Managing Emergency Department Patient Flow and Nurse Staffing: A Fluid Queueing Model with Multiple Customer Classes and Service Pools

Siddhartha Nambiar\textsuperscript{a}, Maria E. Mayorga\textsuperscript{b}, and Yunan Liu\textsuperscript{b}

\textsuperscript{a}National Center for Human Factors in Healthcare, MedStar Health Research Institute; \textsuperscript{b}Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University

\textbf{ARTICLE HISTORY}
Compiled May 26, 2021

\textbf{ABSTRACT}
Efficient patient flow through an emergency department (ED) is a critical factor that contributes to a hospital’s performance, which in turn influences overall patient health outcomes. In this work, we model a multi-class many-server pooled queueing system where patients of varying acuity receive care from one of several nurse pools, each comprising an ED ward. We assume that a patient’s time in a ward is a function of the nurse workload (for which nurse-patient ratio is used as a proxy) in their unit. Our objective is to reduce patient Length-of-Stay (LOS) and to control nurse workload by optimizing patient routing and nurse allocation decisions between wards. We describe the computational challenges in formulating and solving the queueing model representing our problem. We tackle the complexity by first approximating the set of queueing equations via a deterministic fluid model that describes the queueing equations by their first-order behaviors. Next, we formulate and solve an optimization model using the first-order control equations and input the results into a discrete-event simulation to obtain performance measures of interest for the non-approximated system, such as patient LOS and ward workload. Finally, we demonstrate the utility of our approach via a case study using retrospective data from a hospital in North Carolina, USA. Aside from showcasing the flexibility our model offers to healthcare decision-makers, our case study highlights the importance of accounting for nurse workload and service behavior in developing routing and staffing policies.

\textbf{KEYWORDS}
Patient flow; emergency department; staffing; routing; simulation; queueing theory; fluid approximations

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 and 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 and 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 difficult 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) (J 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 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 and Resar, 2004). Furthermore, there is a recent focus on better understanding the impact on 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 workload include increased mortality in the 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 LOS

We motivate our work using findings from preliminary analyses of how workload affects patients LOS based on data from a regional hospital in North Carolina. 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 had 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 Fig. 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 ward nurses and the total patient LOS. As universal measures for nurse workload does not exist (as detailed later in Section 2), we used the ratio of patients to nurses in a ward as a proxy measure for workload.

![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.](image1)

We created a scatter-plot (Fig. 2) to examine the relationship between the average patient-nurse ratio (calculated as a time-weighted average of patient-nurse ratio during a patient’s stay) and LOS for every patient in our dataset. We note here that we excluded data points with LOS values greater than 24 hours to remove outliers. We refer the reader to work by Swan et al. (2019) for details on how the the time-average for patient-nurse ratio is calculated.

![Figure 2. Scatter-plot of average patient time in service against average patient-nurse ratio](image2)

The density of Fig. 2 makes it difficult to see patterns. Thus, we aggregated average patient-nurse ratio (x-axis) into buckets and calculated the average value of patient time in service (y-axis) within each of those buckets, as shown in Fig. 3. We now see that the curve follows a distinct polynomial form. We fit the curve in Fig. 3 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
Figure 3. Average patient time in service 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.

combinations of patient severity types and wards shows similar trends (we expand further on this aspect in Section 4). Considering the form of Fig. 3, we see that a patient’s time in service 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 service begins to decrease. We contend that this decrease is due to the nurses attempting to provide faster service due to the knowledge of the high workload that they are facing with so many patients to care for. Though workload is measured in different ways, this observation is well-supported by prior literature that noted an inverted u-shaped relationship between patient time in system and provider workload (Batt and Terwiesch (2012); Berry Jaeker and Tucker (2017); Keunecke et al. (2019)).

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

1.2. Summary of work

In this paper we model a multi-class many-server pooled queueing system where patients of different acuity levels receive care from one of several nurse pools, each comprising an ED ward. We assume that a patient’s time in service is a function of the ratio of nurses to patients in their ward (as shown in Fig. 2 and Fig. 3). Our model reduces patient Length-of-Stay (LOS) and controls nurse workload by optimizing routing and nurse allocation decisions between the wards. First, we assess the complexity of the queueing model control equations arising due to patients of multiple severity levels existing within the same ward. We note that the time the patients spend in the ward (a measure of the queueing system’s service rate) depends on the patient acuity mix and the nurse workload within their ward. We address this complexity by approximating the queueing system via a deterministic fluid model to describe the control equations via their first-order behaviors and formulate and solve an optimization model using these control equations. The solution to the fluid optimization problem
is input to a discrete-event simulation to obtain performance measures of interest for
the non-approximated system, such as patient LOS and ward workload. The purpose
of the simulation is similar to the work done by Dotoli et al. (2009) where the authors
used a simulation model to show that the policy resulting from the optimal decision
parameters obtained from the fluid model is more efficient than the status quo.

We demonstrate the validity of our approach via a case study using data and system
characteristics from the emergency department of a regional hospital in North Car-
olina. We use this retrospective data to model patient arrivals, patient assignment to
wards, nurse assignment to wards, and patient LOS. Once validated against existing
data, we run various ‘what-if’ scenarios by rerouting patients and re-assigning nurses
based on the information obtained from solving the fluid optimization model. We pro-
vide a detailed analytical treatment for each of the ‘what-if’ scenarios and discuss the
merits and demerits of implementing them. Our results (1) highlight the importance of
accounting for nurse workload and service behavior in developing routing and staffing
policies and (2) show that small changes to patient routing policies could lead to re-
duced patient LOS and better-balanced nurse workloads. We thus develop a framework
to optimize both routing and staffing while controlling individual ward workloads.

An outline of the remainder of this paper is as follows. Section 2 gives an overview of
relevant literature. Section 3 then describes the emergency department setting that we
abstract to form a queuing model followed by an approximation to form a fluid model.
Furthermore, we describe the development of a simulation model and the various
input functions and parameters used to build it using data from the hospital. Finally,
in Section 4, we conduct several experiments by optimizing patient routing and nurse
staffing under different constraints and discuss the optimal solution and its implications
under each case.

2. Review of Literature

The present work is relevant to staffing and routing within multi-class multi-server
problems in healthcare and general service systems, queues considering fairness,
queues with shared server capacity, and fluid approximations of queuing systems. An
overview of these four streams is provided below.

Routing and staffing to manage patient flow. The field of operations re-
search/management science has played a significant role in improving the performance
of hospital EDs. The scope of this work has ranged from developing aggregated perfor-
mance measures to benchmark ED performance measures (Kang et al., 2017), using
predictive analytics to characterize how patient flow is affected by information about
current state of an ED (Peck et al., 2014), and developing analytical models to dy-
namically determine inpatient staffing rates (Broyles et al., 2011). Hall et al. (2006)
discuss strategies, concepts, and methods specific to patient flow that can improve
the delivery of health care by reducing delays. The authors discuss the approach of
analyzing patient flow from a holistic perspective by looking at the effects of manag-
ing individual patients and providers on the entire health system. Zeltyn et al. (2011)
created a simulation model of an emergency department to address several operational
problems related to patient flow. Specifically, they centered their model and questions
about patient flow around the topic of efficient physician and nurse staffing.

The question of staffing is often closely related to routing as optimal staffing de-
PENDs on the choice of routing rule used, and vice versa (Gans et al., 2003). However,
Harrison and Zevei (2005) mention that nurse staffing and patient routing are often treated in a separate but hierarchical manner due to the computational complexity involved. Gans et al. (2003) provides an overview of some of this complexity by invoking cell center models where complex functional forms for customer arrivals, service duration, and customer abandonments can lead to intractable solutions. Furthermore, the authors state that incorporating human behavior such as fatigue, for both customers and servers, adds to the complexity in developing queueing models. We refer the reader to work by Saghaian et al. (2015) for a more comprehensive review of papers that discuss ways of improving a wide range of processes involving patient flow from the initial call to the ED through disposition, discharge home, or admission to hospital.

**Service fairness.** A key feature of the model in our paper is the consideration of nurse workload and its fair distribution across hospital wards. While we will discuss specifics regarding nurse workload in Section 3.2.2, we note here that the topic of “fairness” in large-scale service systems has been recognized by a number of authors in the past. Traditional queueing models focus on ensuring fairness from a customer’s perspective. Fairness towards servers, however, is not easily defined and depends heavily on the nature of the service/industry under consideration. Many call centers follow a longest-idle-server-first (LISF) routing policy; that is, newly arriving calls are routed to the server that has experienced the longest idle time (Armony and Ward, 2010). Other allocation policies considered in recent literature that attempt to achieve fairness among servers include Randomized Most Idle (Mandelbaum et al., 2012a), and Longest Idle Pool First (Atar et al., 2011). Tseytlin (2009) presents various queueing models with heterogeneous servers and also provides a literature review on fairness in service systems from the behavioral sciences point of view. The concept of fairness towards servers can also be found in the set of literature concerning load balancing and communication networks. We refer the reader to references provided by Alanyali et al. (1998), Afek et al. (1999), Bartal et al. (2002), and Rubenstein et al. (1999) for further background on this area.

**Shared service capacity.** Next, we note that few analytical models for resource allocation in healthcare consider the fact that resources within wards are partially shared, central resources (Schmidt et al., 2013). In general service systems, the use of pooled 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 ward 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. (2012b) 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 and Ward, 2010; Ward and Armony, 2013).

**Fluid approximations for queueing systems.** Finally, we review relevant works on fluid approximations for queues. Fluid models have been predominantly adopted to characterize and approximate performance of large-scale queueing systems. Due to the
large body of work on fluid models, 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. 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 and 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 with exponential service times by Liu and Whitt (2014b) and for the overloaded $G/GI/s + GI$ model with exponential service and abandonment times by Aras et al. (2018).

3. Model Description & Formulation

We describe our model formulation over the following four subsections. Section 3.1 describes the abstraction of the process of patient flow through an ED as a queueing model. In Section 3.2 we specify two key salient features of our model (that of pooled patient service and consideration for nurse workload) and define them mathematically. In Section 3.3 we introduce a fluid approximation to the queueing model and develop an optimization model. Finally in Section 3.4 we describe the development of the discrete-event simulation of patient flow.

3.1. Queueing Model

In this section, 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 pooled service from all nurses in the ward.

A pictorial representation of the patient flow process is provided in Fig. 4. 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 $r^j_i$ with $\sum_{j=1}^{J} r^j_i = 1$ for $1 \leq i \leq I$. In other words, a proportion $r^j_i$ of patient 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 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 (further details in Section 3.5). Following ward
Figure 4. An overview of the patient flow process being considered in this paper. 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.

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 \( r_{ij} \) 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.

### 3.2. Salient Model Features

Our model incorporates two important salient features that make it a more accurate abstraction of reality and adds to the novelty of our work. The first is the fact that patients are assumed to receive service from the entire pool of nurses within their ward. As part of this assumption, we also contend that a patient’s time in service is influenced by and dependent on the number of other patients and nurses in the ward they are in. The second salient feature of our model is the consideration of nurse workload. We discuss each of these features in this subsection.

#### 3.2.1. Pooled Patient Service

In modern-day healthcare practices, patients are rarely looked after by just one health professional. Gone are the days when patients were cared for by a single physician or a private nurse residing in the community (Mitchell and Golden, 2012). Modern healthcare is complex and has evolved such that health professionals work as members of teams who share common aims with respect to patient outcomes (Mitchell and
Traditional models of queueing theory deal with service systems in which a customer is serviced by exactly one server (Gross, 2008). Models incorporating multiple servers tend to assume that each one serves a different customer. When there are more customers than servers, there typically exists a waiting room within which customers wait until a server is available to see them. A variation of this type of model is the “Processor Sharing” service discipline (Dudin et al., 2018). Here, customers are all served simultaneously with each receiving an equal fraction of the available service capacity. Processor sharing models have been applied most extensively in the field of multiprogramming computer systems (Baskett and Gomez, 1972).

In this paper, we contend that the service being provided by a pool of nurses in a hospital ward to the patients within the ward is similar to that of a “processor sharing” service discipline. We assume that adding or reducing more patients to a ward changes the service rates of all the patients within the unit. As we discuss in the next subsection, we incorporate this by modeling a patient’s time in the system using an exponential distribution with its rate being a function of the nurse-patient ratio in the unit.

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 of all different severity types in the ward. In other words, the random variable $S_j^i$ corresponding to the amount of time a patient of severity type $i$ spends in service in ward $j$ depends on number of nurses $s^j$ in ward $j$ and the number of patients $n_j = (n_j^1, ..., n_j^I)$ of all severity types $i \in \{1, ..., I\}$ in the ward. We characterise this dependency by assuming the mean value of this random variable $\mathbb{E}[S_j^i]$ to be the function $m_j^i$ as

$$\mathbb{E}[S_j^i] \equiv m_j^i(s^j, n^j_j) \to \mathbb{R}_{>0}.$$ 

Here, we do not specify the form of the function $m_j^i$, this is discussed further in Section 4.2.

### 3.2.2. Nurse Workload

Fairness within a healthcare setting is often characterized by the workload experienced by nurses and other hospital staff (Carayon and Gurses, 2008). Hooey et al. (2017) define workload as “the task demand of accomplishing mission requirements for the human operator”. Several studies have shown that higher nursing workload has adverse effects on patient safety (Al-Kandari and Thomas, 2009; Carayon and Gurses, 2008; Cohen et al., 1999). Aside from being correlated with sub-optimal patient care, high levels of nurse workload have also been shown to lead to reduced patient satisfaction (Anderson et al., 1998). Quantifying nurse workload is not an easy task, as workload is a complex construct. While the idea of nursing workload is broadly related to the number of tasks a nurse needs to complete, a universally accepted definition for nursing workload does not exist (Carayon and Gurses, 2008). Different researchers and healthcare organizations define or calculate the workload experienced by the nurses in a number of different ways. Some use the severity levels of a nurse’s patients to calculate the workload (Lang et al., 2004) while others have the nurses fill out questionnaires to evaluate their workload subjectively (van den Oetelaar et al., 2016). Some objective workload metrics from the literature include ‘nurses per occupied bed’ (Twigg and Duffield, 2009), ‘nurse to patient ratio’ (Hurst, 2008), and ‘relative value units’ (RVUs) (Glass and Anderson, 2002).
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 paper 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 predefined 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 $n_j = (n_{j_1}, n_{j_2}, .., n_{j_I})$ according to the function $\gamma_j$ as

$$\gamma_j(s_j, n_j) \rightarrow \mathbb{R}_{\geq 0}$$

We will use this workload function $\gamma_j$ to define a constraint within our optimization 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 threshold across all pairs of wards under a pre-defined balance threshold.

### 3.3. Fluid Approximation

The queueing model we defined in Section 3.1 is complex and is difficult to express in a closed-form. As a result, it becomes difficult to formulate a model to optimize staffing and routing. To handle this complexity, we approximate our queueing model by scaling up the arrival rates of patients, ward capacity, and nurse staffing, while fixing abandonment distribution, and service time distributions. 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 and Pats, 1995; Whitt, 2004) and for a discrete-time analog of the general $G_t(n)/GI/s + GI$ model (Whitt, 2006b). Furthermore, Whitt (2006c) argues that the discrete-time setting may be used as an approximation for continuous-time setting (of the $G_t(n)/GI/s + GI$ model) as time increments of the discrete-time setting can be arbitrarily short. This characteristic of fluid queues allows us to model complex arrival and service rate functions and to obtain tractable analytic solutions.

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). In our work, we follow the procedure outlined by Whitt (2006c), to perform the fluid scaling. Accordingly, we introduce a sequence of models indexed by a scaling parameter $\eta$, 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 $\eta$ as

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

Thus, $\lambda_i(\eta) \approx \eta \lambda_i$ is the arrival rate of patients into the queueing model indexed by $\eta$. 

10
but $\lambda_i$ 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 seven-tuple $(\lambda, x, F, r, S, s)$ where
\[
\lambda \equiv (\lambda_1, \lambda_2, \ldots, \lambda_J) \text{ is an I-tuple of numbers corresponding to arrivals, } F = (F_1, \ldots, F_I)
\]
is an I-tuple of CDFs corresponding to abandonment, $S \equiv (S^I_i : 1 \leq j \leq J, 1 \leq i \leq I)$ is an $I \times J$ matrix of service time CDFs, $x \equiv (x^I_i : 1 \leq j \leq J, 1 \leq i \leq I)$ is an $I \times J$ matrix of numbers corresponding to number of patients of each severity type in a ward, $r \equiv (r^I_i : 1 \leq j \leq J, 1 \leq i \leq I)$ is an $I \times J$ matrix of numbers corresponding to patient routing proportions, and $s \equiv (s_1, \ldots, s^J)$ 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 $F^i_x(w_i)$ (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 (2006c) 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 dependant on the number of patients, we require a scaling given fluid class independently from other fluid classes present in the same ward.

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 $\lambda_i r^I_i F^i_x(w_i)$. The first term $(\lambda_i)$ is the overall arrival rate of fluid $i$. The second term $(r^I_i)$ is the proportion of fluid $i$ that is routed to ward $j$ while the last term $(F^i_x(w_i))$ 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 $m^i_x(s^j, x^j)$. The rate out thus equals the inverse of the mean service time $(m^i_x(s^j, x^j)^{-1})$ multiplied by the service capacity $(x^j)$ giving us the following control equation:
\[
\lambda_i \times r^I_i \times F^i_x(w_i) = x^j \times m^i_x(s^j, x^j)^{-1}, \forall i, \forall j.
\]
In addition we have the following sets of constraints to prevent fluid loss during routing

\[ \sum_i r_j^i = 1, \forall 1 \leq j \leq J. \]

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

\[ \sum_i x_j^i \leq M_j \forall 1 \leq j \leq J. \]

### 3.3.1. Fluid Optimization Model

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 -

1. **A reward rate** \( v(i,j,t) \) earned per ward for serving class-\( i \) fluid in ward \( j \) after customers have waited for time \( t \). This reward decreases in \( t \). Within the context of our ED setting, a higher reward \( v(\cdot) \) is earned as more patients are served after a shorter waiting period.
2. **Cost** \( c^a(i,t) \) incurred per unit of time for fluid of class \( i \) that abandons after waiting for time \( t \). Within the context of our ED setting, a higher cost \( c^a(\cdot) \) is incurred as more patients are forced to abandon the system after having to wait for longer periods of time.
3. **Holding cost** of \( c^h(i,y) \) incurred per unit time 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 

\[ y = \lambda_i \int_0^w F_{ci}(t) dt. \]

Within the context of our ED setting, a higher holding cost \( c^h(\cdot) \) 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, 2006c). We thus have the following expression for total reward.

\[
R = R(s, r, w) = \sum_i \left( \lambda_i F_{ci}^c(w_i) \sum_j r_{i,j} v(i,j,w_i) - \lambda_i \int_0^w c^a(i,t) dF_i(t) - c^h(i, \lambda_i \int_0^w F_{ci}(t) dt) \right).
\]

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 optimization model may now be written...
as follows

\[
\text{maximize } \sum_i \left( \lambda_i F_i^w(w_i) \sum_j r_{i,j} v(i,j,w_i) - \lambda_i \int_0^{w_i} c^3(i,t) dF_i(t) - c^5 \left( i, \lambda_i \int_0^{w_i} F_i(t) dt \right) \right)
\]

subject to
\[
\lambda_i r_i F_i^c(w_i) m_j^i(s_j, x_{i,j}) = x_{i,j}, \quad 1 \leq j \leq J, \quad 1 \leq i \leq I
\]
\[
\sum_j x_{i,j} \leq M_j, \quad 1 \leq j \leq J
\]
\[
\Psi(\gamma_j^i(s_j, x_j), \forall, j) \in \psi
\]
\[
0 \leq r \leq 1
\]
\[
\sum_j s_j = \Theta
\]
\[
0 \leq w
\]

Here, \( \Theta \) is the maximum number of servers available for assignment. We have not placed any restriction on the nature of the functions \( \gamma_j^i(s_j, x_j) \) and \( m_j^i(s_j, x_j) \). Presumably, \( \gamma_j^i(s_j, x_j) \) would increase with an addition to the amount of fluid \( (x_j) \) 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 \( m_j^i(s_j, x_j) \).

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 not be routed to specific wards.

To test the performance of the optimal strategy obtained via the 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.

### 3.4. Simulation Model

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 Fig. 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 predefined 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 \((m_j)\) 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).

3.5. 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. Optimize fluid model.** Solve the fluid optimization problem in Eq. (3). We note here that the fourth constraint corresponding to the workload constraint may be modified to according to the experiment being considered.

**Step 2. Translate fluid approximation solution to simulated queueing model.** The solution of the fluid optimization problem, specifically the routing proportions and staffing levels \(r_j^i\) and \(s_j^i\), is then input to the simulation model in AnyLogic software.

**Step 3. Simulate solutions and analyze results.** We run multiple replications of the simulation model and store the value of average patient LOS and ward workload for each replication. Record the result for optimal patient LOS and ward workloads as the average across all the replications.

We note here that in Step 2 above, the solution we obtain for staffing level \(s_j^i\) from the fluid optimization model is a continuous value. However, staffing is often discussed in terms of the number of personnel, which is an integer value. To generate the integer values from the output from the fluid model, we consider all possible combinations for floor and ceiling values of the non-integral \(s_j^i\) such that the total number of nurses is the same. We ignore those combinations that lead to infeasibility. Out of all feasible combinations, we calculate the objective value for each and select the combination with the best objective value. However, we also note here that integer solutions for staffing levels are not always necessary as staffing levels can be represented as a fractional value of Full Time Equivalent (FTE) hours. We are able to observe (results not shown) while performing experiments that those policies that do not rely on heuristics to obtain integral staffing values lead to better weighted patient LOS and ward workload values. All experimental results we show use the aforementioned heuristic procedure that determines integer staffing values to allow for ease of interpretation of staffing results.

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 et al. (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. Case Study Data Description

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 feed into the fluid optimization and simulation.

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 ($\lambda_i$), (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 ($r_{ji}$), and (6) rate function for the time spent by a patient in service ($\mu_j^i$). We note here that any mention of status quo henceforth in this paper 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 ($r_{ji}$).

Details regarding arrival rates, staffing levels, ward capacities, and routing proportions are provided in Section 1.1. 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. 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 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 a weibull function for patient abandonment distribution with high values for scale and shape parameters (assumed to be the same and equal to be 18, 14, 13, 12, 10 for patient severity types 5, 4, 3, 2, and 1 respectively). Such a distribution ensured that the probability of patient abandonment remained nearly zero unless a patient waited for a long time. For instance, the probability of patients of severity type 5 abandoning becomes greater than 10% only after waiting 8 hours. This probability peaks to the maximum possible of around 35% at 10 hours before going back to be under 10% after 11.5 hours. We note here that patient wait times were never that high 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 (2006c) for details.
4.2. **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 is 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.

**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.

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

4.3. **Ward Workload Function**

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

\[
\gamma^j(s^j, n^j) = \sum_i u^j_i \frac{n^j_i}{s^j_i},
\]

where \( u^j_i \) 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 \( u^j_i \) is higher for patients of higher severity (say, severity 1 compared to severity 3) within the same ward. Similarly,
the value of $u_j$ 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 $u_j$ across wards and only assumed difference 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 values used for $u_i$ during the experiments in Section 5 are 10, 8, 7, 3, 2 for $i = 1, 2, 3, 4, 5$ respectively, indicating that more severe patients lead to a higher level of workload for the nurses.

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.

5. Numerical Analysis

Before conducting numerical analyses, we first calibrate patient arrival rates within the simulation and fluid model 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 leads to the actual arrival rates varying over time. For instance, the Fast Track ward at the hospital is only open from 7am to 10pm 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 structure characteristic and therefore calibrating patient arrival rates helps in better estimating actual arrival rates. Fig. 5 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 matches the simulated LOS.

![Figure 5](image.png)

**Figure 5.** A comparison of patient LOS using the calibrated arrival rates within the simulation vs actual patient LOS obtained from data.

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 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.

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.

5.1. Numerical Analyses & Optimal Strategies

In this section we attempt to optimize the status quo routing and staffing values (shown in Fig. 1) 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. We note here that we obtained the existing system ward workload values using the function from Eq. (4) applied to the existing system’s nurse staffing values. Under each experimental scenario, we follow the solution procedure from Section 3.5 to obtain optimal values for patient LOS and ward workload. 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, as shown in Table 2.

<table>
<thead>
<tr>
<th>Severity 1</th>
<th>Severity 2</th>
<th>Severity 3</th>
<th>Severity 4</th>
<th>Severity 5</th>
<th>Weighted Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wait times (min)</td>
<td>215</td>
<td>288</td>
<td>203</td>
<td>112</td>
<td>131</td>
</tr>
<tr>
<td>% of total patients</td>
<td>0.8%</td>
<td>24.7%</td>
<td>49.8%</td>
<td>23.4%</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

Table 2. Status quo patient wait time obtained using existing data from the hospital.

<table>
<thead>
<tr>
<th>Workload</th>
<th>Critical Care Ward</th>
<th>Minor Care Ward</th>
<th>Fast Track Ward</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14.5</td>
<td>19.1</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Table 3. Status quo ward workload values obtained using equation (4) on existing data from the hospital.

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 aren’t 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 ($\gamma_j^*$). 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

$$
\gamma^j(s^j, x^j) \leq \gamma_j^*, \ \forall j = 1, 2, 3.
$$

While performing the experiment, the value of $\gamma_j^*$ is kept the same for all three wards and is equal to 15. We provide intuition behind this number by stating...
that under the workload function defined by us in Section 4.3.3, the workload of a ward having one patient of each severity type being cared for by two nurses equals 15. Additionally, we note from Table 3 that one of the three wards experiences a workload value of over 15 and our goal is to manage routing and staffing such that all wards experience workload values under 15.

- **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

\[
|\gamma^j(s^j, x^j) - \gamma^k(s^k, x^k)| \leq \hat{\gamma}, \ \forall (j, k) \in \{(1, 2), (2, 3), (1, 3)\}
\]

While performing the experiment, we chose a value of \(\hat{\gamma} = 2.5\). 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 \(\hat{\gamma}\) could be chosen for each pair of wards being considered for workload balance.

Under each experimental setup, we compare patient LOS (section 4.2.1) and ward workloads (section 4.2.2) across the experiments. Finally, since we assume a fixed values for the total number of nurses in our model, we discuss the possible improvements in system performance on increasing the total number of nurses available for staffing (section 4.2.3).

### 5.1.1. Comparing Patient LOS

On comparing patient LOS across experiments, as shown in Table 4, we note that the weighted average of the optimized models decrease by between 3.7% - 5.6% compared to status quo under all experimental scenarios. However, we also note that the range in reduction to weighted average LOS across the experiments is small. We see that the overall decrease is achieved by decreasing the LOS of severity type 1 and type 2 patients at the expense of increasing the LOS among severity type 4 and 5 patients. This is reasonable as type 4 and 5 patients are the least severe patients and already have the shortest LOS among severity types.

<table>
<thead>
<tr>
<th>Status quo</th>
<th>Severity 1</th>
<th>Severity 2</th>
<th>Severity 3</th>
<th>Severity 4</th>
<th>Severity 5</th>
<th>Weighted Average</th>
<th>Max Ward Workload</th>
<th>Ward Workload Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>215</td>
<td>288</td>
<td>203</td>
<td>112</td>
<td>131</td>
<td>200.81</td>
<td>19.10</td>
<td>9.1</td>
<td></td>
</tr>
<tr>
<td>No workload constraints</td>
<td>196 (8.8%)</td>
<td>288 (0%)</td>
<td>184 (9.4%)</td>
<td>118 (-5.4%)</td>
<td>142 (-4.4%)</td>
<td>191.48 (3.7%)</td>
<td>17.27 (9.6%)</td>
<td>4.27 (33.1%)</td>
</tr>
<tr>
<td>Workload threshold constraint: 15</td>
<td>197 (8.4%)</td>
<td>287 (9.5%)</td>
<td>179 (11.8%)</td>
<td>115 (-2.7%)</td>
<td>115 (12.2%)</td>
<td>189.48 (5.6%)</td>
<td>16 (16.2%)</td>
<td>3.7 (99.3%)</td>
</tr>
<tr>
<td>Workload balance constraint: 2.5</td>
<td>209 (2.8%)</td>
<td>283 (1.7%)</td>
<td>183 (9.9%)</td>
<td>118 (-5.4%)</td>
<td>141 (-7.6%)</td>
<td>191.81 (4.5%)</td>
<td>16.1 (15.7%)</td>
<td>2.5 (72.5%)</td>
</tr>
</tbody>
</table>

Table 4. Results for patient LOS and workload. Status quo shown in grey and simulated results for each experiment shown below. The percent different compared to status quo are shown in parenthesis.
5.1.2. Comparing Ward Workload

We note from Table 5 that on optimizing the model without any constraints, the workloads in the Critical Care and Minor Care ward decrease (compared to the status quo) while the workload in the Fast Track ward increases. The model thus works on achieving a certain level of equity in workload distribution across wards. On implementing a threshold constraint, we note that all workload values continue to decrease; however, we are unable to reach our target of 15 in the Minor Care ward. This may be attributed to the system operating under heavy levels of utilization, thus making it difficult to reduce the workload levels even further. We note here that we don’t see an infeasible solution in such a situation despite setting a highly restrictive threshold constraint. 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.

<table>
<thead>
<tr>
<th></th>
<th>Critical Care</th>
<th>Minor Care</th>
<th>Fast Track</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Status Quo</strong></td>
<td>14.50</td>
<td>19.10</td>
<td>10.00</td>
</tr>
<tr>
<td>No Workload Constraints</td>
<td>13.00</td>
<td>17.27</td>
<td>14.60</td>
</tr>
<tr>
<td>Workload Threshold Constraint: 15</td>
<td>13.50</td>
<td>16.00</td>
<td>12.30</td>
</tr>
<tr>
<td>Workload Balance Constraint: 2.5</td>
<td>13.60</td>
<td>16.10</td>
<td>14.50</td>
</tr>
</tbody>
</table>

Table 5. Results for Ward Workload.

5.1.3. Effects of 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 and 4.2.2 that attempting to optimize for patient LOS and ward workload leads to small improvements. This indicates to us that that the current system is already operating at capacity and to see significant improvements in performance, we may require additional nurses/resources. As is evident from Fig. 6, both weighted average LOS and average ward workload may be reduced by 12.5% and 30% respectively on adding 5
more nurses. This indicates to us 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 if adding staff members is economically viable and that our analysis is focused purely on operational improvements in terms of patient LOS and nurse workload.

5.2. Discussion of Results

The key takeaways from our numerical analysis 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, our model attempts to achieve better balance between ward workloads compared to the status quo even without setting predefined threshold/balance constraints. Third, 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 does 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 patient LOS is concerned but has room for improvement as far as nurse workload is concerned. 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.

6. Conclusions

We developed in this paper a framework to improve patient LOS and nurse workload by adjusting staffing and routing policies for a hospital ED modeled as a multi-class many-server pooled queueing system. 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 regional hospital in North Carolina to conduct case studies 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 about 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 pool’s shift started may impact their
workload. This potential temporal variation in nurse Future research could involve de-
veloping 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 paper. 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 paper 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 analy-
ses within this paper, we considered five patient classes to match the five levels of the
Emergency Severity Index (ESI) triage algorithm adopted by Southeastern Regional
Medical Center to classify incoming 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 sys-

tem functions. However, closer inspection of patient time in service for each of these
combinations indicates that a higher level granularity may be possible. Consider the
scatter-plot in Fig. 2 showing the time in service for patients of severity type 3 receiv-
ing care in the minor care ward. Separating these 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 re-
quire 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 paper can have applications
well beyond the field of healthcare and can benefit any service system that involves
customer arrivals into one of several different server pools, such as in wireless net-
works (Qadir et al., 2016) where resource pooling involves abstracting a collection of
networked resources to behave like a single unified resource pool.

References


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 Confer-

AHRQ (2018). Section 1. the need to address emergency department crowding.

nurses’ workload in medical and surgical wards of selected hospitals in kuwait. *Journal of


