. This may be attributed once again to the U-shaped nature of the patient service time curve. Essentially, adding a large number of nurses means that they are under less pressure to work faster and this effect is seen in average measures of patient LOS. However, the key takeaway here is that in order for the hospital to significantly reduce patient LOS under the current ward system, it is necessary for them to increase the number of available nurses on staff. We note here that a more rigorous cost benefit analysis is required before concluding whether adding staff members is economically viable and that our analysis is focused purely on operational improvements in terms of patient LOS and nurse workload.

4.2.2 Discussion. The key takeaways from our numerical analyses are as follows. First, our modeling framework of

optimizing for staffing and routing within the fluid approximation and inputting the results into a simulation model provides us with an efficient means of improving patient LOS and ward workload estimates for a hospital's ED. Second, we note that in our case study, despite being able to see improvements in ward workload, the optimal staffing and routing policies from our model without workload constraints do not necessarily lead to significant improvements in patient LOS. This indicates that the hospital's existing routing and staffing protocols already perform reasonably well as far as our objective function is concerned but has room for improvement as far as nurse workload is concerned. Restricting workload does lead to both reduced workload and reduced LOS if the constraints are not too loose. Finally, we note that in our case study, achieving significant improvements in both patient LOS and ward workload requires an increase in the number of available staff.

5. Conclusion

We developed in this article a framework to improve patient LOS and nurse workload by adjusting staffing and routing policies for a hospital ED modeled as a multiclass multi-server queueing system with pooled service and state-dependent service times. We used a hybrid method by combining fluid approximations to queues and simulation to solve the combined routing and staffing problem. We used data from the emergency department of a regional hospital in North Carolina to conduct a case study showing the implementation of our framework. Our analyses showed that making small modifications to the routing proportions and staffing policies can lead to reduction and better balance of ward workload levels, without negatively impacting patient LOS. We must note here our data did not provide us with any information about patient recidivism or outcomes. It is likely that patients who are cared for under high patient-nurse ratio values return to the hospital or experience worse outcomes despite departing initially after a smaller time spent in the ward. Furthermore, we note here our assumption that patient outcomes are not dependent on the time they spend in the system. In other words, we do not account for long-term patient health outcomes or whether a patient after discharge left the ED to go home or was admitted to the hospital. We leave this for future research.

A natural question that arises from an implementation standpoint is how to use the new routing proportions to send patients to wards. We suggest that the optimized proportions obtained from running mathematical models such as ours must be implemented on an aggregate scale to account for

temporal and staffing fluctuations that occur as a result of normal operations within an emergency department.

Furthermore, our work assumes that workload, while dynamic, is not explicitly dependent on time. However, the time since a nurse's shift started may impact their workload. Future research could involve developing time-varying proportions that take into account some of these drawbacks.

An important future direction of research could consider the use of transient analysis instead of fluid approximations to analyze the complex queueing models we developed in this article. While fluid approximations are useful in analyzing the average behavior of the system, it does not account for any of the stochastic behavior. Most real systems rarely settle into a steady-state, and the ability to analyze a system in its transient state is often computationally intractable. Furthermore, the results of a transient analysis is a function of the initial conditions of the system, something that a steady-state analysis does not consider. Studying the model developed in this article in its transient state could be a potential future direction of research.

A second important direction for future research stems from consideration for more personalized patient service rates in the queueing model. During the numerical analyses within this article, we considered five patient classes to match the five levels of the Emergency Severity Index (ESI) triage algorithm adopted by the hospital in our case study to classify patients. We then inferred functional forms for patient time in service by separating these five severity types depending on the ward they were in, thus leading to 15 possible combinations for the patient time in system functions. However, closer inspection of patient time in service for each of these combinations indicates that a higher level granularity may be possible. For example, separating the severity type 3 patients into two groups, one requiring higher and the other requiring lower patient times in service, could lead to increased modeling accuracy. Determining the separation threshold (for example, patients requiring more or less than 600 minutes of service) point would require the use of classification algorithms like decision trees trained on data available in the ED such as the primary reason for admission, mode of admission, and initial diagnosis. This framework of increasing the granularity of patient classes within the queueing and fluid models would be better at predicting patient time in service which would then lead to more accurate patient routing and nurse staffing policies.

Finally, we note that the framework established in this article can have applications well beyond the field of health-care and can benefit any service system that involves customer arrivals into one of several different server pools, such as in wireless networks (Qadir et al., 2016) where resource pooling involves abstracting a collection of networked resources to behave like a single unified resource pool.

Consent and approval

This study has been exempt from the requirement for approval by an institutional review board as it uses only secondary data.

Role of the funder

NSF funded the research grant that in part supported Nambiar as a PhD student and some of Mayorga's effort during the grant period.

Disclosure statement

The authors report no conflict of interest

Funding

This work was supported in part by the National Science Foundation (NSF) under NSF award number SCH-1522107. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s), and do not necessarily reflect those of the NSF.

ORCID

Siddhartha Nambiar  <http://orcid.org/0000-0003-1779-564X>
 Maria E. Mayorga  <http://orcid.org/0000-0002-6399-2153>
 Yunan Liu  <http://orcid.org/0000-0001-9961-2610>

References

- Agor, J., McKenzie, K., Mayorga, M. E., Ozaltin, O., Parikh, R. S., & Huddleston, J. (2017). Simulating triage of patients into an internal medicine department to validate the use of an optimization-based workload score. In *Proceedings of the 2017 Winter Simulation Conference* (p. 234). IEEE.
- AHRQ. (2018). *Section 1. The need to address emergency department crowding.* <https://www.ahrq.gov/research/findings/final-reports/ptflow/section1.html>
- Almehdawe, E., Jewkes, B., & He, Q.-M. (2013). A Markovian queueing model for ambulance offload delays. *European Journal of Operational Research*, 226(3), 602–614. <https://doi.org/10.1016/j.ejor.2012.11.030>
- Anderson, R. M. (2014). *Stochastic models and data driven simulations for healthcare operations* [PhD thesis]. Massachusetts Institute of Technology.
- Anderson, D., Price, C., Golden, B., Jank, W., & Wasil, E. (2011). Examining the discharge practices of surgeons at a large medical center. *Health Care Management Science*, 14(4), 338–347. <https://doi.org/10.1007/s10729-011-9167-6>
- Aras, A. K., Chen, X., & Liu, Y. (2018). Many-server Gaussian limits for non-Markovian queues with customer abandonment. *Queueing Systems*, 89(1–2), 81–125. <https://doi.org/10.1007/s11134-018-9575-0>
- Armony, M., & Ward, A. R. (2010). Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research*, 58(3), 624–637. <https://doi.org/10.1287/opre.1090.0777>
- Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., & Yom-Tov, G. B. (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1), 146–194. <https://doi.org/10.1287/14-SSY153>
- Ata, B., & Shneorson, S. (2006). Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science*, 52(11), 1778–1791. <https://doi.org/10.1287/mnsc.1060.0587>
- Ball, J. E., Bruyneel, L., Aiken, L. H., Sermeus, W., Sloane, D. M., Rafferty, A. M., Lindqvist, R., Tishelman, C., Griffiths, P., Consortium, R., & RN4Cast Consortium. (2018). Post-operative mortality, missed care and nurse staffing in nine countries: A cross-sectional study. *International Journal of Nursing Studies*, 78, 10–15.
- Batt, R. J., & Terwiesch, C. (2012). Doctors under load: An empirical study of state-dependent service times in emergency care. <https://doi.org/10.1287/orsc.1120.0848>

- faculty.wharton.upenn.edu/wp-content/uploads/2012/11/DULnew_v6.pdf
- Berry Jaeker, J. A., & Tucker, A. L. (2017). Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science*, 63(4), 1042–1062. <https://doi.org/10.1287/mnsc.2015.2387>
- Carayon, P., & Wood, K. E. (2009). Patient safety. *Information Knowledge Systems Management*, 8(1–4), 23–46. <https://doi.org/10.3233/IKS-2009-0134>
- Carnes, K. M., de Riese, C. S., & de Riese, W. T. (2015). A cost-benefit analysis of medical scribes and electronic medical record system in an academic urology clinic. *Urology Practice*, 2(3), 101–105. <https://doi.org/10.1016/j.urpr.2014.10.006>
- Chan, C. W., Farias, V. F., & Escobar, G. B. (2017). The impact of delays on service times in the intensive care unit. *Management Science*, 63(7), 2049–2395. <https://doi.org/10.1287/mnsc.2016.2441>
- Chan, C. W., Huang, M., & Sarhangian, V. (2021). Dynamic server assignment in multiclass queues with shifts, with applications to nurse staffing in emergency departments. *Operations Research*, 69(6), 1936–1312. <https://doi.org/10.1287/opre.2020.2050>
- Chan, C. W., Sarhangian, V., Talwai, P. M., & Gogia, K. (2020). *Utilizing partial flexibility to improve emergency department flow: Theory and implementation*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4002563
- Cohen, I., Mandelbaum, A., & Zychlinski, N. (2014). Minimizing mortality in a mass casualty event: Fluid networks in support of modeling and staffing. *IIE Transactions*, 46(7), 728–741. <https://doi.org/10.1080/0740817X.2013.855846>
- Considine, J., LeVasseur, S. A., & Villanueva, E. (2004). The Australasian triage scale: Examining emergency department nurses' performance using computer and paper scenarios. *Annals of Emergency Medicine*, 44(5), 516–523. <https://doi.org/10.1016/j.annemergmed.2004.04.007>
- Dai, J., & Shi, P. (2019). Inpatient overflow: An approximate dynamic programming approach. *Manufacturing & Service Operations Management*, 21(4), 894–911. <https://doi.org/10.1287/msom.2018.0730>
- Denton, B. T. Ed. (2013). *Handbook of healthcare operations management*. Springer, pp. 978–971.
- Derlet, R. W., & Richards, J. R. (2000). Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Annals of Emergency Medicine*, 35(1), 63–68. [https://doi.org/10.1016/S0196-0644\(00\)70105-3](https://doi.org/10.1016/S0196-0644(00)70105-3)
- Dinh, M. M., Green, T. C., Bein, K. J., Lo, S., Jones, A., & Johnson, T. (2015). Emergency department clinical redesign, team-based care and improvements in hospital performance: A time series analysis. *Emergency Medicine Australasia*, 27(4), 317–312. <https://doi.org/10.1111/1742-6723.12424>
- Dotoli, M., Fanti, M. P., Mangini, A. M., & Ukovich, W. (2009). A continuous petri net model for the management and design of emergency cardiology departments. *IFAC Proceedings Volumes*, 42(17), 50–55. <https://doi.org/10.3182/20090916-3-ES-3003.00010>
- Fishbein, D., Nambiar, S., McKenzie, K., Mayorga, M., Miller, K., Tran, K., Schubel, L., Agor, J., Kim, T., & Capan, M. (2019). Objective measures of workload in healthcare: A narrative review. *International Journal of Health Care Quality Assurance*, 33(1), 1–17. <https://doi.org/10.1108/IJHCQA-12-2018-0288>
- Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2), 79–141. <https://doi.org/10.1287/msom.5.2.79.16071>
- George, J. M., & Harrison, J. M. (2001). Dynamic control of a queue with adjustable service rate. *Operations Research*, 49(5), 720–731. <https://doi.org/10.1287/opre.49.5.720.10605>
- Haraden, C., & Resar, R. (2004). Patient flow in hospitals: Understanding and controlling it better. *Frontiers of Health Services Management*, 20(4), 3–15.
- Harrison, J. M., & Zeevi, A. (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management*, 7(1), 20–36. <https://doi.org/10.1287/msom.1040.0052>
- Helm, J., AhmadBeygi, S., & Van Oyen, M. (2011). Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management*, 20(3), 359–374. <https://doi.org/10.1111/j.1937-5956.2011.01231.x>
- Holden, R. J., Scanlon, M. C., Patel, N. R., Kaushal, R., Escoto, K. H., Brown, R. L., Alper, S. J., Arnold, J. M., Shalaby, T. M., Murkowski, K., & Karsh, B.-T. (2011). A human factors framework and study of the effect of nursing workload on patient safety and employee quality of working life. *BMJ Quality & Safety*, 20(1), 15–24. <https://doi.org/10.1136/bmjqs.2008.028381>
- Jun, J., Jacobson, S. H., & Swisher, J. R. (1999). Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society*, 50(2), 109–123. <https://doi.org/10.1057/palgrave.jors.2600669>
- Kc, D., & Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9), 1486–1498. <https://doi.org/10.1287/mnsc.1090.1037>
- Kc, D., & Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*, 14(1), 50–65. <https://doi.org/10.1287/msom.1110.0341>
- Kim, S.-H., Pinker, E. J., & Rimar, J. (2021). *Do care providers take an individual patient perspective or a system perspective? A study of the effect of ICU capacity strain on patient discharge*. Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2644600
- Kleinrock, L. (1967). Time-shared systems: A theoretical treatment. *Journal of the ACM*, 14(2), 242–261. <https://doi.org/10.1145/321386.321388>
- Lamy Filho, F., da Silva, A. A., Lopes, J., Lamy, Z. C., Simões, V. M., & Santos, A. M. d (2011). Staff workload and adverse events during mechanical ventilation in neonatal intensive care units. *Jornal de Pediatria*, 87(6), 487–492. <https://doi.org/10.2223/JPED.2140>
- Lee, C., Liu, X., Liu, Y., & Zhang, L. (2021). Optimal control of a time-varying double-ended production queueing model. *Stochastic Systems*, 11(2), 140–173. <https://doi.org/10.1287/stsy.2019.0066>
- Liu, Y. (2018). Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research*, 66(2), 514–534. <https://doi.org/10.1287/opre.2017.1678>
- Liu, Y., & Whitt, W. (2011). A network of time-varying many-server fluid queues with customer abandonment. *Operations Research*, 59(4), 835–846. <https://doi.org/10.1287/opre.1110.0942>
- Liu, Y., & Whitt, W. (2012a). The $G_t/GI/s_t/GI$ many-server fluid queue. *Queueing Systems*, 71(4), 405–444.
- Liu, Y., & Whitt, W. (2012b). A many-server fluid limit for the $G_t/GI/s_t/GI$ queueing model experiencing periods of overloading. *Operations Research Letters*, 40, 307–312.
- Liu, Y., & Whitt, W. (2012c). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research*, 60(6), 1551–1564. <https://doi.org/10.1287/opre.1120.1104>
- Liu, Y., & Whitt, W. (2014a). Algorithms for time-varying networks of many-server fluid queues. *INFORMS Journal on Computing*, 26(1), 59–73. <https://doi.org/10.1287/ijoc.1120.0547>
- Liu, Y., & Whitt, W. (2014b). Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability*, 24(1), 378–421.
- Liu, Y., & Whitt, W. (2014c). Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, 28(4), 419–449. <https://doi.org/10.1017/S0269964814000084>
- Liu, Y., & Whitt, W. (2017). Stabilizing performance in a service system with time-varying arrivals and customer feedback. *European Journal of Operational Research*, 256(2), 473–486. <https://doi.org/10.1016/j.ejor.2016.07.018>
- Liu, Y., Sun, X., & Hovey, K. (2022). Scheduling to differentiate service in a multiclass service system. *Operations Research*, 70(1), 527–544. <https://doi.org/10.1287/opre.2020.2075>

- Magalhães, A. M. M., Costa, D. G., Riboldi, C. O., Mergen, T., Barbosa, A. S., & Moura, G. M. S. S. (2017). Association between workload of the nursing staff and patient safety outcomes. *Revista da Escola de Enfermagem da USP*, 51, e03255. <https://doi.org/10.1590/s1980-220x2016021203255>
- Mandelbaum, A., & Pats, G. (1995). State-dependent queues: Approximations and applications. *Stochastic Networks*, 71, 239–282.
- Mandelbaum, A., Massey, W. A., & Reiman, M. (1998). Strong approximations for Markovian service networks. *Queueing Systems*, 30(1/2), 149–201. <https://doi.org/10.1023/A:1019112920622>
- Mandelbaum, A., Momčilović, P., & Tseytlin, Y. (2012). On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science*, 58(7), 1273–1291. <https://doi.org/10.1287/mnsc.1110.1491>
- Mazur, L. M., Mosaly, P. R., Moore, C., Comitz, E., Yu, F., Falchook, A. D., Eblan, M. J., Hoyle, L. M., Tracton, G., Chera, B. S., & Marks, L. B. (2016). Toward a better understanding of task demands, workload, and performance during physician-computer interactions. *Journal of the American Medical Informatics Association*, 23(6), 1113–1120. <https://doi.org/10.1093/jamia/ocw016>
- Milburn, A. B. (2012). Operations research applications in home healthcare. In *Handbook of healthcare system scheduling* (p. 281–302). Springer.
- Mitchell, P., & Golden, R. (2012). *Core principles & values of effective team-based health care*. National Academy of Sciences.
- Murray, M. J. (2003). The Canadian triage and acuity scale: A Canadian perspective on emergency department triage. *Emergency Medicine*, 15(1), 6–10.
- Nambiar, S. (2020). *Improving hospital operational efficiency when staff workload affects patient outcomes*. <https://www.proquest.com/openview/068241c01115851f4ef5d5ef5dd3ce80/1?pq-origsite=gscholar&cbl=18750&diss=y>
- Nicosia, F. M., Park, L. G., Gray, C. P., Yakir, M. J., & Hung, D. Y. (2018). Nurses' perspectives on lean redesigns to patient flow and inpatient discharge process efficiency. *Global Qualitative Nursing Research*, 5, 2333393618810658. <https://doi.org/10.1177/2333393618810658>
- Parenti, N., Reggiani, M. L. B., Iannone, P., Percudani, D., & Dowding, D. (2014). A systematic review on the validity and reliability of an emergency department triage scale, the Manchester triage system. *International Journal of Nursing Studies*, 51(7), 1062–1069. <https://doi.org/10.1016/j.ijnurstu.2014.01.013>
- Powell, S. G., & Schultz, K. L. (2004). Throughput in serial lines with state-dependent behavior. *Management Science*, 50(8), 1095–1105. <https://doi.org/10.1287/mnsc.1040.0233>
- Punnakitikashem, P., Rosenberger, J. M., Behan, D. B., Baker, R. L., & Goss, K. (2006). An optimization-based prototype for nurse assignment. In *Proceedings of the 7th Asian Pacific Industrial Engineering and Management Systems Conference* (pages 17–20). Citeseer.
- Qadir, J., Sathiaselan, A., Wang, L., & Crowcroft, J. (2016). "Resource pooling" for wireless networks: Solutions for the developing world. *ACM SIGCOMM Computer Communication Review*, 46(4), 30–35. <https://doi.org/10.1145/3027947.3027953>
- Ross, S. (2019). *Introduction to probability models*. Academic Press.
- Saghafian, S., Austin, G., & Traub, S. J. (2015). Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering*, 5(2), 101–123. <https://doi.org/10.1080/19488300.2015.1017676>
- Saghafian, S., Hopp, W. J., Van Oyen, M. P., Desmond, J. S., & Kronick, S. L. (2012). Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60(5), 1080–1097. <https://doi.org/10.1287/opre.1120.1096>
- Saghafian, S., Hopp, W. J., Van Oyen, M. P., Desmond, J. S., & Kronick, S. L. (2014). Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management*, 16(3), 329–345. <https://doi.org/10.1287/msom.2014.0487>
- Sir, M. Y., Dundar, B., Steege, L. M. B., & Pasupathy, K. S. (2015). Nurse-patient assignment models considering patient acuity metrics and nurses' perceived workload. *Journal of Biomedical Informatics*, 55, 237–248.
- Swan, B., Ozaltin, O., Hilburn, S., Gignac, E., & McCammon, G. (2019). Evaluating an emergency department care redesign: A simulation approach. In *2019 Winter Simulation Conference (WSC)* (pp. 1137–1147). IEEE. <https://doi.org/10.1109/WSC40007.2019.9004947>
- Tanabe, P., Gimbel, R., Yarnold, P. R., & Adams, J. G. (2004). The emergency severity index (version 3) 5-level triage system scores predict ED resource consumption. *Journal of Emergency Nursing*, 30(1), 22–29. <https://doi.org/10.1016/j.jen.2003.11.004>
- Upenieks, V. V., Kotlerman, J., Akhavan, J., Esser, J., & Ngo, M. J. (2007). Assessing nursing staffing ratios: Variability in workload intensity. *Policy, Politics, & Nursing Practice*, 8(1), 7–19. <https://doi.org/10.1177/1527154407300999>
- Ward, A. R., & Armony, M. (2013). Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research*, 61(1), 228–243. <https://doi.org/10.1287/opre.1120.1129>
- Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50(10), 1449–1461. <https://doi.org/10.1287/mnsc.1040.0279>
- Whitt, W. (2006a). Fluid models for multiserver queues with abandonments. *Operations Research*, 54(1), 37–54. <https://doi.org/10.1287/opre.1050.0227>
- Whitt, W. (2006b). A multi-class fluid model for a contact center with skill-based routing. *AEU - International Journal of Electronics and Communications*, 60(2), 95–102. <https://doi.org/10.1016/j.aeue.2005.11.005>
- Wright, P. D., Bretthauer, K. M., & Côté, M. J. (2006). Reexamining the nurse scheduling problem: Staffing ratios and nursing shortages. *Decision Sciences*, 37(1), 39–70. <https://doi.org/10.1111/j.1540-5414.2006.00109.x>
- Yom-Tov, G. B., & Mandelbaum, A. (2014). Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2), 283–299. <https://doi.org/10.1287/msom.2013.0474>
- Yousefi, N., Hasankhani, F., & Kiani, M. (2019). Appointment scheduling model in healthcare using clustering algorithms. *arXiv Preprint. arXiv:1905.03083*