Utilizing Partial Flexibility to Improve Emergency Department Flow: Theory and Implementation

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School, cwchan@columbia.edu

Vahid Sarhangian
Department of Mechanical and Industrial Engineering, University of Toronto, sarhangian@mie.utoronto.ca

Prem Talwai
Operations Research Center, Massachusetts Institute of Technology, talwai@mit.edu

Kriti Gogia
Department of Emergency Medicine, Weill Cornell Medicine, krg2015@med.cornell.edu

Emergency Departments (EDs) typically have multiple areas where patients of different acuity levels receive treatments. In practice, different areas often operate with fixed nurse staffing levels. When there are substantial imbalances in congestion among different areas, it could be beneficial to deviate from the original assignment and reassign nurses. However, reassignments typically are only feasible at the beginning of 8-12-hour shifts, providing *partial flexibility* in adjusting staffing levels. In this work, we propose a stochastic queueing network model of patient flow in the ED and study an associated fluid control problem to guide the reassignment decision for two types of nursing staff. We propose a heuristic solution approach and investigate its performance both analytically and using simulation. Analytical results and simulation experiments suggest a significant reduction of waiting times in parameter regimes relevant to the ED setting. We further implement the staffing approach at a large ED. This pilot study highlights several challenges of implementing operational interventions in the ED, including the difficulty of establishing a clean statistical environment in such setting. Despite these challenges, we find that guiding reassignment decisions using our approach is associated with significant improvements to patient flow including a reduction in average total ED length-of-stay (LOS) of 1.7 hours.

Key words: Healthcare, staffing, emergency department, queueing, partial flexibility, empirical evaluation

1. Introduction

Emergency Departments (EDs) provide care for patients for whom timely access to care is often paramount. Unfortunately, waiting times in EDs are often quite long and this can have substantial impact on quality of care, patient satisfaction, and employee morale. In this work, we propose an approach to manage nurse staffing resources with the goal of mitigating time spent in the ED. In addition, we report the results of a pilot implementation of the approach at the New York Presbyterian (NYP) Weill Cornell Medical Center (WCMC) ED.

In North America, it is common for EDs to have several areas (also referred to as bays, pods, or zones). Patients are often routed to these areas based on the severity of their condition and/or

resource requirements, as measured e.g. by the Emergency Severity Index (ESI) or the Canadian Triage and Acuity Scale (CTAS). For instance, low-acuity patients (ESI/CTAS 4 and 5) are typically treated in a separate area referred to as the "Fast Track". In practice, different areas often operate with fixed staffing levels. Nurse schedules, which determine the shifts and the areas the nurses will work, are typically specified weeks in advance so that the nurses are able to plan their lives appropriately. Despite the attractive simplicity of such a *dedicated* staffing approach, in instances when there are substantial imbalances in demand among different areas (e.g., due to regular stochastic fluctuations or external "shocks" caused by pandemics or flu outbreaks), and/or there are substantial imbalances in supply among different areas (e.g. due to nurse absenteeism), it could be beneficial to deviate from the original assignment and reassign nurses among the areas. Despite potential operational benefits, mid-shift reassignments are seldom feasible due to the setup time nurses require to arrange their work area. More importantly, since nurses often provide care for multiple patients at a time, mid-shift reassignments could result in handing off the care of patients to other nursing staff and hence impact continuity of care. However, at the beginning of shifts, when new nurses arrive, there is an opportunity to reassign nurses to different areas. This paper is concerned with guiding this reassignment decision.

Our work builds on Chan et al. (2021) which examines how a fixed number of servers should be assigned to different classes of a queueing system at the beginning of discrete time-intervals (shifts). They consider a standard parallel queueing system (see, e.g., Dai and Tezcan 2011) with a fixed number of flexible servers and develop asymptotically optimal reassignment policies with the objective of minimizing the total expected holding cost of the system incurred over a finite horizon. Through analysis of an associated fluid control problem, they develop insights into the properties of "good" reassignment policies. Specifically, they demonstrate the importance of balancing efficiency – as measured by myopic cost reduction – and excessive idleness – due to the inability to reassign servers mid-shift. In addition, it is shown that when the parameters are such that the system does not reach steady-state early into the shift, this *partial flexibility* with respect to reassignment of servers can lead to significant cost reduction compared to dedicated staffing.

Nevertheless, there are gaps in the model of Chan et al. (2021) that do not allow for direct implementation in practice. In particular, the model there is a single-stage queueing system with a single server type, whereas the provision of care in ED involves a secondary stage for patients admitted to the hospital, which may utilize a different nursing type. The WCMC ED utilizes two types of nursing staff: Emergency Department nurses (ED) – who can care for all patients in the ED – and Emergency Department Inpatient Nurses (EDIN) – who can only care for *boarding* patients, i.e., patients who have been treated in the ED and are waiting to be admitted to the hospital and do not have the credentials to provide care for *emergent* patients. When the initial treatment of a

patient concludes and a decision is made to admit the patient to the hospital, the care is handed over from an ED nurse to an EDIN nurse, unless all EDIN nurses are busy. In the latter case, the ED nurse continues to be the caregiver until either the patient is admitted to an inpatient unit or an EDIN nurse becomes available. Since the EDIN staffing levels affect the number of ED nurses available to serve patients in their "treatment" phase, the reassignment decisions for the two nursing types are inter-dependent and cannot be made independently, rendering the staffing solution developed in Chan et al. (2021) not directly applicable. We note that utilizing inpatient nurses for boarding patients in the ED is a common strategy and not exclusive to WCMC (e.g., Baker and Esbenshade 2015).

To address these gaps, we model each area of the ED as a two-stage tandem queueing network, with the two stages corresponding to the treatment and boarding phases of care, respectively. The model incorporates a number of important features of patient flow including time-varying arrivals. The problem is to determine the optimal assignment of available ED and EDIN nurses to the two phases of care at the beginning of each shift, in order to minimize the total expected average queue-length over a finite horizon. We present a heuristic solution by analyzing an associated fluid control problem and evaluate its performance both analytically and using simulation experiments.

The proposed approach provides an assignment that can in principle be implemented in practice. Nevertheless, our model does not capture all the complexities of a real ED. As such, it is not guaranteed that its recommendations will always be followed in practice. In addition, the actual impact of the intervention may differ from what is estimated using simulation experiments. Therefore, we conduct a pilot study of our proposed staffing approach in the WCMC ED. Our objectives are two-fold: (1) to understand the potential translational challenges and nuances of implementing our (and more broadly any) dynamic staffing algorithm in practice; and (2) to empirically measure the impact of such an implementation on patient flow metrics in a real ED. For the latter, one would ideally use an experimental design and randomly assign patients to periods with and without our staffing approach implemented. Given the infeasibility of this approach in the ED setting, we conduct a pre/post comparison of the outcome metrics using reduced form models and supplement our analysis with extensive sensitivity analysis and robustness checks. In doing so, we also aim to gain insights on the adequacy and challenges of using this approach in measuring the performance of complex operational interventions.

Our main contributions and results can be summarized as follows.

A queueing control problem to guide nursing reassignment decisions at the ED: Our queueing network model explicitly accounts for the two different phases of care (treatment and boarding) and the two types of nursing staff, and captures the differential impact of ED and EDIN nurses on

the patient flow. This leads to a new discrete-time stochastic control problem for determining the assignment of available nursing staff (for each phase of care) to the different areas of the ED.

Solution approach and performance evaluation: We propose an associated fluid control problem and present a heuristic algorithm to approximately solve it. The heuristic relies on "decoupling" the assignment decisions for the two types of nursing staff. By analyzing the dynamics of the fluid model during a shift, we identify parameter regimes where the proposed approach is expected to perform well. We further use simulation experiments to evaluate the performance of a translation of our approximate solution for the original stochastic queueing system. The results suggest that in relevant parameter regimes the proposed reassignment policy can lead to significant reduction of average queue length compared to the best dedicated staffing.

Pilot implementation at WCMC and empirical study: we conduct a pilot implementation of the proposed staffing approach and collect pre/post implementation data. Using a linear regression model, we find that our intervention is associated with a significant improvement to the patient flow, including a 1.7 hours reduction in the total LOS. These benefits are observed despite real-world challenges that arise when implementing such an operational intervention, such as an increase in the arrivals during the ON period and partial compliance with our staffing recommendations. Robustness checks and sensitivity analysis suggest that our findings are to a large degree robust to model specifications and potential biases introduced by unobserved confounders. The pilot study also highlights several challenges of implementing and empirically evaluating the proposed approach in practice. These findings point to important future research directions and provide useful insights for implementing other operational interventions in the ED.

The rest of the paper is organized as follows. After discussing the related literature in Section 2, we provide background information for the ED setting in Section 3. In Section 4, we present the queueing network model and its fluid approximation under dedicated staffing. In Section 5, we present and analyze the associated fluid control problem under partial flexibility and the proposed solution approach. Details of the pilot implementation are presented in Section 6 and the pre/post comparison results are discussed in Section 7. Finally, we conclude in Section 8 with a discussion of the limitations of our study and future research directions.

2. Related Literature

There have been substantial efforts within the medical community to understand and quantify the impact of waiting time and LOS in the ED. For instance, increases in waiting have been shown to be associated with increased mortality rate (Plunkett et al. 2011); increases in total ED LOS are associated with an increase risk of adverse events (Ackroyd-Stolarz et al. 2011); and, longer boarding times have been shown to be associated with increased mortality and hospital LOS (Singer et al. 2011).

Recognizing the need to improve patient flow in EDs, the Operations Research / Management community has proposed a number of different strategies to manage patient flow in the ED. In particular, queueing models of patient flow have been used to guide various control and design decisions in the ED. Yildiz et al. (2019) use a queueing network model and its steady-state fluid approximation to investigate the parameter regimes where a provider triage (i.e., an initial assessment performed by a physician) can be economically and operationally beneficial. Zayas-Cabán et al. (2016) use tandem queues with abandonment to study a Triage-Treat-and-Release program where nurses or physician assistants are responsible for both triage and treatment of low-acuity patients. Helm et al. (2011), Saghafian et al. (2012) and Saghafian et al. (2014) use dynamic programming to develop patient prioritization strategies to improve flow in the ED. Another stream of work, e.g., Dobson et al. (2013), Huang et al. (2015) and He et al. (2019), focuses on developing strategies to balance the tension between treating new patients (i.e., triaged with limited information about their health state) and "in-process" patients (i.e., have been assigned a bed and are awaiting test results and/or additional physician checks).

More closely related to our work are studies that focus on managing ED providers: nurses and physicians. A large body of literature focuses on shift scheduling and rostering, often assuming deterministic demand and ignoring queueing dynamics. Other papers focus on determining appropriate provider staffing levels. Yom-Tov and Mandelbaum (2014) uses queueing models to determine staffing levels to stabilize performance metrics given time-varying arrival rates and customers may have multiple entries into service. Green et al. (2013) develop a Newsvendor-like model to determine ED nurse staffing levels, which account for the endogenous absenteeism behavior of nurses who are more likely to not show up for a shift if it is understaffed. Yankovic and Green (2011) use a queueing model to estimate the impact of nurse-to-patient ratios and nursing capacity on ED overcrowding. In contrast, in this work we assume that staffing levels and schedules are already determined, and focus on dynamic assignment of the providers to different areas of the ED. Dynamic assignment of providers within the ED has also been considered in Cohen et al. (2014), but the focus there is on continuous allocation of surgeons to the shock rooms and operating rooms in the ED during a mass casualty incident.

There is a vast literature on the value of flexibility in operations management (see, e.g., Jordan and Graves 1995 and Tsitsiklis and Xu 2017). In particular, many studies have investigated the benefits of flexibility in server capabilities in parallel queueing systems. Related examples include Gurumurthi and Benjaafar (2004), Bassamboo et al. (2006), Iravani et al. (2005, 2011) and Tsitsiklis and Xu (2013). In this work, we assume that nurses (servers) are capable of working in all areas (server pools). However, their assignments to different areas can only change at the beginning of 8-12-hour shifts, providing partial flexibility with respect to adjusting the staffing levels for different

areas. This also relates our work to the literature on flexible staffing. In this literature, flexibility is often with respect to delaying the staffing decision until more information regarding the workload is available. This is common in, for example, the retail and restaurant industries, but has been also considered in healthcare. For instance, He et al. (2012) utilizes a Newsvendor framework to study the value of delaying staffing decisions for operating room nurses until case information is revealed. The partial flexibility considered in this work is different, in that the flexibility is with respect to the area where nurses work, without the option of adjusting the total number of nurses available.

Despite the growing literature on queueing models of patient flow in the ED, little work has been done to empirically demonstrate and quantify the benefits of prescriptions obtained from the models in practice. There are a few examples in the literature that do so, mainly focusing on improving provider schedules, e.g., Green et al. (2006) and Sir et al. (2017). Our work appears to be the first to attempt to empirically measure the benefits of a dynamic operational policy in the ED setting, and highlight the translational challenges.

3. Setting

In order to provide context of our model and the subsequent pilot study, we start by describing the operations at the WCMC ED. NYP Weill Cornell is an academic medical center and its ED treats more than 70,000 patients annually at the WCMC ED.

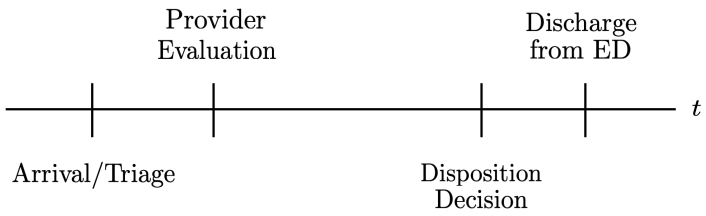


Figure 1 **A typical timeline for patients in the ED.**

A typical timeline for patients is depicted in Figure 1. Patients arrive to the ED and are seen by a triage nurse. After some (possibly 0) wait, the patient is seen by a provider. The provider works to stabilize the patient and may order some tests and/or labs in order to determine a disposition recommendation. Once the disposition decision is made the patient is prepared for transfer into the hospital or is discharged home. If a patient is admitted into the hospital, the time from disposition decision to ED discharge is referred to as the boarding time. A patient's total LOS is measured from arrival to discharge from the ED and includes various 'waits' throughout the process. Our model and analysis will primarily focus on the waiting time from triage to provider evaluation. However, in our pilot evaluation, we also consider other measures of patient flow, including boarding time and total LOS.

Table 1 Patient Areas at WCMC at 7am and 7pm shift

	Area A (24 hours)	Area B (24 hours)	Area C (24 hours)	Area U (7am-12am)
Percentage of patients	26.6%	27.4%	25.7%	20.2%
Average Number of ED nurses per shift	3.867 (0.623)	2.983 (0.624)	3.067 (0.607)	0.617 (0.691)
Average Number of EDIN nurses per shift	0.867 (0.650)	1.633 (0.551)	1.750 (0.508)	0.000 (0.000)
Average ESI	2.793 (0.541)	2.968 (0.431)	2.887 (0.476)	3.865 (0.577)

Note. Mean and standard deviation (in parentheses).

Patients are treated in 4 areas (A, B, C, and U) that are physically separated from each other by hallways and/or doors. In general, patients are routed to the different areas based upon assessed severity at triage, i.e., their ESI levels. Areas A, B, and C are staffed 24 hours a day, 7 days a week. Patients with ESI 1, 2, or 3 are treated in one of these three areas. Area A predominantly treats trauma patients. Area U, sometimes referred to as the "FastTrack", is staffed from 7am to 12am, 7 days a week and is primarily used to treat patients with ESI 4 or 5. As is typical with most EDs, patient arrivals exhibit substantial temporal variation during the day. See Appendix A.4 for an illustration of the empirical arrival rates.

Nursing schedules for all four areas are set weeks in advanced and are fixed across days and shifts. One can think of this staffing approach as "dedicated staffing" to each area. Still, if there is nurse absenteeism and/or large imbalances in patient demand, the charge nurse may reassign nurses to different areas based on his/her intuition. The majority of nurses are scheduled to 12-hour shifts that begin at 7am or 7pm. However, there are 2 to 3 nurses that begin their shift at 5pm. Our analysis focuses on the nurses that begin at 7am or 7pm. Table 1 summarizes the differences across the four areas with respect to staffing levels, patient load, and ESI levels. The values are averages over 4 weeks.

4. The Queueing Model and Its Fluid Limit under Dedicated Staffing

To capture salient features of an ED to develop insights into effective nurse staffing, we propose a stylized queueing system. In particular, we consider a queueing network with I tandem queues in parallel, each assigned to a corresponding patient class and representing an area in the ED. Under dedicated staffing, different classes operate independently. Hence, we begin by describing the model for an arbitrary patient class $i \in \mathcal{I} = \{1, 2, \ldots, I\}$ under dedicated staffing. Figure 2 provides a schematic representation of the model.

Patients in class i arrive to the ED according to a non-stationary Poisson process with a piece-wise continuous, periodic rate function $\{\lambda_i(t); t \geq 0\}$, which is commonly used to model patient arrivals to the ED (see, e.g., Kim and Whitt 2014). Without loss of generality we assume that the period is 24 hours; that is, $\lambda_i(t+24) = \lambda_i(t)$ for $t \geq 0$. We denote the average arrival rate of class i over a period by Λ_i, i.e.,

$$\Lambda_i = \frac{1}{24} \int_0^{24} \lambda_i(s) ds. \tag{1}$$

The service time of each patient in class i can be partitioned into two *phases*. The first phase corresponds to the stabilization and disposition assessment of the patient and is exponentially distributed with rate μ_i. After completing Phase 1, a class i patient will require Phase 2 service with probability $p_i \in [0,1]$. Thus, p_i is the probability that a class i patient is admitted to the hospital from the ED. The 'service' (boarding) time of class i customers in Phase 2 is exponentially distributed with rate ν_i. A patient who completes Phase 1 of service and remains in the system (i.e. must be admitted to the hospital) must immediately begin Phase 2 of service. That is, while new patients may wait in an infinite buffer queue to begin Phase 1 of service, there is no queue for Phase 2 of service.

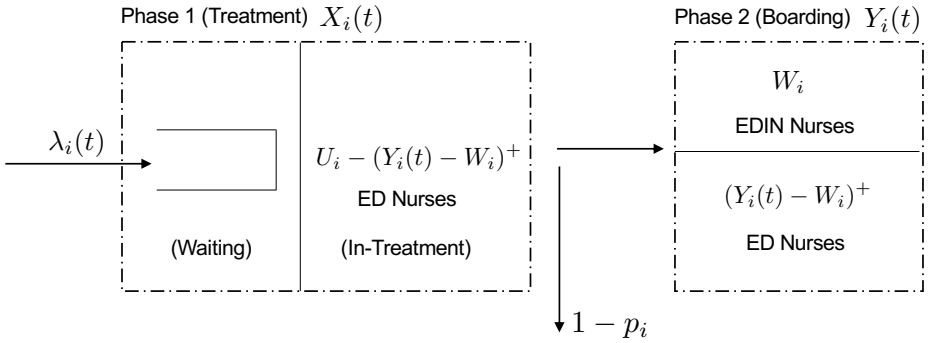


Figure 2 **A schematic representation of class i of the queueing network and Phase 1 and 2 of service. The number of ED nurses available in Phase 1 depends on the number of patients in boarding and the number of available EDIN staff.**

There are two types of servers: 1) ED nurses who can treat patients in either phase of service and 2) EDIN nurses who can only treat boarding (Phase 2) patients. We use the term 'server' to capture the nurse capacity required to treat a single patient. In many EDs, nurses can handle up to 7 patients at a time, so one can think of a single server as 1/7 of a nurse's capacity. While both sets of servers are co-located and treat patients who are in the area for class i patients, they are distinguished by their capabilities. To highlight this differentiation, we refer to the pool of W_i servers allocated to provide care for boarding patients as Station 2, and to the pool of U_i servers that provide care for emergent patients in Phase 1 as Station 1. Servers at Station 1 are flexibly trained, so they can serve both phases of a patient's service. In particular, if a patient who just completed Phase 1 service requires Phase 2 service, but there are no available servers in Station 2, the same Station 1 server provides the service for the patient in Phase 2. The service is "handed-over" to a server in Station 2 according to a preemptive-resume policy as soon as it becomes available. At this point, the "borrowed" Station 1 server becomes available to serve Phase 1 patients again.

Denote by $X(t) = (X_1(t), \ldots, X_I(t))$ the stochastic process that keeps track of the number of patients in Phase 1 with $X_i(t)$ denoting the number of class i patients in Phase 1 (in service or waiting in queue) at time t. Denote by $Q(t) = (Q_1(t), \ldots, Q_I(t))$ the process that keeps track of the number of customers in the queue to begin Phase 1 service in Station 1, with $Q_i(t)$ denoting the number of class i patients in queue. Further, let $Y(t) = (Y_1(t), \ldots, Y_I(t))$ denote the process that keeps track of the number of patients in Phase 2 with $Y_i(t)$ denoting the number of class i customers in Phase 2 at time t. Define $3 \times I$ independent unit-rate Poisson processes denoted by $\{N_{ij}(t); t \geq 0\}$ for $i \in \mathcal{I}$ and $j \in \{1, 2, 3\}$. Assume that at time $t = 0$, $X(0)$ and $Y(0)$ satisfy $X_i(0) \geq 0$ and $0 \leq Y_i(0) \leq U_i + W_i$ for all $i \in \mathcal{I}$. Then, the sample paths of the processes $X(t)$, $Y(t)$ and $Q(t)$ satisfy the following for all $i \in \mathcal{I}$ and $t \geq 0$:

$$X_i(t) = X_i(0) + N_{i1}\left(\int_0^t \lambda_i(u)du\right) - N_{i2}\left(\mu_i \int_0^t \left(X_i(u) \wedge \left(U_i - (Y_i(u) - W_i)^+\right)\right) du\right), \tag{2}$$

$$Y_i(t) = Y_i(0) + N_{i2}\left(p_i\mu_i \int_0^t \left(X_i(u) \wedge \left(U_i - (Y_i(u) - W_i)^+\right)\right) du\right) - N_{i3}\left(\nu_i \int_0^t Y_i(u)du\right), \tag{3}$$

$$Q_i(t) = \left(X_i(t) - \left(U_i - (Y_i(t) - W_i)^+\right)\right)^+. \tag{4}$$

In (2)–(3) the number of customers in each phase are obtained by subtracting the total departures from the total arrivals up to any time $t \geq 0$ (see, e.g., Mandelbaum et al. 1998 and Pang et al. 2007 for details). Note that at any time $t \geq 0$, the departure rate from Phase 1 service is determined by the number of active servers available in Station 1, i.e., $(U_i - (Y_i(t) - W_i)^+)$, with $(Y_i(t) - W_i)^+ \leq U_i$ ED servers borrowed to treat Phase 2 patients. This highlights the inter-dependency between the staffing decisions of the two stations: if Station 2 is understaffed, as patients complete service in Station 1 and move to Phase 2, Station 1 capacity decreases during the shift, leading to higher congestion in the queue.

We now formally define the primary metric with respect to which we assess the quality of various staffing allocations. We measure the performance of the stochastic system based on the expected time-average queue length over a finite horizon of length T, that is,

$$\mathbb{E}\left[\frac{1}{T}\int_0^T Q(t) \cdot \mathrm{e}\ dt\right], \tag{5}$$

where e denotes an I-dimensional vector with all values equal to 1. By Little's Law, minimizing the above metric is equivalent to minimizing the total expected waiting time. Moreover, as the horizon length goes to infinity, the metric converges to the long-run average total queue length (assuming that it exists). Note that if there are substantial differences in wait-tolerance across the different classes, this could be incorporated by introducing a weight vector and optimizing over the expected waiting cost. Alternatively, one could introduce staffing constraints for different classes to ensure minimum service levels are met.

Next we describe a deterministic fluid approximation for the transient dynamics of the system under dedicated staffing. In the fluid approximation, the discrete processes are replaced by continuous deterministic flows and the capacities are also assumed to be continuous. Denote by vectors $x(t) = (x_1(t), \ldots, x_I(t))$ and $y(t) = (y_1(t), \ldots, y_I(t))$, the amount of fluid in Phase 1 (waiting or in service) and Phase 2 at time $t \geq 0$, respectively, and let $q(t) = (q_1(t), \ldots, q_I(t))$ denote the amount of fluid waiting in queues. Further, denote the vectors of Station 1 and 2 capacities by $u = (u_1, \ldots, u_I)$ and $w = (w_1, \ldots, w_I)$, respectively. The fluid trajectories satisfy the following system of ordinary differential equations (ODEs) for all $i \in \mathcal{I}$ and $t \geq 0$, analogous to the sample path equations,

$$\dot{x_i}(t) = \lambda_i(t) - \mu_i \left(x_i(t) \wedge \left(u_i - (y_i(t) - w_i)^+ \right) \right), \tag{6}$$

$$\dot{y_i}(t) = p_i \mu_i \left(x_i(t) \wedge \left(u_i - (y_i(t) - w_i)^+ \right) \right) - \nu_i y_i(t), \tag{7}$$

$$q_i(t) = \left(x_i(t) - \left(u_i - (y_i(t) - w_i)^+ \right) \right)^+, \tag{8}$$

starting from any initial condition satisfying $x_i(0) \geq 0$ and $0 \leq y_i(0) \leq (u_i + w_i)$ for $\forall i \in \mathcal{I}$. The fluid approximation can formally be justified as a Functional Strong Law of Large Number (FSLLN) for the stochastic sample paths in (2)–(3) using Theorem 2.2 of Mandelbaum et al. (1998). To this end, one considers a sequence of systems indexed by η in which the number of servers, arrival rates and the initial conditions linearly increase, while the service rates at both phases of service remain unscaled. For a system with a large number of servers, the fluid equations approximate the transient dynamics of the sample paths well on compact subsets of $\mathbb{R}_+$, with an error of size $o(\eta)$.

Before closing this section, we make a remark regarding the assumptions of the model. The assumption of exponential service time distributions, although common in the literature, are less likely to hold in practice. In Section 5.4 we examine the performance of our reassignment policies for the case where service time are Log-Normally distributed and show that the policies can still be effective. Beyond the parametric choice of distribution, the Phase 2 service times depend on the discharge process from the inpatient department of the hospital and hence can be time-dependent. Nevertheless, we will provide evidence through our pilot study that the prescriptions obtained from our model could still useful for guiding the assignment decisions in practice. We provide a more detailed discussion of the model assumptions in Section 8.

5. The Associated Fluid Control Problem under Partial Flexibility

We now consider the case where nurses can be reassigned at the beginning of each shift. Under this *partially flexible* design, we consider a total of n_1 and n_2 servers that can be assigned to Stations 1 and 2 of any patient class. The assignment of the servers can be changed at the beginning of shifts of length τ. Once the servers are assigned to a patient class, the dynamics of the system

are identical to that explained above under dedicated staffing with U_i and W_i determined by the assignment policy in that shift for $i \in \mathcal{I}$.

Recall that we refer to the amount of nursing capacity required to treat a single patient as a single server. Thus, if a single nurse can treat multiple patients at once, the set of feasible allocations needs to appropriately account for such constraints, i.e., they will be scalar multiples of the patient-to-nurse ratio. For simplicity of exposition, we assume here that each nurse treats one patient at a time, with the understanding that the ability of nurses to treat multiple patients simultaneously can easily be incorporated into the control problem by appropriately defining constraints on the feasible set of allocations.

Ideally, we would like to find an "admissible" control policy that minimizes the metric in (5). Due to the multidimensionality of the problem, time-varying dynamics, and the complex interplay between Stations 1 and 2, solving the stochastic control problem is very challenging. As such, in the following, we first present an associated fluid control problem that is directly motivated by the fluid approximations under dedicated staffing. We then propose a solution approach and investigate its performance in the subsequent sections. For other recent examples of policies derived from fluid control problems see Chen et al. (2021), Hu et al. (2021) and Chan et al. (2022).

5.1. The Fluid Control Problem

The fluid model under dedicated staffing serves as the initial foundation for the proposed deterministic control problem under the partially flexible design. The problem is to allocate the n_1 and n_2 available Station 1 and 2 capacity to each of the I customer classes, at the beginning of shifts of length τ. Denote by $u(t) = (u_1(t), \ldots, u_I(t))$ and $w(t) = (w_1(t), \ldots, w_I(t))$ the vector of capacity allocations, where $u_i(t)$ and $w_i(t)$ are respectively the amount of capacity allocated to Station 1 and 2 of class $i \in \mathcal{I}$ at time $t \geq 0$. The objective is to determine $u(t)$ and $w(t)$ subject to the capacity constraints and the restriction to control at the beginning of shifts, in order to minimize the time-average queue length over a finite horizon of length T and starting from a given initial condition $x(0) = x^0$ and $y(0) = y^0$, satisfying $x^0 \geq 0$ and $0 \leq y^0 \leq (u + w)$. For any vector $a \in \mathbb{R}^I_+$ denote $\hat{a} = \text{diag}(a_1, \ldots, a_I)$, i.e., its corresponding diagonal matrix and recall that $\mathbf{e}$ denotes the I–dimensional unit vector. The fluid control problem is:

$$\min_{(u(\cdot),w(\cdot))} \frac{1}{T} \int_0^T q(s) \cdot \mathrm{e}\, ds \tag{9}$$

$$\text{s.t. } q(t) = \Big(x(t) - \Big(u(t) - (y(t) - w(t))^+\Big)\Big)^+, \quad \forall t \in [0,T], \tag{10}$$

$$\dot{x}(t) = \lambda(t) - \hat{\mu} \cdot \Big(x(t) \wedge \Big(u(t) - (y(t) - w(t))^+\Big)\Big), \quad \forall t \in [0,T], \tag{11}$$

$$\dot{y}(t) = \hat{p} \cdot \hat{\mu} \cdot \Big(x(t) \wedge \Big(u(t) - (y(t) - w(t))^+\Big)\Big) - \hat{\nu} \cdot y(t), \quad \forall t \in [0,T], \tag{12}$$

$$x(0) = x^0,\ y(0) = y^0, \tag{13}$$

$$u(t)\cdot \mathrm{e} \le n_1,\ w(t)\cdot \mathrm{e} \le n_2, \quad \forall t \in [0,T], \tag{14}$$

$$u(t) \ge 0,\ w(t) \ge 0 \quad \forall t \in [0,T], \tag{15}$$

Denote by $\{t_k; k \in \mathcal{K} \equiv \{0,\ldots,N-1\}\}$ the sequence of times corresponding to the start of the shifts with $t_{k+1} - t_k = \tau > 0$ for $\forall k$. For simplicity, we assume that $T = N\tau$. A control policy $(u(t), w(t))$ is said to be an admissible policy if in addition to (11)–(15) it satisfies,

$$u(t) = u(k\tau) \equiv u[k] \quad \text{and} \quad w(t) = w(k\tau) \equiv w[k], \tag{16}$$

for $t \in [t_k, t_{k+1})$ and $\forall k \in \mathcal{K}$. That is, we require the allocations to remain constant over shifts of length τ. This defines a discrete-time control problem where the goal is to find a sequence of admissible allocations $\{(u[k], w[k]); k \in \mathcal{K}\}$ that minimizes the cost function in (9).

In principle, the proposed fluid control problem can be solved numerically, and the solution, after rounding, can be used as an assignment for the stochastic system. Unfortunately, there are several challenges associated with this approach. The time-varying dynamics and the large number of control variables make finding an optimal solution computationally demanding. In addition, a naive translating the solution for the stochastic system may perform poorly as shown in Chan et al. (2021) for the simpler single-station system. Motivated by these challenges, we next propose an intuitive heuristic solution that can be implemented with minimal computational effort and has good performance in the original stochastic system in parameter regimes relevant to the nurse staffing application.

5.2. The Proposed Solution Approach

Our objective is to construct a server allocation algorithm that (1) performs well, i.e. achieves queue length reductions relative to the status-quo dedicated staffing, and (2) is simple to explain and implement. The second criteria is essential to help facilitate implementation in practice.

We propose a solution approach that utilizes a decoupling of the staffing decisions for Station 1 and 2. The idea is to first determine the required number of Station 2 servers (possibly borrowing some from Station 1) to manage the boarding phase of service for each class and allocate the available n_2 servers accordingly. The remaining Station 1 servers are then allocated to Phase 1 of service for the I customer classes. The decoupling reduces the complexity of the allocation problem and facilitates an intuitive understanding of the underlying allocation algorithm. However, because Station 1 servers allocated to Phase 2 of service may free up during the shift and can then be used to serve Phase 1 of service, it is not obvious that such an approach will perform well.

An important consideration of the proposed approach for allocating servers is the possibility of incurring idleness during the shift. In particular, greedy allocation of servers based on the current

congestion levels may lead to poor solutions due to the inability to reassign servers during the shift. If one is too aggressive in allocating servers to a heavily loaded area, the servers in that area may incur excessive idleness during the shift while other areas have many patients waiting in the queue. This is particularly important for allocation of ED nurses, since the service times are relatively shorter in Phase 1 compared to Phase 2, increasing the possibility of mid-shift idleness. For allocation of Station 2 capacity, this is less of an issue since Phase 2 service times (i.e., boarding times) are typically longer and in the case of our ED comparable to the length of the shift. In addition, as the number of available EDIN nurses is typically lower, larger imbalances are required to result in moving an EDIN nurse to another area. As such, our proposed approach ignores the possibility of idleness for Station 2 capacity, but carefully accounts for it when allocating the remaining ED nurses.

Consider the system at the beginning of shift $k \in \mathcal{K}$ with $x[k]$ and $y[k]$ denoting the current value of the fluid trajectories for Phase 1 and 2 of service. In the following, we describe an algorithm to obtain an admissible pair $(u[k], w[k])$ for the shift.

1. *Assign Servers to Station 2.* For each class, $i \in \mathcal{I}$, compute $l_i = y_i[k] + p_i(x_i[k] + \Lambda_i)$, i.e., the current amount of fluid in Phase 2 plus the average amount of Phase 1 fluid to enter the system during the shift that will eventually require Phase 2 of service. Allocate the available Station 2 servers proportional to the computed l_i's, i.e., let

$$w_i[k] = \frac{l_i}{\sum_{i\in\mathcal{I}} l_i} n_2. \tag{17}$$

2. *Assign Station 1 servers to unattended Phase 2 patients.* Since all patients in Phase 2 of service must be in service (there is no queue), Station 1 servers have to be "borrowed" to manage the remaining patients in Phase 2. Set,

$$n_1' = n_1 - \sum_{i\in\mathcal{I}} (y_i[k] - w_i[k])^+, \tag{18}$$

i.e., the number of remaining Station 1 servers after lending the required number to Station 2.

3. *Assign the remaining Station 1 servers to the I customer classes* according to the balancing heuristic of Chan et al. (2021). For each $i \in \mathcal{I}$, calculate the maximum capacity $\tilde{u}_i$ such that no idleness is incurred during the shift. This can be done using bi-section and numerical evaluation of the ODEs in (11)–(12). If $\sum_{i\in\mathcal{I}} \tilde{u}_i[k] < n_1$, set,

$$u_i[k] = \tilde{u}_i[k] + \frac{1}{I} \sum_{i\in\mathcal{I}} \tilde{u}_i[k], \quad \forall i \in \mathcal{I}, \tag{19}$$

i.e., allocate the remaining capacity equally among the I classes. Otherwise, set

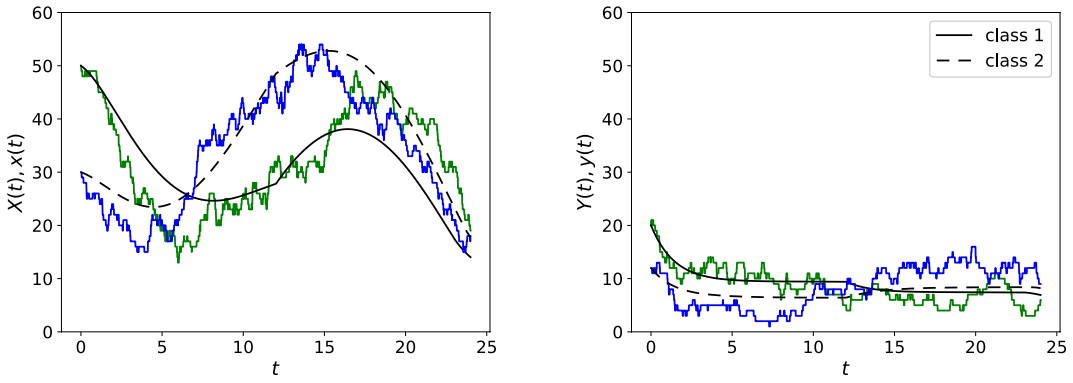
$$u_i[k] = \frac{\tilde{u}_i[k]}{\sum_{i\in\mathcal{I}} \tilde{u}_i[k]} n_1', \quad \forall i \in \mathcal{I}, \tag{20}$$

i.e., re-normalize the allocations so that they add up to n_1'.

We note that the proposed solution does not provide universally good solutions for the fluid control problem of Section 5.1. Rather, it provides "good" solutions in a regime that is relevant to the application considered in this work. We provide analytical and numerical evidence of this in the following sections. Here, we present an example to illustrate the sample paths and fluid trajectories of the system under the proposed heuristic; see Figure 3. We consider a two-class system with $n_1 = 40, n_2 = 16$ which correspond to 10 ED nurses and 4 EDIN nurses assuming a 1/4 nurse-to-patient ratio. The system is initialized with $X(0) = (50, 30)$ patients in Phase 1 and $Y(0) = (20, 12)$ patients in Phase 2. Service rates are $\mu = (0.5, 0.5)$ and $\nu = (0.5, 0.5)$, $p = (0.4, 0.4)$ and the shift length is $\tau = 12$. The arrival rates are given by the sinusoidal function,

$$\lambda_i(t) = \Lambda_i + \beta_i \sin\left(\frac{\pi(t+7)}{12}\right), \tag{21}$$

with $\Lambda_i = 8.6, \beta_i = -4.3$ for $i \in \{1, 2\}$. Observe that the fluid trajectories approximately capture the dynamics of the system under the heuristic policy.



(a) Phase 1 (b) Phase 2

Figure 3 **A sample path of $(X(t), Y(t))$ and the approximating fluid trajectories $(x(t), y(t))$ for $t \in [0, 24]$ under the heuristic policy for a two-class system with $n_1 = 40$, $n_2 = 16$, $\mu = (0.5, 0.5)$, $p = (0.4, 0.4)$, $\nu = (0.5, 0.5)$, $X(0) = (50, 30)$, $Y(0) = (20, 12)$, and shift length $\tau = 12$.**

5.3. The Performance of the Proposed Approach: Analytical Results

To provide analytical justification for our proposed solution approach, we focus on understanding how the Phase 2 dynamics influence staffing decisions in Station 1. Since all patients in Phase 2 of service must be assigned a server, inadequate Station 2 capacity during the shift would result in loss of capacity in Station 1. As such, it is not immediately clear whether the decoupling approach that we propose leads to good solutions.

We start by considering the case where the servers can be continuously allocated, i.e., without the constraints in (16). Not surprisingly, in this case, the decoupling approach can result in the optimal

server allocation (See Proposition 3 in Appendix A.1 for a formal statement.) More specifically, given the optimal allocation in Station 2, one can recover the optimal Station 1 allocation by optimally allocating the remaining Station 1 capacity to different classes. This, of course, requires knowing the optimal allocation vector in Station 2 which is, in general, unknown – but it motivates the approximation idea in our proposed solution method: if we have a "good" allocation for Station 2, we can solve the Station 1 problem by lending the required Station 1 servers to Station 2 and then allocating the remaining Station 1 servers.

In the presence of the shift constraint, we cannot obtain the optimal discrete-time policy by simply decoupling as described above. Even if we are given the optimal Phase 2 allocations $\{w^*[k] : k \in \mathcal{K}\}$, finding the optimal Station 1 allocation requires considering the evolution of Station 2 fluid over the entire shift, not just its value at the beginning of the shift. Therefore, in the following, we provide a characterization the accuracy of this decoupling strategy under the partially flexible design, i.e., under the constraints in (16).

To this end, we examine the dynamics of the system over a single shift and for a given allocation vector. Recall that in our solution approach we ignore the potential loss of capacity in Station 1 as servers move to Station 2. As such, we compare the actual Phase 1 dynamics of class i to an "isolated" system in which $y(t)$ is held fixed at its initial value $y[k] = y(k\tau)$ during the shift. In comparing the trajectories we seek to understand the extent to which staffing decisions for the two stations can be made "independently" at the beginning of each shift. An understanding of the deviation between these two systems provides insight into the parameter regimes where our proposed heuristic policy in 5.2 is expected to perform well.

For ease of exposition and without loss of generality, we consider the first shift starting at time $t = 0$, and denote the allocations for class i with $u_i \equiv u_i[1]$ and $w_i \equiv w_i[1]$ for $\forall i \in \mathcal{I}$. Since the dynamics are independent across different classes during the shift, we focus on an arbitrary customer class $i \in \mathcal{I}$. Denote the trajectory of the "isolated" system by $\tilde{x}_i(t)$, and note that it satisfies the following ODE for $\forall i \in \mathcal{I}$ and $t \in [0, \tau]$:

$$\dot{\tilde{x}}_i(t) = \lambda_i(t) - \mu_i \left(\tilde{x}_i(t) \wedge \left(u_i - (y_i^0 - w_i)^+ \right) \right), \tag{22}$$

starting at $\tilde{x}_i(0) = x_i(0) = x_i^0$. Note that the Phase 2 trajectory is fixed at its initial value y_i^0. Our first result provides conditions under which the Phase 1 dynamics of the isolated system coincides with that of the original system during a shift.

Proposition 1 *Suppose that at the beginning of the shift, there is enough capacity for the initial Phase 2 fluid, i.e., $y_i^0 < w_i$ and the Station 1 queue is non-empty, i.e., $x_i^0 > u_i$. If the Phase 2 fluid is initially decreasing, i.e., $\dot{y}_i(0) < 0$, then $x_i(t) = \tilde{x}_i(t)$, $\forall t \in [0, \tau]$.*

The result holds because a non-empty queue implies that the input rate to Phase 2 is at its maximum at the beginning of the shift, i.e., $p_i\mu_i u_i$. If enough capacity in allocated to Station 2 to process this maximal input, no Station 1 capacity will be borrowed during the shift. Hence, under these conditions, we can allocate Station 1 servers without consideration for Station 2 dynamics.

In contrast, if some of the Station 1 capacity is borrowed to cover for the insufficient capacity at Station 2, the trajectories of the two systems will deviate. Our next result provides an upper-bound on the deviation between $x_i(t)$ and $\tilde{x}_i(t)$ when the assumptions of Proposition 1 are not satisfied. Since the proposed solution approach allocates Station 1 capacity in a way that avoids incurring idleness (unless there is excess capacity available), we assume that the Station 1 queue remains nonempty during the shift in the isolated system. Let S denote the set of all allocations (u_i, w_i), such that $\tilde{x}_i(t)$ in (22) does not incur any idleness during the shift, i.e.,

$$S = \{(u_i, w_i) |\ \tilde{x}_i(t) > u_i - (y_i^0 - w_i)^+ \ \forall t \in [0, \tau]\}. \tag{23}$$

Proposition 2 *Suppose that the allocation for the shift satisfies $(u_i, w_i) \in S$ and that Station 2 capacity is insufficient to serve all initial class i patients, i.e., $y_i^0 > w_i$. Then for all $t \in [0, \tau]$, if $\dot{y}_i(0) \geq 0$,*

$$x_i(t) - \tilde{x}_i(t) = \frac{\mu_i \dot{y}_i(0)}{(p_i\mu_i + \nu_i)^2}(e^{-(p_i\mu_i+\nu_i)t} - 1 + (p_i\mu_i + \nu_i)t), \tag{24}$$

and otherwise,

$$0 \leq \tilde{x}_i(t) - x_i(t) \leq \frac{\mu_i}{\nu_i^2} \max\{|\dot{y}_i(0)|,\ |p_i\tilde{\lambda}_i - \nu_i y_i^0|\}(e^{-\nu_i t} - 1 + \nu_i t), \tag{25}$$

where $\tilde{\lambda}_i \equiv \min_{s\in[0,\tau]} \lambda_i(s)$ denotes the minimum arrival rate during the shift.

In the first case, when Phase 2 trajectory is initially non-decreasing, i.e., $\dot{y}(0) \geq 0$, $x_i(t)$ grows faster than $\tilde{x}_i(t)$ as "active" Station 1 capacity $(u_i - (y_i(t) - w_i)^+)$ decreases from its initial value. Since by assumption $\tilde{x}_i(t)$ remains above capacity during the shift, it follows that the input rate into Phase 2 remains independent of the arrival rate, allowing us to exactly obtain the difference between $\tilde{x}(t)$ and $x(t)$. When $\dot{y}_i(0) < 0$, the Phase 1 trajectory $x_i(t)$ could remain above capacity for the entire shift or fall below at some point during the shift. The upper-bound in (25) is obtained by considering the maximum deviation between the two cases. In either case, we show that both $y_i(t)$ and the total number of active servers at Station 1 deviate from their initial value like $O(1 + e^{-\nu_i t})$. Thus, since the discrepancy between $\tilde{x}_i(t)$ and $x_i(t)$ is a result of assuming a constant number of available Station 1 servers for $\tilde{x}_i(t)$, the error is in the order of $O(t + e^{-\nu_i t})$.

The characterization of error in Proposition 2 provides insights on parameter regimes where the proposed approach can be expected to perform well. In (24), it can be easily verified that

for sufficiently large t, the magnitude of the deviation grows with the boarding rate ν_i, while in (25) the error grows with ν_i for all $t \geq 0$. Moreover, in (24) the magnitude of the error increases with service rate μ_i, while in (25) the magnitude of the error increases with the service rate μ_i for sufficiently small p_i. Therefore, for a sufficiently long shift, we expect the error to be smaller for longer boarding times. Further, if the admission probability is sufficiently small, we also expect the performance of our approach to improve with increasing treatment times.

Since the above observations are based on an upper-bound (second case of Proposition 2) we discuss and illustrate their validity in Appendix A.2 by obtaining a tighter estimate of the error for the special case of a sinusoidal arrival rate function.

5.4. The Performance of the Proposed Approach: Numerical Experiments

In this section, we evaluate the performance of the proposed solution for the stochastic queueing network presented in Section 4 using simulation experiments. We compare the performance to that of the optimal dedicated staffing, which we obtain using simulation. We present the results for two sets of experiments. In the first set, we evaluate the performance of the solution approach for parameters calibrated to our implementation setting. In the second set, we vary the system parameters to illustrate the parameter regimes where the proposed solution does (and does not) perform well.

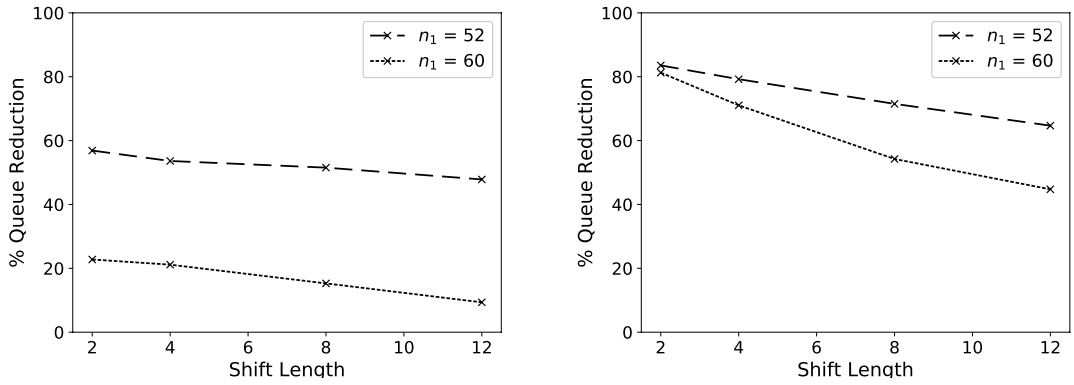
We use the proposed solution approach to obtain an allocation for each shift, given the observed state of the system at the beginning of the shift and the time the shift starts. We then convert this allocation to an integer assignment using the largest remainder method, i.e., we first apply the floor function to the allocations and then sequentially assign the remaining capacity in order of the largest difference between the original assignment and the floored value. In addition, we implement a non-preemptive version of the reassignment policy, in which nurses are required to complete their current service before being re-assigned to another class at the beginning of each shift.

We evaluate the performance of the proposed solution both over a finite horizon of length T and as $T \to \infty$. In both cases, we estimate the expected relative reduction in total queue length in comparison with the best dedicated staffing obtained using the following approach. For the transient experiments, we simulate all (feasible) dedicated staffing scenarios, and compare the estimated performance obtained using 100 replications. In the infinite horizon experiments, we first simulate 10 replications for each scenario to screen out the unstable ones, and keep the top 10 performing scenarios. We then run 40 additional replications for the 10 remaining scenarios and choose the one with the best estimate.

In all experiments, we consider a four-class system, i.e., $\mathcal{I} = \{1, 2, 3, 4\}$ and use the sinusoidal arrival rate in (21) with $\beta_i = (-\Lambda_i/2)$. We denote the Station 1 and 2 utilizations using ρ_1 and ρ_2

where $\rho_1 = (1/n_1)\sum_{i\in\mathcal{I}}(\Lambda_i/\mu_i)$ and $\rho_2 = 1/(n_2)\sum_{i\in\mathcal{I}}(p_i\Lambda_i/\nu_i)$. Further, we denote the offered class i load as $R_{1,i} = \Lambda_i/\mu_i$ for Station 1 and $R_{2,i} = p_i\Lambda_i/\mu_i$ for Station 2 and let $r_j = (r_{j,1}, \ldots, r_{j,I})$ with $r_{j,i} = R_{j,i}/\sum_{i\in\mathcal{I}} R_{j,i}$, where $\sum_{i\in\mathcal{I}} R_{j,i} = n_j\rho_j$ for $j \in \{1, 2\}$.

5.4.1. Calibrated Experiment We consider a system with parameters calibrated using data from the WCMC ED. We estimate the arrival rate parameters to be $\Lambda = (1.79, 1.75, 1.73, 2.34)$ and $\beta = (-0.67, -0.70, -0.70, -0.98)$. The estimated service rates are $\mu = (0.161, 0.158, 0.162, 0.238)$ in Phase 1 and $\nu = (0.063, 0.059, 0.057, 0)$ for Phase 2, and $p = (0.45, 0.36, 0.45, 0)$. The Phase 1 service times are estimated using "provider evaluation to decision" samples from the four areas.



(a) Finite Horizon (b) Infinite Horizon

Figure 4 Performance of the proposed reassignment policy in terms of expected reduction in total queue length compared to dedicated staffing for a system with parameters calibrated using data.

We also consider a nurse-to-patient ratio, which determines the number of patients that can be treated by a single server at each station. In the data, the number of available nurses and nurse-to-patient ratios vary across different shifts. In our experiments, we consider 13 ED nurses, and 8 EDIN nurses and use a ratio of 4 for Station 1 (ED) and 6 for Station 2 (EDIN), which roughly corresponds to the average capacity observed in our data. We then set the utilizations to $\rho_1 = 0.82$ and $\rho_2 = 0.77$ so that the system is heavily loaded but can still be stabilized with dedicated staffing. To observe the effect of system size we also consider a case with 15 ED nurses, and 9 EDIN nurses, and utilizations $\rho_1 = 0.75$ and $\rho_2 = 0.70$. For the transient experiments, we set the time horizon to $T = 720$ and initialize the system at a congested initial state of $X(0) = (1.25n_1, 0.75n_1, 1.25n_1, 0.75n_1)$ and $Y(0) = (0.5n_2, 0.5n_2, 0.5n_2, 0)$.

Figure 4 presents the results for the two system sizes and different shift-lengths. We observe a significant reduction of total queue length for both the transient and steady-state experiments. The reductions are largest for $\tau = 2$ (which approximates a continuous-time control) and decrease as the shift length increases. Still, we observe more than 40% reduction in the long-run average

total queue length for $\tau = 12$ hour shifts. In the transient experiments, the reductions are much larger for the smaller system. This can be explained noting that for the small system there are only 4 stable dedicated staffing options due to the nurse-to-patient ratio. This increases the value of adding some flexibility, although with a lower ratio, the possible reassignments are also further limited.

We also repeat the experiments assuming a Log-Normal distribution for the service times in the simulation and with parameters estimated from the data. We find that although the service times are more variable than exponential, the heuristic still achieves significant queue-length reductions (see Appendix A.3 for details). Overall, the results suggest that the proposed heuristic may have the ability to improve patient flow in an ED in practice.

5.4.2. Illustrative Experiments To further illustrate the performance of the proposed approach we conducted additional experiments. We summarize the setting and results here and relegate the details to Appendix A.3. In Experiment 1, we consider a similar setting to that of the calibrated experiments, but with shorter service and boarding times and higher admission probability. In doing so, we aim to understand the performance degradation in regimes where the decoupling approach is expected to have higher error as identified in the results of Section 5.3. We still observe a significant queue length reduction in both steady-state and transient experiments, although the gains are smaller compared to the calibrated experiment. In Experiment 2, we consider a case where our heuristic is expected to perform poorly. The utilizations are the same as in Experiment 1 but the load is unevenly distributed for both stations. Classes 3 and 4 face a smaller load in Phase 2 despite having a higher input rate. As such, the heuristic under-allocates capacity to the first two class. We observe that our heuristic performs worse than the best dedicated staffing for the larger system size and shift length $\tau = 12$. We can expect the heuristic's performance to further worsen as the discrepancy between the arrival rate and offered load in Phase 2 widens.

6. Implementation Setting

Beginning with the 7am shift on March 19, 2018, nurses at the NYP Weill Cornell ED were allocated to areas based on recommendations provided by our heuristic described in Section 5.2. We now provide details of the implementation.

6.1. Pilot Study

We developed an online Web App (see Appendix A.4 for screenshots) that would take as inputs, ED and boarding patient census in each area; number of ED and EDIN nurses available; the maximum number of patients to be assigned to a single ED and EDIN nurse, and provide a recommendation for assignment of the available nurses to each area. The Web App is also able to take and record the

actual staffing used during a shift. The recommendation is obtained using the approach proposed in Section 5.2 and the same parameters as in the calibrated example of Section 5.4. In consultation with the charge nurse, we allocated a minimum number of ED nurses to each area, and then allocated the remaining ED nurses according to our solution approach. The minimum assigned nurses for areas A, B, C, and U were respectively 4, 2, 2, 1, if the number of available ED nurses was greater than or equal to 9; equal to 1 for all areas if it was between 3 and 9; and 0 otherwise.

For the shifts beginning from 7:00am March 5, 2018 to 7:00pm March 18, 2018 – the 'OFF' period – the charge nurse would input the system state (census and staffing). No staffing recommendation was provided and the charge nurse would allocate nurses to the different areas using his/her own experience and knowledge. The implemented allocation was reported through the Web App. For the shifts beginning from 7:00am March 19, 2018 to 7:00pm April 3, 2018 – the 'ON' period – the charge nurse would input the census and the number of nurses available. The Web App would output a staffing recommendation, specifying the number of ED and EDIN nurses to allocated to each area. Finally, to ensure willingness of nurse staffing to participate in the pilot, the charge nurse had full discretion to follow or not follow the recommended allocation; thus, she would also report the staffing actually used and if it did not coincide with the recommendation, the reason why. We note that the same team of charge nurses rotated throughout the pilot and were able to successfully record the data for all shits. Further, the nursing manager who oversaw the pilot was the same throughout the pilot.

During the OFF period the actual staffing used coincidentally matched with the full recommendation of our proposed solution 1 out of 28 shifts (3.6%). The recommendation for the allocation of ED (EDIN) nurses coincided 7 (5) out of 28 shifts (25% (17.9%)). During the ON period, the recommended staffing allocation was followed fully for 38.9% of the shifts. The ED (EDIN) allocation was followed 48.6% (55.8%) of shifts. The primary reason provided for why the recommendation was not followed related to the acuity level of the patients currently in the ED or anticipated to arrive. That is, either the acuity level of the current patients in one area was higher than usual, or the charge nurse expected a higher-than-usual number of high-acuity patients to arrive (e.g., due to bad weather). As we further discuss in Section 8, inability of our model in capturing the acuity level of patients within areas could be a main contributor to its partial implementation.

6.2. Data

Our data consists of all 7,933 visits who began or ended their visit to the WCMC ED from 12:00am March 1, 2018 - 12:59pm April 13, 2018. This was done in order to ensure that we had full records for all patients who were in the ED during the pilot period. The data contains operational level information (e.g., times stamps for arrival, first treatment, discharge, change to admit status as

well as assigned area and disposition) and patient level demographic and medical information (e.g., age, gender, race and ethnicity, insurance type, ESI level, and Elixhauser score; see Elixhauser et al. 1998). We restricted our analysis to patients who arrived and departed from the ED from 7:00am March 5, 2018 - 6:59am April 4, 2018. In addition, we excluded patients who were missing ESI or Elixhauser scores. The final cohort for analysis had 2,413 treated during the OFF period and 2,644 treated during the ON period.

Table A. 3 in Appendix A.4 presents means and standard deviations for patient characteristics. For most of them, there are no statistically significant differences during the ON versus OFF period. However, it appears that patients who arrived to the ED during the ON period were older, had more chronic diseases (as measured by Elixhauser), and were more likely to be on Medicare.

Table 2 summarizes the outcomes during the ON and OFF period. There is a statistically significant (at the $p < 0.001$ level) increase in waiting time of just under 7 minutes, and a statistically significant (at the $p < 0.01$ level) increase in provider evaluation to decision time of 35 minutes (7.7%). We note that during this time period, there was also a 9.6% increase in the number of patients arriving to the ED during the ON period. More specifically, 10.6% for Area A, 11.9% for Area B, 11% for Area C, and 3.8% for Area U. On the other hand, *all other measures of time* decrease with significance at the $p < 0.05$ level. This reduction is more than 2 hours for boarding time and 45 minutes for total ED LOS, as measured from arrival time to discharge. There is no statistically significant difference in the number of patients who walkout. There is a statistically significant (at the $p = 0.005$ level) increase in the number of patients who return to the WCMC ED within 10 days of initial ED visit. Note that we look at returns within 10 days to avoid censoring our outcome variable since our data set extends through the end of day of April 13, 2018, which is 10 days following the end of our pilot. While these comparisons do not control for various patient characteristics and/or staffing and census metrics, they are suggestive that the intervention may have improved patient flow in the ED, though perhaps at the cost of increasing returns to the ED. In the next section, we provide more rigorous analysis to explore this hypothesis.

7. Pilot Evaluation

To evaluate the impact of the staffing changes, we conduct a number of regression-based analyses. Before presenting the results, we discuss some of the challenges of the evaluation.

7.1. Evaluation Challenges

First, the intervention was not implemented in all shifts. Therefore, we cannot measure the impact of a full implementation (as in our simulation experiments). Instead, we consider the intervention to be the availability of a decision support tool that provides recommendations regarding the assignment of nurses to different areas of the ED. Note that even if the recommendations were not

Table 2 Summary Statistics of Patient Outcomes at WCMC

	OFF Period ($N = 2,413$)	ON Period ($N = 2,644$)	Total ($N = 5,057$)	p-value
Waiting time (min)	33.28 (32.78)	40.00 (44.17)	36.80 (39.28)	< 0.001
Prov eval. to decision (hrs)	4.50 (3.78)	4.85 (4.45)	4.68 (4.15)	0.003
Boarding time (hrs)	16.72 (15.58)	14.11 (12.22)	15.37 (14.01)	< 0.001
Prov eval. to discharge (hrs)	10.09 (12.32)	9.42 (10.29)	9.74 (11.31)	0.036
Total ED LOS (hrs)	10.43 (16.84)	9.68 (10.03)	10.04 (11.16)	0.001
Returns within 10 days	7.0%	9.2%	8.2%	0.005
Walkout	3.2%	3.4%	3.3%	0.619

Note. Mean and standard deviation (in parentheses) shown for continuous variables. p-values provided for testing differences between the mean (or distribution for categorical variables) in the Off versus On periods. Returns captures the percentage of patients who returned to the WCMC ED within 10 days of the first ED visit (regardless of the patient's disposition). Walkout includes all patients who left without being seen, walked out after evaluation, walked out before evaluation, and left against medical advice.

fully followed, they could have influenced the staffing decision, e.g. the recommendation could be partially followed. Second, we know that the number of patients increased substantially during the ON period. As such, the expected value of the metrics would have likely changed during the ON period, even in the absence of the intervention. Finally, we assume patients are exogenously assigned to treatment (ON) or control (OFF). This seems reasonable as patients (and even physicians and nurses, except for those in leadership positions who signed off on the pilot program) were unaware of our pilot, so could not have made decisions about visiting the ED based on whether it was during the ON or OFF period. That said, if patients are systematically less sick in the ON period, this could violate our assumption and bias our results. In addition, other unobserved confounders, such as changes in physician staffing or how the staff responded to higher workload during the ON period, could also potentially bias our results. By our summary statistics in Table A. 3, patients did not appear to be less sick; if anything they appeared to be more sick during the ON period (see, e.g., the Elixhauser score). Further, we are not aware of any systematic changes in staffing or other operational policies in the ED during the ON period. However, we cannot completely rule out the possibility that unobserved confounders were present during the study period.

When considering the implementation of an operational intervention that impacts staffing and, therefore, all patients who arrive to the ED during the intervention period, it is not possible to randomize patients to treatment versus control. This allocation is completely determined by their independent decision of when to visit the ED. Such realities introduce challenges in establishing clear, clean statistical inference on the impact of our proposed algorithm. Despite these challenges, we conduct a standard pre/post implementation analysis followed by a number of additional analyses to examine the robustness of our results.

7.2. Pre and Post Implementation Analysis and Results

For our continuous outcome measures (Waiting time, provider evaluation to disposition decision, boarding time, provider evaluation to discharge, and total ED LOS), we start with the following

Table 3 Pre and Post Implementation Results

	Waiting Time (min)	Prov Eval to decision (hrs)	Boarding Time (hrs)	Prov Eval to discharge (hrs)	Total LOS (hrs)	Returns within 10 days	Walkout
ON (θ)	4.03***	0.16	-3.96***	-1.73***	-1.74***	0.01	0.00
95% CI	[1.86,6.21]	[-0.09,0.41]	[-5.61,-2.31]	[-2.35,-1.12]	[-2.34,-1.14]	[-0.00,0.03]	[-0.01,0.01]
Num. of obs.	4758	4753	1286	4758	4817	4817	4817
R^2	0.177	0.082	0.081	0.211	0.220	0.049	0.048

Note. Includes controls for age, gender, race, ethnicity, insurance, ESI, Elixhauser, shift, number of nurses, and ED census. The model for returns also controls for disposition. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Table A. 4 in the appendix provides the coefficients and 95% confidence intervals for all control variables.

reduced form equation:

$$y_i = \beta X_i + \theta ON_i + \xi AM_i + \psi Num\_Nurses_i + \gamma ED\_census_i + \epsilon_i, \tag{26}$$

where y_i is the continuous outcome of interest. For discrete outcome measures (Returns and Walkouts), we use a Linear Probability Model with the same specification as the continuous outcomes models; the main difference being that y_i can only take on values of 0 or 1. For completeness, we also consider Probit and Logistic models for these discrete outcomes.

In equation (26) X_i is a vector of patient characteristics: age, gender, race, ethnicity, insurance, ESI, and Elixhauser. AM_i is an indicator which is equal to 1 if patient i arrives during the day shift (i.e., between 7:00am and 6:59pm). Num_Nurses_i indicates the number of nurses working during the shift. ED_census_i denotes that number of patients in the ED, including those waiting to see a provider, those "in treatment", and those boarding. ON_i is an indicator which is equal to 1 if the patient arrives to the ED during the ON period. ϵ_i denotes the standard error term. For Returns to the ED within 10 days, we also control for discharge disposition, as whether or not a patient is admitted to the hospital could have a substantial impact on the likelihood of return. Due to data limitations, we are not able to consider longer periods for which returns may occur.

7.2.1. Results Table 3 summarizes our main results by providing the coefficient estimate θ for patients arriving to the ED during the ON period relative to the OFF period. Full regression results are provided in Appendix A.4. After controlling for various patient characteristics and operational metrics, we see that the availability of the new staffing approach is associated with an increase of 4.03 minutes in waiting time, and a decrease of 3.96, 1.73, and 1.74 hours in boarding time, provider evaluation to ED discharge time, and total ED LOS, respectively. The impact on provider evaluation to decision was not statically significant (p-value: 0.200). The impact on all other continuous outcomes is significant at the $p<0.001$ level, although the main driver of the impact on provider evaluation to discharge and total LOS appears to be the reduction in boarding time for admitted patients.

Recall that our queueing model assumes that the service times of patients in each phase of service are exogenous to the staffing allocation. Thus, the sole improvement to ED flow predicted by our model is for waiting time. However, the results of our pilot suggest that the waiting time increased by several minutes. That said, the arrivals to the ED during the ON period increased significantly, so without our intervention it is possible the waiting time would have increased even more during the ON Period. We carefully examine this hypothesis in Appendix A.3 using our calibrated simulation model. We find that under dedicated staffing, the observed increase in arrival rates can lead to instability and results in a much larger increase in waiting times than the observed 12%, whereas our algorithm can still stabilize all areas and reduce queue-lengths significantly compared to dedicated staffing. It is important to note that although in our regression model we control for census, the impact of increased load on waiting times is highly nonlinear, in particular when the utilization is high. As such, our linear model is unable to properly account for the impact of large increases in census. We further discuss this point in Section 8.

While our data does not allow us to identify exactly the mechanism which led to the decrease in other measures, we developed conjectures through discussion with our clinical collaborators. We see that boarding time is significantly reduced in the ON period. At first glance, this may be unexpected because the prevailing sentiment is that ED boarding is largely driven by congestion in the inpatient units, but our intervention did not involve the inpatient units. Our clinical collaborators suggested that better alignment of nursing staff with patient demand may have reduced the time required for nurses to complete the hospital admission process. This process is very paper-work intensive and requires substantial coordination between the nursing staff and the inpatient teams to find a bed and facilitate hand-offs. Thus, we conjecture that the nurse staffing allocation resulted in reductions in waiting times between nursing tasks to facilitate the inpatient admission.

We also observe a large decrease in the time spent in the ED once a patient is seen by a provider, as measured by provider evaluation to discharge and total LOS. Although the majority of the decrease in provider evaluation to discharge can be attributed to the reduction in boarding time, the total LOS measure incorporates both the initial waiting time (which appears to increase) and the time from provider evaluation to discharge (which appears to decrease). Hence, the reduction in total LOS suggests a reduction in service time which we did not account for in our model. However, provider evaluation to discharge often involves a number of internal "waits", e.g., for a nurse to review the result of a lab or rotate between patients (e.g. Dobson et al. 2013) or finalize the discharge process for a patient after the discharge decision is made by a physician. Thus, by better matching nursing availability with patient demand via our new staffing approach, these mid-process waits could have been reduced.

Finally, we do not find a statistically significant impact at the $p < 0.05$ level of the new staffing approach on returns to the ED in 10 days (p-value = 0.116) or on walkouts (p-value = 0.808). We see similar results with the logistic and probit models.

We further explore the robustness of these results by: (1) conducting sensitivity analysis with respect to a hypothetical confounder to see how large of an effect it must have to render the results statistically insignificant (Appendix A.5); and (2) robustness check with respect to model specification using alternative sets of covariates (Appendix A.6). These analyses suggest that the results for all continuous outcomes are quite robust, except for the small increase in waiting time. The results for binary outcomes are also robust except for the returns to the ED in 10 days, for which an unobserved confounder could result in a positive and statistically significant coefficient.

8. Discussion and Future Work

In this work, we propose a queueing network model of patient flow to guide the assignment of two types of nursing staff to different areas of the ED at the beginning of each shift. We formulate an associated fluid control problem for this queueing network and propose a heuristic solution. By analyzing the dynamics of the fluid model over a shift, we gain insights into parameter regimes where the proposed solution can perform well. Simulation experiments confirm these observations and suggest that the proposed approach can reduce waiting times in a parameter regime relevant to our ED application.

Through a pilot implementation of our approach and empirical comparison of the pre/post data, we find that guiding the reassignment of nursing staff using our proposed approach is associated with significant improvement of patient flow in the ED. These gains are observed despite the fact that the solution proposed by our model was not fully implemented during all shifts and the number of patients increased significantly during the ON period. The presence of these real-world challenges highlight the difficulty of deriving clean statistical inferences of operational interventions in complex settings such as EDs.

We believe that the simplicity of the proposed approach played an important role in convincing the ED leadership to conduct the pilot study. In addition, making sure the practical constraints (e.g., restriction to reassignment at the beginning of the shift, reserving capacity for each area) were taken into consideration, the easy-to-use Web App, and allowing the practitioners to overrule the proposed assignment, also contributed to the success of the pilot. At the same time, some of these modeling choices and simplifications have resulted in discrepancies between the model and empirical setting and made it more difficult to asses the true impact of our staffing approach. In particular, since the recommendations of the model were only partially implemented, we cannot draw any conclusions regrading the impacts of fully following the recommendations and how these

impacts may generalize to other EDs. In the following, we discuss the limitations of our study and propose directions for future work to address them.

Model limitations: Some of the discrepancies between the theoretical and empirical results can be attributed to our modeling assumptions. First, we assume that service times are exogenous and stationary. Phase 1 service times could however vary depending on the workload assigned to each nurse and hence may not be exogenous to the nursing assignment. In addition, Phase 2 service times depend on the congestion level and operations of the inpatient wards. As such, considering a state-dependent service time in Phase 1 and a non-stationary service rate in Phase 2 to capture the dynamics of inpatient ward admissions (see, e.g., Shi et al. 2016 and Dong and Perry 2020) would improve the fidelity of our model and make it more comparable to the real ED settings. These extensions can be incorporated rather easily in the fluid model, but solving the control problem would then become more complicated, requiring a new approach. Second, when evaluating the performance of our algorithm in the simulation experiments, we assume that the status-quo policy is (an optimal) dedicated staffing. Our pilot demonstrates deviations from both policies. That is, during the OFF period the assignments did not always follow a uniform dedicated staffing, and during the ON period our recommendations were only partially followed. A primary reason for deviations from both policies appear to be the acuity of the patients within each area which was not captured in our model. Therefore, an important extension of our model would be to account for different acuity levels of patients, e.g., by introducing multiple customer classes within each area. In this case, a control policy would map, not only the number of patients in each area, but their acuity levels (classes) to a reassignment decision. We note that even by considering observable variables related to patient acuity, the resulting policy is unlikely to be perfectly followed in practice. Further, our results suggest that even partial implementation may be effective. As such, examining the robustness of the performance to level of conformity would be an interesting topic to consider for future work.

Limitation of the empirical study: Our empirical study highlights several challenges of evaluating the benefits of an operational intervention in the complex ED environment. In particular, our results suggest that accounting for the nonlinear queueing effects could be critical in accurate measurements of the benefits of operational policies aimed at reducing congestion. This limitation can be addressed using more general models such as the non-parametric partially linear model (Robinson 1988) which may require larger sample sizes and hence a longer pilot. Further, although sensitivity analysis suggests that our results are to some extent robust to unobserved patient characteristics and operational variables, the inability to account for the potential effect of confounders is a major limitation of observational studies like ours. In particular, although we have made conjectures on the mechanisms behind the observed improvements, we cannot rule out other possible

mechanisms due to data limitations. Since randomization is practically infeasible in such operational interventions, using more detailed patient-level and operational variables (e.g. utilization of inpatient wards) as well as utilizing other study designs such as quasi-experimental designs should be considered in future studies.

Other extensions: Beyond above points, there are a number of additional extensions that can be considered in future work. Here, we briefly summarize a few. (1) Our data points to a considerable variability in the number of available nursing staff (e.g., due to absenteeism) as well as the arrival rate, which were ignored in our formulation. An extended problem can explicitly account for the parameter uncertainty and investigate its impact along the same lines as done in the call center literature, e.g., Harrison and Zeevi (2005) and Atlason et al. (2008) (arrival rate uncertainty) and Whitt (2006) (arrival and capacity uncertainty). In addition to reassignment decision, a two-stage formulation can be considered to determine the number of scheduled staff as well as determining a control policy to assign them to different areas once the available number of staff is realized. (2) Another mechanism for balancing congestion between the areas is the routing decision. Although our observations from the data suggest that routing is primarily determined based on the care requirements of the patients, congestion of different areas does also affect the routing decision. Investigating the benefits of (partial) congestion-based routing by itself or jointly with server reassignment could be an interesting area for future work. Again, there are connections to the call center literature on skill-based routing (see, e.g., Harrison and Zeevi 2005, Armony 2005 and the references therein), but in the ED context a contextual routing policy which maps, not only congestion, but also patient characteristics to different areas would be more practical.

Acknowledgments

The authors would like to thank the anonymous referees and Associate Editor, and the Department Editor Diwakar Gupta for their constructive comments and suggestions. The authors are grateful to Dr. Rahul Sharma, Dr. Sunday Clark, Christel Vasquez, and Brian Miluszusky at Weill Cornell Medicine, without whom this project would not have been possible. The authors also gratefully acknowledge Michael Huang who developed the web application for the pilot study.

References

Ackroyd-Stolarz S, Guernsey JR, Mackinnon N, Kovacs G (2011) The association between a prolonged stay in the emergency department and adverse events in older patients admitted to hospital: a retrospective cohort study. *BMJ quality & safety* 20(7):564–569.

Armony M (2005) Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* 51(3):287–329.

Atlason J, Epelman MA, Henderson SG (2008) Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Management Science* 54(2):295–309.

Baker SJ, Esbenshade A (2015) Partnering effectively with inpatient leaders for improved emergency department throughput. *Advanced Emergency Nursing Journal* 37(1):65.

Bassamboo A, Harrison JM, Zeevi A (2006) Design and control of a large call center: Asymptotic analysis of an LP-based method. *Operations Research* 54(3):419–435.

Carnegie NB, Harada M, Hill JL (2016) Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness* 9(3):395–420.

Chan CW, 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):1936–1959.

Chan TC, Huang SY, Sarhangian V (2022) Dynamic control of service systems with returns: Application to design of post-discharge hospital readmission prevention programs. *arXiv preprint arXiv:2203.00093* .

Chen J, Dong J, Shi P (2021) Optimal routing under demand surges: The value of future arrival rates. *Available at SSRN 3980227* .

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.

Dai J, Tezcan T (2011) State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research* 36(2):271–320.

Dobson G, Tezcan T, Tilson V (2013) Optimal workflow decisions for investigators in systems with interruptions. *Management Science* 59(5):1125–1141.

Dong J, Perry O (2020) Queueing models for patient-flow dynamics in inpatient wards. *Operations Research* 68(1):250–275.

Elixhauser A, Steiner C, Robert DRH, Coffey RM (1998) Comorbidity measures for use with administrative data. *Medical care* 36(1):8–27.

Green LV, Savin S, Savva N (2013) "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* 59(10):2237–2256.

Green LV, Soares J, Giglio JF, Green RA (2006) Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* 13(1):61–68.

Gurumurthi S, Benjaafar S (2004) Modeling and analysis of flexible queueing systems. *Naval Research Logistics (NRL)* 51(5):755–782.

Harrison JM, Zeevi A (2005) A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management* 7(1):20–36.

He B, Dexter F, Macario A, Zenios S (2012) The timing of staffing decisions in hospital operating rooms: incorporating workload heterogeneity into the newsvendor problem. *Manufacturing & Service Operations Management* 14(1):99–114.

He S, Sim M, Zhang M (2019) Data-driven patient scheduling in emergency departments: A hybrid robust-stochastic approach. *Management Science* 65(9):4123–4140.

Helm JE, AhmadBeygi S, Van Oyen MP (2011) Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management* 20(3):359–374.

Hu Y, Chan CW, Dong J (2021) Optimal scheduling of proactive service with customer deterioration and improvement. *Management Science* .

Huang J, Carmeli B, Mandelbaum A (2015) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* 63(4):892–908.

Imbens GW (2003) Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* 93(2):126–132.

Iravani SM, Kolfal B, Van Oyen MP (2011) Capability flexibility: a decision support methodology for parallel service and manufacturing systems with flexible servers. *IIE Transactions* 43(5):363–382.

Iravani SM, Van Oyen MP, Sims KT (2005) Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Science* 51(2):151–166.

Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. *Management Science* 41(4):577–594.

Kim SH, Whitt W (2014) Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management* 16(3):464–480.

Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for markovian service networks. *Queueing Systems* 30(1):149–201.

McNabb A (1986) Comparison theorems for differential equations. *Journal of Mathematical Analysis and Applications* 119(1-2):417–428.

Pang G, Talreja R, Whitt W, et al. (2007) Martingale proofs of many-server heavy-traffic limits for markovian queues. *Probability Surveys* 4:193–267.

Plunkett PK, Byrne DG, Breslin T, Bennett K, Silke B (2011) Increasing wait times predict increasing mortality for emergency medical admissions. *European Journal of Emergency Medicine* 18(4):192–196.

Robinson PM (1988) Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society* 931–954.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5):1080–1097.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2014) Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* 16(3):329–345.

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

Singer AJ, Thode Jr HC, Viccellio P, Pines JM (2011) The association between length of emergency department boarding and mortality. *Academic Emergency Medicine* 18(12):1324–1329.

Sir MY, Nestler D, Hellmich T, Das D, Laughlin Jr MJ, Dohlman MC, Pasupathy K (2017) Optimization of multidisciplinary staffing improves patient experiences at the mayo clinic. *Interfaces* 47(5):425–441.

Tsitsiklis JN, Xu K (2013) On the power of (even a little) resource pooling. *Stochastic Systems* 2(1):1–66.

Tsitsiklis JN, Xu K (2017) Flexible queueing architectures. *Operations Research* 65(5):1398–1413.

Whitt W (2006) Staffing a call center with uncertain arrival rate and absenteeism. *Production and operations management* 15(1):88–102.

Yankovic N, Green LV (2011) Identifying good nursing levels: A queuing approach. *Operations Research* 59(4):942–955.

Yildiz O, Kamali M, Tezcan T (2019) When to use provider triage in emergency departments. *Management Science* 65(3):1003–1019.

Yom-Tov GB, 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.

Zayas-Cabán G, Xie J, Green LV, Lewis ME (2016) Dynamic control of a tandem system with abandonments. *Queueing Systems* 84(3-4):279–293.

Online Appendix

A.1. Additional Technical Results and Proofs

Proposition 3 *Suppose* $(u^*(t), w^*(t))$ *and* $(x^*(t), y^*(t))$ *solve* (9)–(15). *Then* $u^*(t)$ *also solves:*

$$\min_{u(\cdot)} \ \frac{1}{T}\int_0^T q(t)\cdot e\, dt, \tag{A1}$$

$$s.t.\ q(t) = \Big(x(t) - \Big(u(t) - (y^*(t) - w^*(t))^+\Big)\Big)^+, \quad \forall t \in [0,T], \tag{A2}$$

$$\dot{x}(t) = \lambda(t) - \hat{\mu}\cdot\Big(x(t) \wedge \Big(u(t) - (y^*(t) - w^*(t))^+\Big)\Big), \quad \forall t \in [0,T], \tag{A3}$$

$$u(t)\cdot e \le n_1, \quad \forall t \in [0,T], \tag{A4}$$

$$u(t) \ge 0, \quad \forall t \in [0,T], \tag{A5}$$

$$x(0) = x^0,\ y(0) = y^0. \tag{A6}$$

Proof of Proposition 3. Let $(u^*(t), w^*(t))$ be the solution to (9)–(15). Then, redefining $u(t)$ as $u(t) - (y(t) - w(t))^+$, the claim follows by definition. □

In the following proofs, we make use of the following differential comparison result proved in McNabb (1986).

Lemma 1 *Suppose* $z_1(a) \le z_2(a)$ *and* $f(t,\cdot)$ *is Lipschitz continuous. If for all* $t > a$*:*

$$\dot{z}_1(t) - f(t, z_1(t)) \le \dot{z}_2(t) - f(t, z_2(t)), \tag{A7}$$

then $z_1(t) \le z_2(t)$ *for all* $t > a$.

Proof of Proposition 1. Consider the dynamics of Phase 2 fluid during the shift and observe that,

$$\dot{y}_i(t) = p_i\mu_i(x_i(t) \wedge (u_i - (y_i - w_i)^+) - \nu_i y_i(t) \tag{A8}$$

$$\le p_i\mu_i u_i - \nu_i y_i(t). \tag{A9}$$

Now consider the ODE,

$$\dot{\tilde{y}}_i(t) = p_i\mu_i u_i - \nu_i \tilde{y}_i(t), \tag{A10}$$

starting from $\tilde{y}_i(t) = y_i^0$, which has the following solution:

$$\tilde{y}_i(t) = \Big(y_i^0 - \frac{p_i\mu_i u_i}{\nu_i}\Big)e^{-\nu_i t} + \frac{p_i\mu_i u_i}{\nu_i}. \tag{A11}$$

Since $x_i^0 > u_i$ by assumption, it follows from (A9) that $\dot{\tilde{y}}_i(0) = \dot{y}_i(0)$ which then using (A10) implies that $\tilde{y}_i(t) < y_i^0$ for all $t \in [0,\tau]$. Next, by (A9) and (A10) we have $\dot{y}_i(t) - p_i\mu_i u_i + \nu_i y_i(t) \le \dot{\tilde{y}}_i(t) - p_i\mu_i u_i + \nu_i\tilde{y}_i(t)$, which implies using Lemma 1 that $y_i(t) \le \tilde{y}_i(t) < y_i^0 < w_i$. It follows that $(y_i(t) - w_i)^+ = (y_i^0 - w_i)^+ = 0$ for all $t \in [0,\tau]$ and hence the claim follows from (6) and (22) noting that $x_i(t)$ and $\tilde{x}_i(t)$ are initialized at the same value and have the same dynamics during the shift. □

Proof of Proposition 2. We begin with the case where $\dot{y}_i(0) > 0$ and derive the exact distance between $x_i(t)$ and $\tilde{x}_i(t)$. Consider the Phase 2 dynamics during the shift governed by the ODE,

$$\dot{y}_i(t) = p_i\mu_i(x_i(t) \wedge (u_i - (y_i(t) - w_i)^+) - \nu_i y_i(t). \tag{A12}$$

Note that by assumption $y_i^0 > w_i$ and $x_i^0 > u_i - (y_i^0 - w_i)$ since $(u_i, w_i) \in S$. Therefore, there exists a $t_0 > 0$, such that for all $t \in [0, t_0)$:

$$\dot{y}_i(t) = p_i\mu_i(u_i - (y_i(t) - w_i)) - \nu_i y_i(t), \tag{A13}$$

which has the following solution:

$$y_i(t) = \Big(y_i^0 - \frac{p_i\mu_i(u_i + w_i)}{p_i\mu_i + \nu_i}\Big)e^{-(p_i\mu_i+\nu_i)t} + \frac{p_i\mu_i(u_i + w_i)}{p_i\mu_i + \nu_i}. \tag{A14}$$

We now claim that (A14) holds for all $t \in [0, \tau]$. To show this note that by assumption $\dot{y}_i(0) \geq 0$ and hence $y(t)$ is nondecreasing in $t \in [0, t_0]$. It follows that,

$$\begin{aligned}
\dot{\tilde{x}}_i(t) &= \lambda_i(t) - \mu_i(u_i - (y_i^0 - w_i)) && \text{(A15)}\\
&\leq \lambda_i(t) - \mu_i(u_i - (y_i(t) - w_i)) && \text{(A16)}\\
&= \dot{x}_i(t), && \text{(A17)}
\end{aligned}$$

for all $t \in [0, t_0]$ which together with $\tilde{x}_i(0) = x_i(0) = x_i^0$ implies that for all $t \in [0, t_0]$,

$$u_i - (y_i(t) - w_i) \leq u_i - (y_i^0 - w_i) \leq \tilde{x}_i(t) \leq x_i(t). \tag{A18}$$

Applying the same argument iteratively starting at $t = t_0$ establishes the claim for all $t \in [0, \tau]$. Thus, we have for $t \in [0, \tau]$:

$$\begin{aligned}
\dot{x}_i(t) &= \lambda_i(t) - \mu_i(x_i(t) \wedge (u_i - (y_i(t) - w_i)^+)) && \text{(A19)}\\
&= \lambda_i(t) - \mu_i(u_i - (y_i(t) - w_i)) && \text{(A20)}\\
&= \lambda_i(t) - \mu_i(u_i - (y_i^0 - w_i) - (y_i - y_i^0)) && \text{(A21)}\\
&= \lambda_i(t) - \mu_i\left(u_i - (y_i^0 - w_i) - \left(y_i^0 - \frac{p_i\mu_i(u_i + w_i)}{p_i\mu_i + \nu_i}\right)(e^{-(p_i\mu_i+\nu_i)t} - 1)\right). && \text{(A22)}
\end{aligned}$$

Integrating and substituting for $\tilde{x}_i(t)$ we have:

$$x_i(t) - \tilde{x}_i(t) = \Big(\frac{p_i\mu_i(u_i + w_i)}{p_i\mu_i + \nu_i} - y_i^0\Big)\Big(\frac{e^{-(p_i\mu_i+\nu_i)t} - 1}{p_i\mu_i + \nu_i} + t\Big). \tag{A23}$$

Next, we turn to the case with $\dot{y}_i(0) \leq 0$ where $y_i(t)$ is nonincreasing in a neighborhood of 0. We first note that since $\dot{y}_i(0) \leq 0$, we have using the same argument as in the proof of Proposition 1 that $y_i(t) < y_i^0$ for all $t \in [0, \tau]$. Thus,

$$\begin{aligned}
\dot{x}_i(t) &= \lambda_i(t) - \mu_i(x_i(t) \wedge (u_i - (y_i(t) - w_i)^+)) && \text{(A24)}\\
&\leq \lambda_i(t) - \mu_i(x_i(t) \wedge (u_i - (y_i^0 - w_i)^+)). && \text{(A25)}
\end{aligned}$$

Using Lemma 1 we then have that $x_i(t) \le \tilde{x}_i(t)$ for all $t \in [0,\tau]$. Next note that,

$$\dot{x}_i(t) = \lambda_i(t) - \mu_i(x_i(t) \wedge (u_i - (y_i(t) - w_i)^+)) \tag{A26}$$

$$\geq \lambda_i(t) - \mu_i x_i(t) \tag{A27}$$

$$\geq \tilde{\lambda}_i - \mu_i x_i(t), \tag{A28}$$

and let $z_i(t)$ be a solution to the initial value problem,

$$\dot{z}_i(t) = \tilde{\lambda}_i - \mu_i z_i(t), \tag{A29}$$

with $z_i(0) = x_i^0$. We have,

$$z_i(t) = \Big(x_i^0 - \frac{\tilde{\lambda}_i}{\mu_i}\Big)e^{-\mu_i t} + \frac{\tilde{\lambda}_i}{\mu_i} \tag{A30}$$

$$\geq x_i^0 \wedge \frac{\tilde{\lambda}_i}{\mu_i}. \tag{A31}$$

Applying Lemma 1 we can conclude that $x_i(t) \geq z_i(t) \geq x_i^0 \wedge (\tilde{\lambda}_i/\mu_i)$ on $[0,\tau]$. Then, noting that $y_i(t) < y_i^0$, and $x_i^0 \geq u_i - (y_i^0 - w_i)$, we have,

$$\dot{y}_i(t) = p_i\mu_i(x_i(t) \wedge (u_i - (y_i(t) - w_i)^+)) - \nu_i y_i(t) \tag{A32}$$

$$\geq p_i\mu_i \left(\frac{\tilde{\lambda}_i}{\mu_i} \wedge (u_i - (y_i^0 - w_i))\right) - \nu_i y_i^0. \tag{A33}$$

Again using Lemma 1 we can then conclude that,

$$y_i(t) \geq \Big(y_i^0 - \frac{p_i\mu_i(u_i + w_i - y_i^0) \wedge p_i\tilde{\lambda}_i}{\nu_i}\Big)e^{-\nu_i t} + \frac{p_i\mu_i(u_i + w_i - y_i^0) \wedge p_i\tilde{\lambda}_i}{\nu_i}. \tag{A34}$$

Going back to $\dot{x}_i(t)$ and substituting from (A34) we have,

$$\dot{x}_i(t) = \lambda_i(t) - \mu_i(x_i(t) \wedge (u_i - (y_i(t) - w_i)^+)) \tag{A35}$$

$$\geq \lambda_i(t) - \mu_i(u_i - (y_i(t) - w_i)) \tag{A36}$$

$$= \lambda_i(t) - \mu_i(u_i - (y_i^0 - w_i) - (y_i(t) - y_i^0)) \tag{A37}$$

$$\geq \lambda_i(t) - \mu_i\Big(u_i - (y_i^0 - w_i) - \Big(y_i^0 - \frac{p_i\mu_i(u_i + w_i - y_i^0) \wedge p_i\tilde{\lambda}_i}{\nu_i}\Big)(e^{-\nu_i t} - 1)\Big). \tag{A38}$$

Integrating and substituting for $\tilde{x}_i(t)$, we obtain:

$$x_i(t) - \tilde{x}_i(t) \geq -\frac{\mu_i}{\nu_i^2}\big(\nu_i y_i^0 - (p_i\mu_i(u_i + w_i - y_i^0) \wedge p_i\tilde{\lambda}_i)\big)(e^{-\nu_i t} + \nu_i t + C) \tag{A39}$$

$$= \frac{\mu_i}{\nu_i^2}\big((p_i\mu_i(u_i + w_i - y_i^0) \wedge p_i\tilde{\lambda}_i) - \nu_i y_i^0\big)(e^{-\nu_i t} + \nu_i t + C) \tag{A40}$$

$$= \frac{\mu_i}{\nu_i^2}\big(\dot{y}_i(0) \wedge (p_i\tilde{\lambda}_i - \nu_i y_i^0)\big)(e^{-\nu_i t} + \nu_i t + C), \tag{A41}$$

for some constant C, where (A41) follows since $\dot{y}_i(0) = p_i\mu_i(u_i + w_i - y_i^0) - \nu_i y_i^0$. Substituting the initial value $x_i(0) = \tilde{x}_i(0) = x_i^0$, and solving for C we obtain the claimed result. □

A.2. Validity of the observations made in Section 5.3

We first obtain a tighter estimate of the error presented in Proposition 2 for sinusoidal arrivals.

Corollary 1 *Let $\lambda_i(t) = \alpha_i + \beta_i \sin(\pi(t+s)/m)$ for $m > 0$ and suppose the assumptions of Proposition 2 hold. Then, in the case with $\dot{y}_i(0) < 0$, we have for all $t \in [0, \tau]$,*

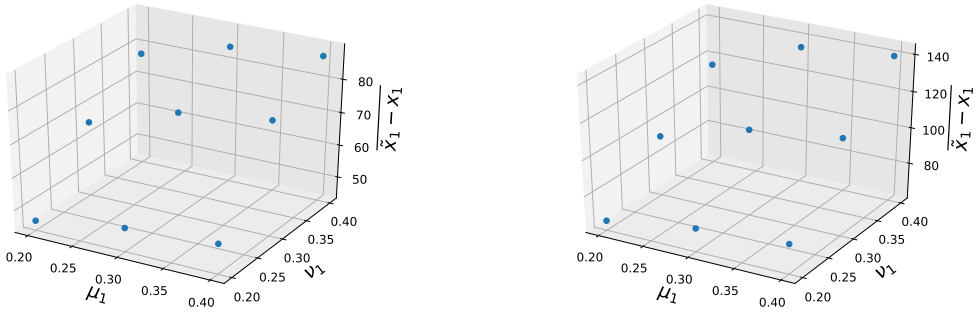
$$0 \le \tilde{x}_i(t) - x_i(t) \le \frac{\mu_i}{\nu_i^2} \max\left\{ |\dot{y}_i(0)|, \left| p_i \left(\alpha_i - \frac{m\mu_i\beta_i}{\sqrt{\pi^2 + m^2\mu_i^2}} \right) - \nu_i y_i^0 \right| \right\} (e^{-\nu_i t} - 1 + \nu_i t). \tag{A42}$$

Proof of Corollary 1. For sinusoidal arrivals, we can obtain an explicit solution to (A29):

$$\begin{aligned} z_i(t) &= \frac{\alpha_i}{\mu_i} + \left(x_i^0 - \frac{\alpha_i}{\mu_i} - \frac{m\beta_i K}{\sqrt{\pi^2 + m^2\mu_i^2}} \right) e^{-\mu_i t} + \frac{m\beta_i}{\sqrt{\pi^2 + m^2\mu_i^2}} \sin\left(\frac{\pi(t+s)}{m} - \tan^{-1}\left(\frac{\pi}{m\mu_i} \right) \right) \\ &\ge x_i^0 \wedge \frac{\alpha_i}{\mu_i} - \frac{m\beta_i}{\sqrt{\pi^2 + m^2\mu_i^2}}, \end{aligned} \tag{A43}$$

where $K = \sin\left(\frac{\pi s}{m} - \tan^{-1}\left(\frac{\pi}{m\mu_i} \right) \right)$. Replacing (A31) with this lower-bound in the proof of Proposition 2, we obtain the result. □

Next, we numerically study the effect of system parameters on the average discrepancy between $x(t)$ and $\tilde{x}(t)$ over the length of a $\tau = 12$ hour shift. In Figure A. 1, we present the average deviation (under the conditions of Corollary 1) for different average boarding and service times and Phase 2 initial conditions. Other parameters are fixed at $n_1 = 32$ and $n_2 = 16$, $p_1 = 0.25$, $x_1^0 = 24$. The arrival rate is the same sinusoidal function in (21) with $\Lambda_1 = 5.6, \beta_1 = -2.8$. We observe that consistent with the discussion in Section 5.3, the average deviation increases with μ_i and ν_i. Also, the error increases with the Phase 2 initial condition, which aligns with our expectation that number of active Station 1 servers diverges faster from its initial value for larger $|\dot{y}_1(0)|$.

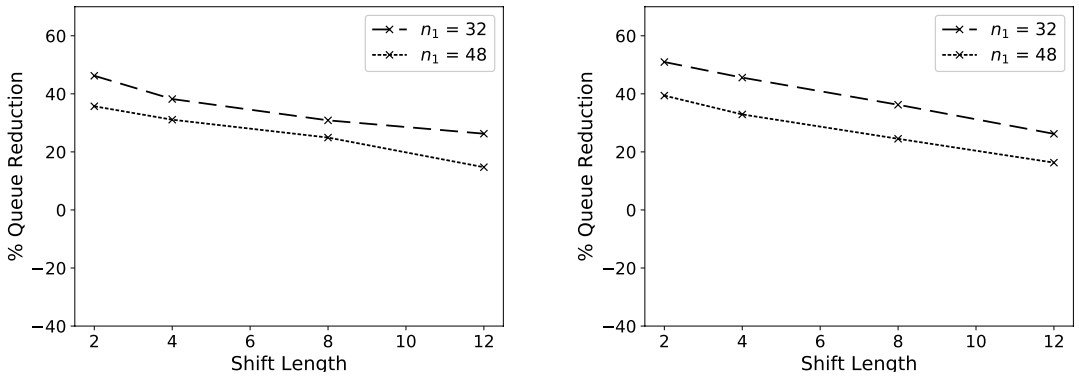


(a) $y_1^0 = 12$ (b) $y_1^0 = 14$

Figure A. 1 **Average deviation between $\tilde{x}_1(t)$ and $x_1(t)$ over a $\tau = 12$ hour shift for different boarding and service rates and two Phase 2 initial conditions of 1(a) $y_1^0 = 12$ and 1(b) $y_1^0 = 14$. Other parameters are $n_1 = 32$, $n_2 = 16$, $p_1 = 0.25$, $x_1^0 = 24$. The arrival rate is the same as in (21) with $\Lambda_1 = 5.6, \beta_1 = -2.8$**

A.3. Additional Simulation Experiments

Experiment 1. We consider a similar setting to that of the calibrated experiments, but with shorter service and boarding times and higher admission probability. In doing so, we aim to understand the performance degradation in regimes where the decoupling approach is expected to have higher error as identified in the results of Section 5.3. More specifically, the service times are identical across the four classes in both phases with $\mu_i = \nu_i = 0.5$ for $i \in \mathcal{I}$. Phase 1 and 2 utilization is set to $\rho_1 = 0.87$ and $\rho_2 = 0.86$, respectively, and $p = (0.60, 0.48, 0.60, 0)$. The Station 1 load is distributed evenly among the classes, i.e., $r_1 = (1/4, 1/4, 1/4, 1/4)$ and Station 2 load is set to $r_2 = (5/14, 4/14, 5/14, 0)$. For the transient experiments, the system is initialized at a similarly congested state as in the calibrated experiments, starting with $X(0) = (1.25n_1, 0.75n_1, 1.25n_1, 0.75n_1)$ and $Y(0) = (0.714n_2, 0.571n_2, 0.714n_2, 0)$. For simplicity we assume a nurse-to-patient ratio of 1 for both stations. Figure A. 2 presents estimates of the relative total queue length reduction achieved under the proposed staffing approach for different shift lengths and two system sizes; $n_1 = 32, n_2 = 14$ and $n_1 = 48, n_2 = 21$. We still observe a significant queue length reduction in both steady-state and transient experiments, although the gains are smaller compared to the calibrated experiment.

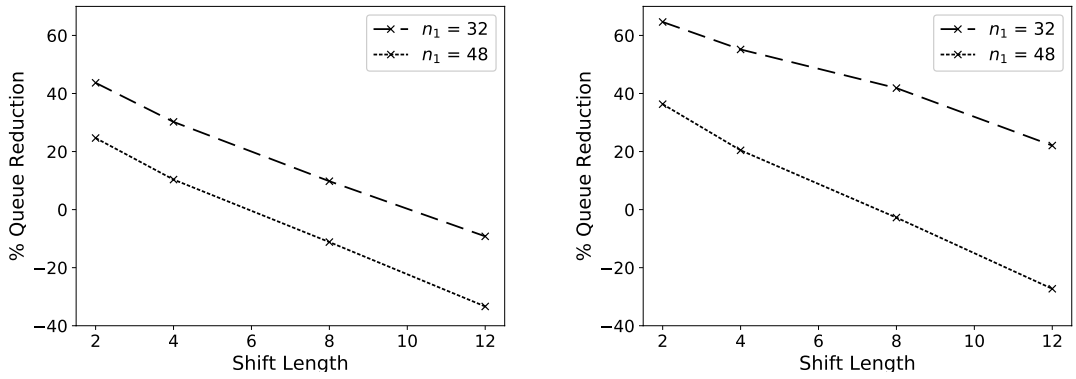


(a) Finite Horizon (b) Infinite Horizon

Figure A. 2 **Experiment 1: Performance of the proposed reassignment policy in terms of expected total average queue reduction for a system with** $\rho_1 = 0.87$, $\rho_2 = 0.86$, $r_1 = (1/4, 1/4, 1/4, 1/4)$ **and** $r_2 = (5/14, 4/14, 5/14, 0)$, $\mu_i = \nu_i = 0.5$ **for** $i \in \{1, 2, 3, 4\}$.

Experiment 2. In the second experiment, the utilizations are the same as in Experiment 1 but the load is unevenly distributed for both stations with $r_1 = (1/3, 1/3, 1/6, 1/6)$ and $r_2 = (3/8, 3/8, 1/4, 0)$. Service rates are $\mu_i = \nu_i = 0.34$ for $i \in \{1, 2\}$, and $\mu_i = \nu_i = 0.75$ for $i \in \{3, 4\}$. The fraction of patients who require Phase 2 treatment is $p = (0.49, 0.49, 0.65, 0)$. Note that classes 3 and 4 face a smaller load than 1 and 2, but have higher average arrival rates and admission probabilities. The system is initialized at $X(0) = (1.25n_1, 0.75n_1, 1.25n_1, 0.75n_1)$ and $Y(0) = (0.857n_2, 0.857n_2, 0.714n_2, 0)$. Figure A. 3 presents the results for different shift lengths and the same system sizes as in Figure A. 2. We observe that our heuristic performs worse than the best dedicated staffing for the larger system size and shift length $\tau = 12$. This can be attributed to the heuristic's strategy for assigning Station 2 servers in (17) without consideration for the boarding times.

In the second experiment, classes 3 and 4 face a smaller load in Phase 2 despite having a higher input rate. As such, the heuristic under-allocates capacity to the first two class. We can expect the heuristic's performance to further worsen as the discrepancy between the arrival rate and offered load in Phase 2 widens. On the other hand, when the discrepancy is small – as it is in our implementation setting – we expect the proposed approach to perform well.

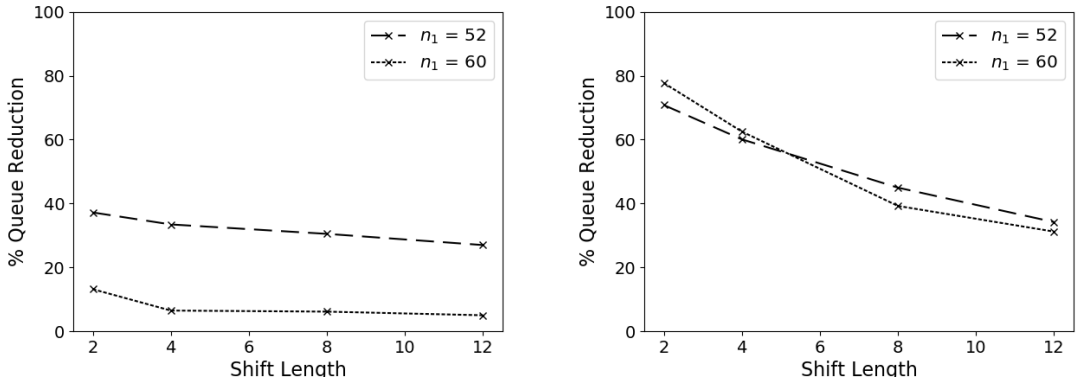


(a) Finite Horizon (b) Infinite Horizon

Figure A. 3 **Experiment 2: Performance of the proposed reassignment policy in terms of expected total average queue reduction for a system with** $\rho_1 = 0.87$, $\rho_2 = 0.86$, $r_1 = (1/3, 1/3, 1/6, 1/6)$ **and** $r_2 = (3/8, 3/8, 1/4, 0)$, $\mu_i = \nu_i = 0.34$ **for** $i \in \{1, 2\}$, **and** $\mu_i = \nu_i = 0.75$ **for** $i \in \{3, 4\}$.

Calibrated experiment with Log-Normal service times. To examine the robustness of our policies to variability in service times, we repeat the calibrated experiment but instead of exponential, use Log-Normal service times with the same means but standard deviations estimated using data from the OFF period. The estimated standard deviations were $(10.38, 12.15, 7.86, 18.7)$ and $(11.82, 13.29, 11.84, -)$ for Phase 1 and Phase 2, respectively. The results are presented in Figure A. 4. Although the Log-Normal service times in Phase 1 are more variable than the exponential ones used in the main experiments, we still observe a significant reduction in queue-lengths under our heuristic policy which only utilizes first-order (mean) service time information.

Calibrated experiment with inflated arrivals. We use our calibrated simulation model to evaluate the impact of the increased arrival rate in the absence of the intervention (i.e., under the optimal dedicated staffing) and for a shift-length of 12 hours. More specifically, we inflated the (average) arrival rates in the simulation by the same percentage the number of arrivals increased for each area during the ON period, i.e., 10.6% for Area A, 11.9% for Area B, 11% for Area C, and 3.8% for Area U. Table A. 1 presents the relative increase in expected waiting times under dedicated staffing as a result of inflating the arrival rates. For $n = 52$, Areas B and C become unstable in the infinite horizon experiment, and the total waiting time in Areas C and U increases by above 100%. Similarly, in the finite horizon experiments, we see the total waiting times increase by more than 100%. For $n = 62$, we also observe large increases in waiting times. In particular, for the finite-horizon experiments the waiting times increase by 43%, and for the infinite horizon



(a) Finite Horizon (b) Infinite Horizon

Figure A. 4 Performance of the proposed reassignment policy with Log-Normal service times.

experiment they increase by 91%. The results suggest that without the intervention the waiting times would have increased much more than the observed 12%.

To further examine the hypothesis, we investigate the robustness of our staffing approach to changes in arrival rate. To this end, we repeat the calibrated experiment with the inflated arrival rates but use the non-inflated arrival rates to obtain the heuristic staffing approach. In doing so we aim to understand whether our staffing approach could have been able to reduce waiting times in spite of the increase in arrival rates. Table A. 2 presents the estimated expected queue-lengths under dedicated staffing and the heuristic policy (which uses the non-inflated arrival rates) for the calibrated experiment with inflated arrivals. Interestingly, we find that the heuristic is still able to perform well and not only maintain stability in the infinite-horizon experiments, but also balance and reduce queue-lengths across the areas. In particular, for 12-hour shifts lengths, the relative queue reductions are comparable to those reported in Section 5.4.

	$n = 52$		$n = 60$	
Area	Finite-horizon	Infinite-horizon	Finite-horizon	Infinite-horizon
A	190%	(unstable)	57%	161%
B	127%	259%	37%	164%
C	198%	(unstable)	35%	163%
U	58%	36%	32%	36%

Table A. 1 Percentage increase in expected waiting time under dedicated staffing and with inflated arrivals.

	$n = 52$				$n = 60$			
	Finite-horizon		Infinite-horizon		Finite-horizon		Infinite-horizon	
Area	Dedicated	Heuristic	Dedicated	Heuristic	Dedicated	Heuristic	Dedicated	Heuristic
A	94.48	18.29	(unstable)	4.35	1.43	0.95	11.18	6.25
B	9.46	16.85	4.06	4.46	1.22	1.02	3.37	4.68
C	71.33	17.61	(unstable)	4.35	1.37	0.94	9.59	5.93
U	5.74	18.09	3.65	4.52	3.57	0.99	5.41	5.55
Total	181.01	70.84	(unstable)	17.68	7.59	3.9	29.55	22.41

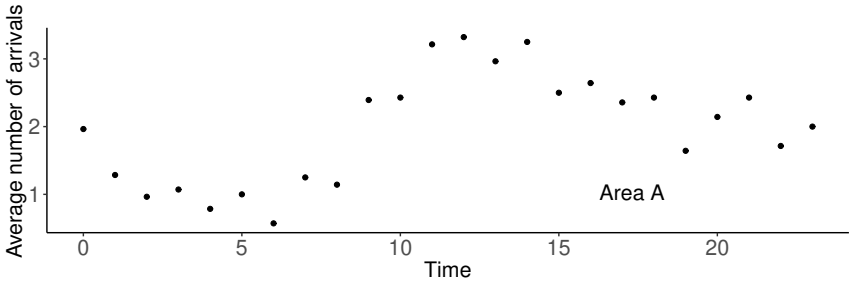
Table A. 2 Estimated expected queue-lengths in the calibrated experiment with inflated arrival rates.

A.4. Supplementary Figures and Tables

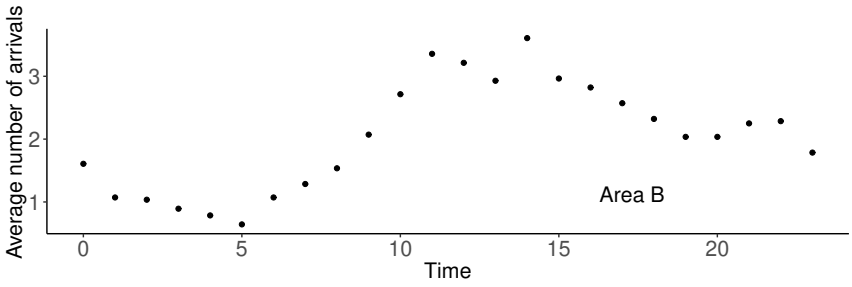
Table A. 3 Summary Statistics of Patient Covariates at WCMC

	OFF Period ($N = 2,413$)	ON Period ($N = 2,644$)	Total ($N = 5,057$)	p-value
Age	53.47 (19.79)	54.93 (20.01)	54.23 (19.92)	0.010
Female	54.7%	55.5%	55.1%	0.577
Race/Ethnicity				
White	46.7%	46.8%	46.8%	0.978
Black	16.5%	17.7%	17.1%	0.225
Asian	4.5%	4.8%	4.7%	0.538
Pacific Islander	1.2%	1.0%	1.1%	0.455
Hispanic/Latino	13.7%	14.0%	13.8%	0.773
Other	15.7.7%	16.3%	16.0%	0.565
Unknown	15.4%	13.3%	14.3%	0.036
Insurance				0.004
Private	40.7%	36.6%	38.6%	
Medicaid	23.1%	22.3%	22.7%	
Medicare	30.2%	34.5%	32.4%	
Self-pay	6.0%	6.6%	6.3%	
ESI	3.098 (0.65)	3.07 (0.64)	3.08 (0.64)	0.089
Elixhauser Score	2.64 (3.14)	2.86 (3.32)	2.75 (3.24)	0.015
Disposition				0.478
Discharged	65.9%	66.5%	66.2%	
Admitted	31.0%	29.8%	30.4%	
Other	3.2%	3.7%	3.4%	

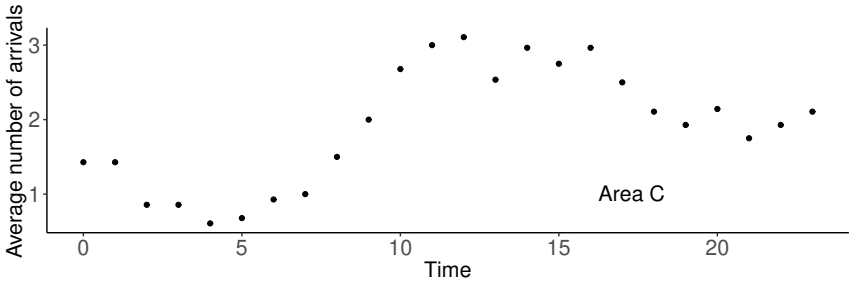
Note. Mean and standard deviation (in parentheses) shown for continuous variables. p-values provided for testing differences between the mean (or distribution for categorical variables) in the Off versus On periods.



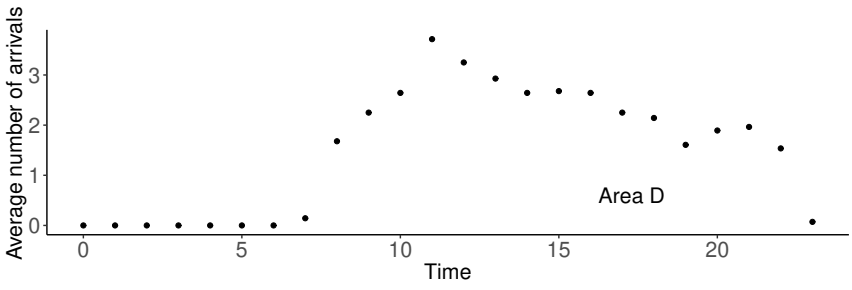
(a) Area A



(b) Area B

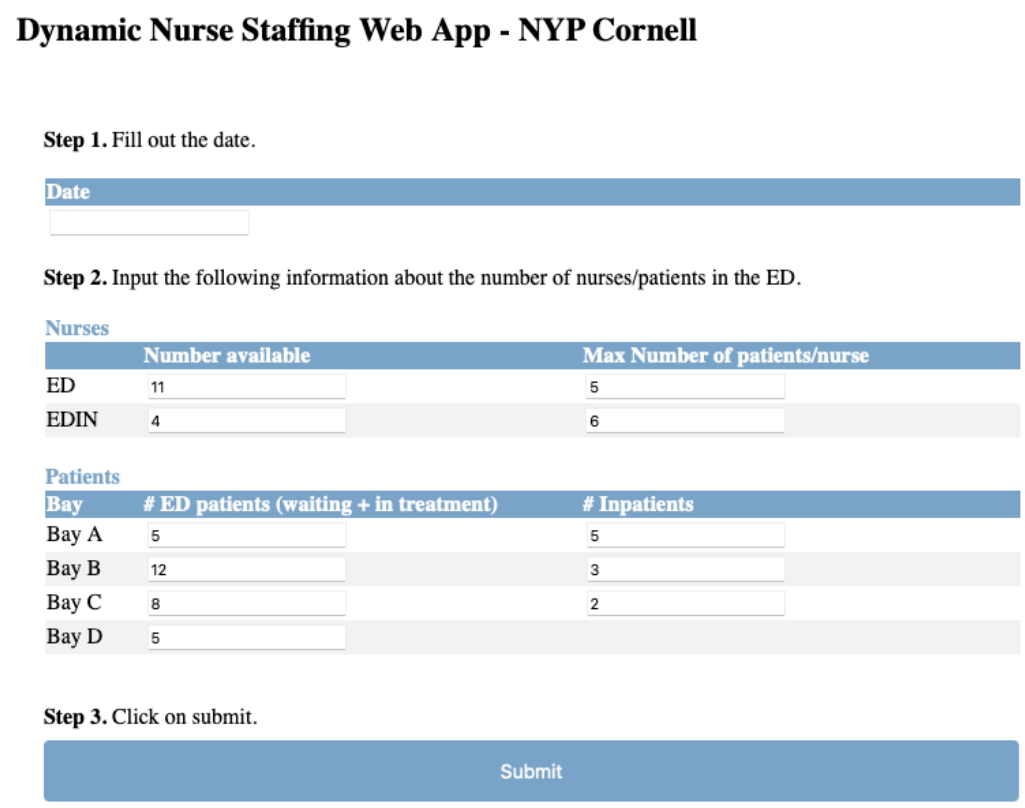


(c) Area C

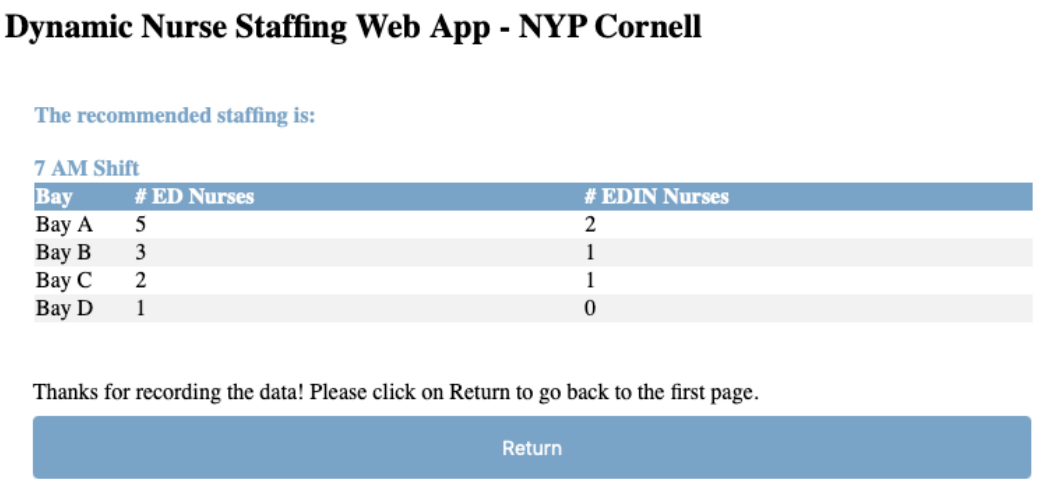


(d) Area U

Figure A. 5 Average hourly arrival rates at the four areas of WCMC ED over 4 consecutive weeks.



(a) Inputting census



(b) Inputting the staffing used

Figure A. 6 **Screenshots of the Web App used in the Pilot study - inputting the census and generating a recommendation. The Web App is available at http://nursestaffing.pythonanywhere.com.**

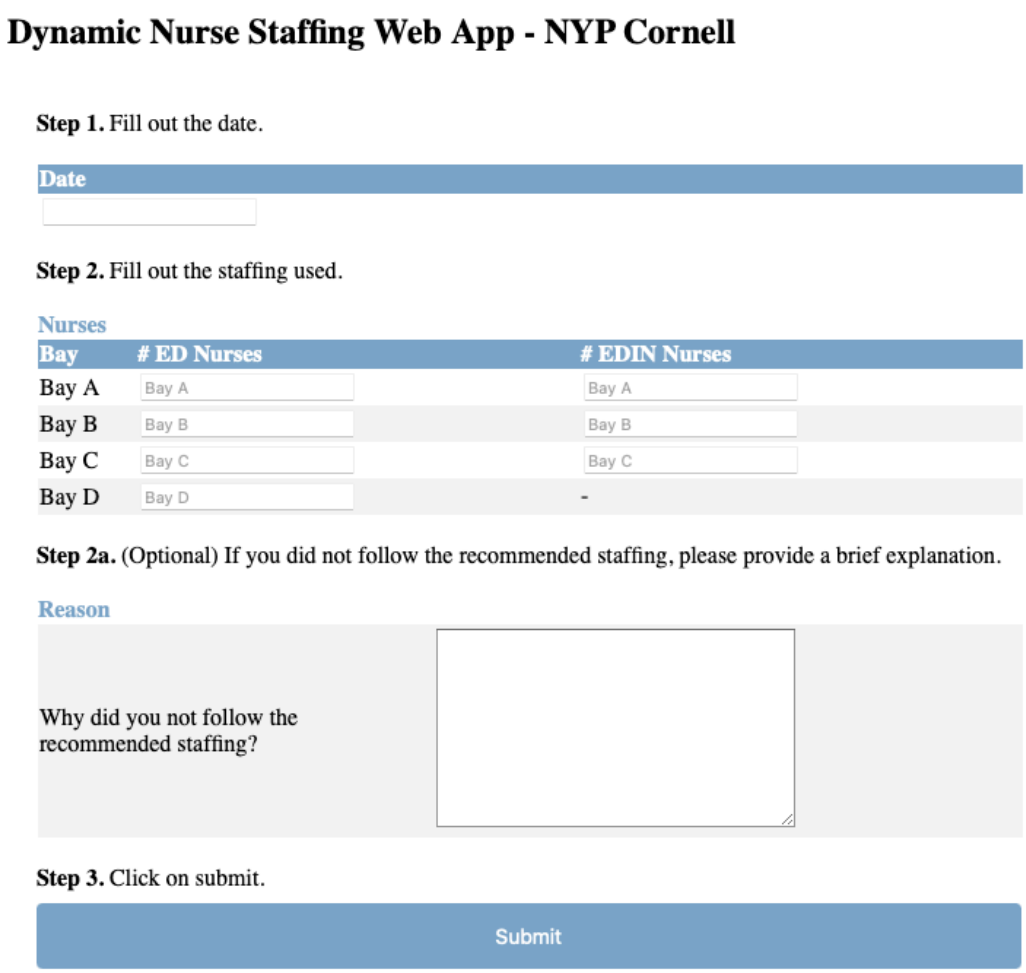


Figure A. 7 **Screenshots of the Web App used in the Pilot study - inputting the staffing used and the reasons for deviation from the recommendation. The Web App is available at http://nursestaffing.pythonanywhere.com.**

Table A. 4 Pre and Post Implementation Full Results

	Waiting Time (min)	Prov Eval to decision (hrs)	Boarding Time (hrs)	Prov Eval to discharge (hrs)	Total LOS (hrs)
ON (θ)	4.03***	0.16	-3.96***	-1.73***	-1.74***
95% CI	[1.86-6.21]	[-0.09,0.41]	[-5.61,-2.31]	[-2.35,-1.12]	[-2.34,-1.14]
intercept	5.92	9.09***	16.30**	21.60***	20.98***
95% CI	[-7.40,19.23]	[7.58,10.60]	[5.80,26.81]	[17.84,25.36]	[17.32,24.64]
Age	0.05	0.00	-0.05	0.01	0.01
95% CI	[-0.02,0.13]	[-0.00,0.01]	[-0.11,0.02]	[-0.01,0.03]	[-0.01,0.03]
Male	-0.68	0.00	-0.05	0.01	0.01
95% CI	[-2.73,1.37]	[-0.23,0.23]	[-2.82,0.30]	[-0.03,0.86]	[-0.33,0.80]
ESI	-7.69***	-1.51***	2.05*	-3.72***	-3.75***
95% CI	[-9.33,-6.04]	[-1.70,-1.32]	[0.48,3.61]	[-4.18,-3.25]	[-4.20,-3.29]
Elixhauser	0.67**	0.03	0.68***	0.99***	0.98***
95% CI	[0.29,1.05]	[-0.02,0.07]	[0.44,0.92]	[0.89, 1.10]	[0.88,1.09]
White	3.31*	0.29	0.02	0.49	0.44
95% CI	[0.14,6.47]	[-0.07,0.65]	[-2.64,2.69]	[-0.41, 1.38]	[-0.43,1.31]
Black	4.92*	0.50*	2.53	1.13*	0.92
95% CI	[1.14,8.70]	[0.08,0.93]	[-0.67,5.73]	[0.06, 2.20]	[-0.11, 1.96]
Asian	6.28*	-0.10	1.67	1.32	1.35
95% CI	[0.88,11.68]	[-0.72,0.51]	[-2.55-5.89]	[-0.20,2.85]	[-0.14,2.84]
Pacific Islander	-2.69	-0.69	-7.00	-0.78	-0.35
95% CI	[-12.50,7.12]	[-1.80,0.42]	[-15.30,1.31]	[-3.55,2.00]	[-3.04,2.35]
Other Race	2.81	0.46*	2.92	1.29*	1.11
95% CI	[-1.23,6.85]	[0.01,0.92]	[-0.59,6.44]	[0.15, 2.44]	[-0.00,2.22]
Hispanic/Latino	-0.75	-0.12	-0.74	-0.52	-0.48
95% CI	[-4.16,2.66]	[-0.51,0.26]	[-3.62,2.14]	[-1.48, 0.45]	[-1.41,0.46]
Medicaid	4.78**	0.53**	2.15	0.64	0.63
95% CI	[1.97,7.58]	[0.21,0.85]	[-0.45,4.74]	[-0.15,1.44]	[-0.15,1.40]
Medicare	0.46	0.19	2.50	2.04***	2.11***
95% CI	[-2.92,3.85]	[-0.19,0.57]	[-0.07,5.07]	[1.09, 3.00]	[1.18,3.04]
Self-pay	2.32	0.15	4.72	-0.06	-0.32
95% CI	[-2.26,6.90]	[-0.36,0.66]	[-1.39,10.83]	[-1.35,1.23]	[-1.54,0.90]
PM Shift	3.59**	0.90***	0.08	0.16	-0.65*
95% CI	[1.39-5.79]	[0.65-1.15]	[-1.67,1.84]	[-0.46,0.78]	[-1.26, -0.05]
Nurses Available	-0.98**	-0.10**	-0.86**	-0.44***	-0.36***
95% CI	[-1.63,-0.32]	[-0.17,-0.03]	[-1.36,-0.36]	[-0.62, -0.25]	[-0.54,-0.18]
ED Census	0.71***	0.01*	0.08***	0.02**	0.03***
95% CI	[0.66,0.76]	[0.00,0.01]	[0.04,0.12]	[0.01,0.04]	[0.01,0.04]
Num. of obs.	4758	4753	1286	4758	4817
R^2	0.177	0.082	0.081	0.211	0.220

Note. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A. 5 Pre and Post Implementation Results for Returns within 10 days

	Linear Probability Model	Logit Model	Probit Model
ON (θ)	0.01	1.18	1.08
95% CI	[-0.00,0.03]	[0.94,1.49]	[0.96,1.22]
intercept	-0.06	0.01^{***}	0.08^{***}
95% CI	[-0.16,0.04]	[0.00,0.04]	[0.04,0.17]
Age	0.00	1.00	1.00
95% CI	[-0.00,0.00]	[0.99,1.01]	[1.00,1.01]
Male	0.02^{*}	1.29^{*}	1.14^{*}
95% CI	[0.00,0.03]	[1.04,1.61]	[1.02,1.27]
ESI	0.02^{**}	1.27^{**}	1.12^{**}
95% CI	[0.01,0.03]	[1.06,1.52]	[1.03,1.23]
Elixhauser	0.01^{***}	1.15^{***}	1.07^{***}
95% CI	[0.01,0.02]	[1.11, 1.19]	[1.05, 1.09]
White	0.06^{***}	3.17^{***}	1.70^{***}
95% CI	[0.03,0.08]	[2.00,5.32]	[1.38,2.12]
Black	0.04^{**}	2.55^{***}	1.54^{***}
95% CI	[0.01,0.07]	[1.54,4.41]	[1.22,1.96]
Asian	0.04^{*}	2.65^{**}	1.59^{**}
95% CI	[0.00,0.08]	[1.31,5.29]	[1.15,2.18]
Pacific Islander	0.09^{*}	4.42^{**}	2.06^{**}
95% CI	[0.02,0.16]	[1.60,11.06]	[1.25,3.29]
Other Race	0.04^{**}	2.70^{**}	1.57^{***}
95% CI	[0.01,0.07]	[1.57,4.83]	[1.22,2.04]
Hispanic/Latino	-0.03^{**}	0.61^{*}	0.79^{*}
95% CI	[-0.06,-0.01]	[0.41,0.90]	[0.65,0.96]
Medicaid	0.05^{***}	1.91^{***}	1.38^{***}
95% CI	[0.03,0.07]	[1.42,2.57]	[1.19,1.60]
Medicare	0.02	1.34	1.12
95% CI	[-0.01,0.04]	[0.94,1.93]	[0.93,1.34]
Self-pay	0.01	1.18	1.08
95% CI	[-0.03,0.04]	[0.70,1.92]	[0.84,1.37]
PM Shift	0.03^{**}	1.38^{**}	1.18^{**}
95% CI	[0.01,0.04]	[1.11,1.73]	[1.05,1.32]
Nurses Available	-0.00	0.98	0.99
95% CI	[-0.01,0.00]	[0.91,1.05]	[0.95,1.02]
ED Census	0.00	1.00	1.00
95% CI	[-0.00,0.00]	[1.00,1.01]	[1.00,1.00]
Admitted	-0.09^{***}	0.24^{***}	0.51^{***}
95% CI	[-0.11,-0.07]	[0.17,0.34]	[0.43,0.60]
Other disposition	0.06^{**}	1.69^{*}	1.36^{*}
95% CI	[0.01,0.10]	[1.03,2.67]	[1.05,1.74]
Num. of obs.	4817	4817	4817
R^2	0.049	0.065	0.106

Note. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Instead of coefficient estimate, Odds Ratio reported for Logistic Model and Risk Ratio reported for Probit Model.

Table A. 6 Pre and Post Implementation Results for Walkouts

	Linear Probability Model	Logit Model	Probit Model
ON (θ)	0.00	1.05	1.03
95% CI	[-0.01,0.01]	[0.74,1.50]	[0.88,1.21]
intercept	0.00	0.05**	0.21**
95% CI	[-0.02,0.11]	[0.01,0.38]	[0.08,0.53]
Age	0.00	1.01	1.00
95% CI	[-0.00,0.00]	[0.99,1.02]	[1.00,1.01]
Male	0.00	1.02	1.02
95% CI	[-0.01,0.01]	[0.73,1.42]	[0.88,1.18]
ESI	-0.01**	0.67**	0.81**
95% CI	[-0.02,-0.00]	[0.51,0.88]	[0.72,0.92]
Elixhauser	0.00	1.01	1.00
95% CI	[-0.00,0.00]	[0.95,1.08]	[0.97,1.03]
White	0.00	1.14	1.05
95% CI	[-0.01,0.02]	[0.70,1.92]	[0.84,1.33]
Black	0.02	1.55	1.21
95% CI	[-0.00,0.04]	[0.90,2.72]	[0.94,1.57]
Asian	-0.01	0.43	0.69
95% CI	[-0.04,0.01]	[0.10,1.29]	[0.39, 1.11]
Pacific Islander	-0.04	0.43	0.70
95% CI	[-0.09,0.01]	[0.07,1.62]	[0.33,1.32]
Other Race	-0.01	0.60	0.79
95% CI	[-0.03,0.01]	[0.30,1.19]	[0.58,1.07]
Hispanic/Latino	0.01	1.54	1.20
95% CI	[-0.00,0.03]	[0.89,2.55]	[0.94,1.51]
Medicaid	0.01	1.53	1.21
95% CI	[-0.00,0.03]	[0.96,2.44]	[0.99,1.49]
Medicare	-0.02*	0.52*	0.79
95% CI	[-0.03,-0.00]	[0.28,0.69]	[0.61,1.02]
Self-pay	0.15***	9.37***	2.94***
95% CI	[0.13,0.17]	[6.03,14.64]	[2.37,3.66]
PM Shift	0.01	1.28	1.09
95% CI	[-0.00,0.02]	[0.91,1.79]	[0.94,1.27]
Nurses Available	-0.00	0.98	1.00
95% CI	[-0.00,0.00]	[0.99,1.01]	[1.00,1.00]
ED Census	0.00	1.00	1.00
95% CI	[-0.00,0.00]	[0.99,1.01]	[1.00,1.00]
Num. of obs.	4817	4817	4817
R^2	0.048	0.054	0.117

Note. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Instead of coefficient estimate, Odds Ratio reported for Logit Model and Risk Ratio reported for Probit Model.

A.5. Sensitivity Analysis

A core assumption of our analysis is unconfoundedness. We now explore how sensitive our results are to this assumption. In our analysis, we assume patients are exogenously assigned to treatment (ON) or control (OFF). This seems reasonable as patients (and even physicians and nurses, except for those in leadership positions who signed off on the pilot program) were unaware of our pilot, so could not have made decisions about visiting the ED based on whether it was during the ON or OFF period. That said, if patients are systematically less sick in the ON period, this could violate our assumption and bias our results. In particular, this would make it seem like the intervention reduced time spent in the ED, when the reductions were actually because less sick patients did not require as much time in the ED. In addition, other unobserved confounders, such as changes in physician staffing or how the staff responded to higher workload during the ON period, could also potentially bias our results. By our summary statistics in Table A. 3, patients did not appear to be less sick; if anything they appeared to be more sick during the ON period (see, e.g., the Elixhauser score). Further, we are not aware of any systematic changes in staffing or other operational policies in the ED during the ON period. However, we cannot completely rule out the possibility that unobserved confounders were present during the study period.

To examine the sensitivity of our results to unobserved variables, we conduct a sensitivity analysis using the framework of Carnegie et al. (2016) which builds on the approach proposed in Imbens (2003). The idea is to augment the regression model with a hypothetical unobserved variable and vary its correlation with both the intervention and outcome variables. It is natural to think of this variable as a linear combination of all unmeasured confounders. Re-estimating the coefficient of the treatment variable for various correlations, one can examine how large of an effect size the unobserved variable needs to have in order to invalidate the results. That is, if the initial coefficient estimate for our treatment (θ) is negative and statistically significant at the $p < 0.05$ level, the sensitivity analysis determines how large an effect size the unobserved variable need to have, in order to make the estimate for θ statistically indistinguishable from 0. One can also consider how large the effect size needs to be for θ to be positive and statistically significant at the $p < 0.05$ level.

Figure A. 8 summarizes the sensitivity analysis for our outcome models. The x-axis is the coefficient specifying the effect size of the unobservable variable on the likelihood of treatment (being admitted during the ON period). The y-axis is the coefficient specifying the effect size of the unobservable variable on the outcome (e.g., ED LOS). The origin $(0, 0)$ corresponds to our baseline analysis, which assumes there is no unobserved confounder. Each line corresponds to a contour (magnitude and statistical significance) of the coefficient estimate of the treatment effect of our nurse staffing intervention. The blue line indicates when the estimated treatment effect is 0. The space between the black 'N.S.' lines indicate where the estimated treatment effect is statistically no different than 0 at the $p < 0.05$ level. Anything to the upper-right or upper-left of these lines indicates when the estimated treatment effect would be the *opposite sign* of the baseline model. The inverted blue triangles represent the coefficient estimates for the observable variables which are negative for the outcome. The red crosses represent the coefficient estimates for the observable variables which are positive for the outcome. This calibrates one to understand how large the coefficient of the unobservable confounder must be relative to those of the observable variables.

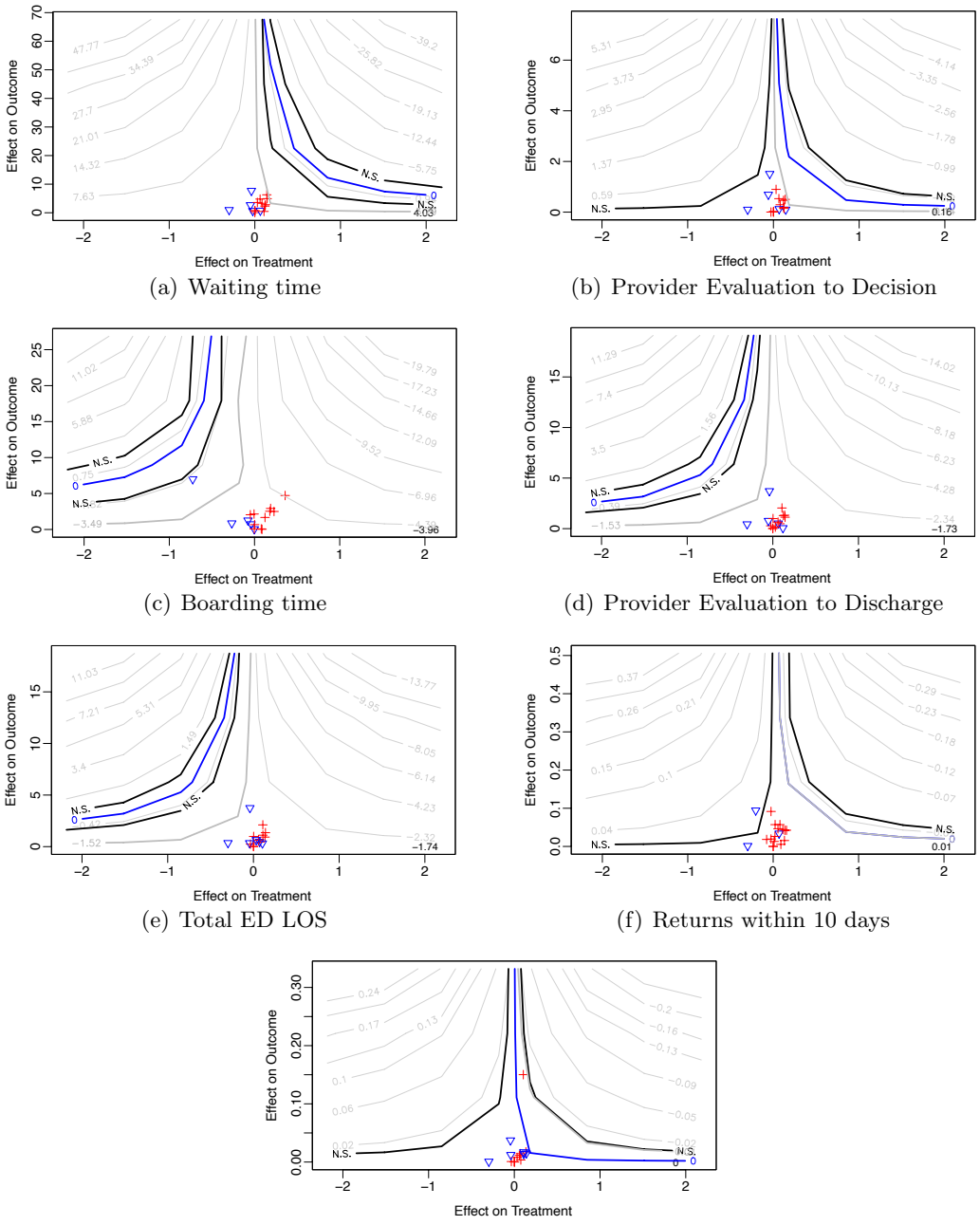
We can see that the estimates presented in Table 3 are quite robust across the 5 continuous outcomes. For instance, for the estimated impact of the treatment on total ED LOS to be insignificant, the unobserved variable must have had a much larger impact on *both* the treatment decision and the ED LOS, compared to those of observable variables. The boarding time model (Figure 8(c)) has one variable that is close to the N.S. line. Nevertheless, the unobservable variable still needs to have a relatively large effect on treatment and outcome in order to make the treatment effect statistically indistinguishable from 0. Therefore, while the sensitivity analysis cannot completely rule out the possibility of an unobservable variable invalidating our results for the continuous outcome models, it suggests that if such a factor exists, it needs to have substantially more explanatory power than the observable characteristics.

We also observe that the estimated effect on walkouts appears to be robust. The red cross close to the N.S. line in Figure 8(g) indicates that, in case a confounder exists, the intervention might have even reduced the probability of walkouts. On the other hand, there are a couple of covariates that are very close, or even crossing the N.S. line for returns to the ED. This suggests that the risk of an unobserved confounder resulting in a positive and statistically significant result is high. That our main results could be considered marginally significant (e.g. at the $p < 0.15$ level), suggests that the results for the returns to the ED within 10 days are not very robust. Thus, while it seems we have compelling evidence that the other measures are quite robust, we cannot say the same for returns.

A.6. Robustness Checks

To further understand the robustness of our results, we conduct a number of additional regressions with alternative control variables. Recall that our main model controlled for age, gender, race, ethnicity, insurance, ESI, Elixhauser, the shift of arrival, the number of nurses working during the shift, and the number of patients in the ED. More specifically, in addition to the covariates considered in our main model, we conduct additional analysis controlling for: 1) discharge disposition; 2) average ED census during the ED visit; 3) the area the patient was assigned; 4) the number of patients waiting, the number of patients in treatment, and the number of patients boarding instead of ED census; and, 5) the number of patients waiting in the assigned area, the number of patients in treatment in the assigned area, and the number of patients boarding in the assigned area instead of ED census.

Table A. 7 summarizes these robustness checks. We can see that, with the exception of waiting time, the results for the continuous measures are very robust, with respect to both magnitude and statistical significance. Thus, there seems to be reasonably robust evidence that the new approach to nurse staffing is associated with a 1 to 3 hour reduction in 'service time', as measured by the time from provider evaluation to discharge, boarding time, and total ED LOS. For waiting time, when replacing ED census with a more fine grained definition of congestion, the association appears to decrease and in one case no longer statistically significant. As mentioned earlier, the average arrival rate of patients appeared to be significantly larger during the ON period compared to the OFF period, and so it is likely that the treatment effect actually captured some of the increased queueing effects due to the increase load. While controlling for more granular measures of census reduces the estimated increase, the nonlinear impact of increased congestion is still not captured in the models.



(a) Waiting time

(b) Provider Evaluation to Decision

(c) Boarding time

(d) Provider Evaluation to Discharge

(e) Total ED LOS

(f) Returns within 10 days

(g) Walkouts

Figure A. 8 **Sensitivity analysis following the approach of Carnegie et al. (2016). Contours indicate where the estimated treatment effect is equal to 0, or statistically indistinguishable from 0 (N.S.) at the** $p < 0.05$ **level.**

We also see that the results for returns within 10 days is robust to model specification, with the estimated coefficient being equal to 0.01 in all cases, but not statistically significant at the $p < 0.05$ level. We observe consistent results for the probit and logistic models, with the p-values being larger than 0.15 for all models. Similarly, the results for walkouts are robust with the estimates varying between 0.01 and 0.05, but remaining statistically insignificant across all models.

Table A. 7 Robustness Checks of Pre and Post Implementation

	(1) discharge disposition	(2) Avg ED census	(3) assigned area	(4) waiting, in treatment, boarding	(5) waiting, in treatment, boarding in area
Waiting Time (min)					
ON (θ)	3.88***	4.44***	4.26***	1.58	2.17*
95% CI	[1.70,6.06]	[2.16,6.73]	[2.24,6.28]	[-0.60,3.76]	[0.25,4.09]
Num. of obs.	4758	4758	4756	4758	4756
R^2	0.179	0.094	0.291	0.204	0.369
Prov Eval to dec. (hrs)					
ON (θ)	0.13	0.16	0.18	0.17	0.22
95% CI	[-0.12,0.38]	[-0.08,0.41]	[-0.05,0.42]	[-0.08,0.43]	[-0.03,0.46]
Num. of obs.	4753	4753	4751	4753	4751
R^2	0.088	0.081	0.149	0.082	0.126
Boarding Time (hrs)					
ON (θ)	-3.96***	-4.19***	-3.99***	-3.47***	-3.97***
95% CI	[-5.61,-2.31]	[-5.85,-2.53]	[-5.64, -2.34]	[-5.14,-1.80]	[-5.62,-2.32]
Num. of obs.	1286	1286	1286	1286	1286
R^2	0.081	0.072	0.084	0.091	0.085
Prov Eval to dis. (hrs)					
ON (θ)	-1.08***	-1.75***	-1.70***	-1.68***	-1.67***
95% CI	[-1.61,-0.54]	[-2.37,-1.14]	[-2.31,-1.09]	[-2.30,-1.05]	[-2.28,-1.05]
Num. of obs.	4758	4758	4756	4758	4756
R^2	0.412	0.211	0.228	0.212	0.226
Total LOS (hrs)					
ON (θ)	-1.11***	-1.76***	-1.70***	-1.72***	-1.67***
95% CI	[-1.62,-0.59]	[-2.36,-1.16]	[-2.29,-1.10]	[-2.33,-1.11]	[-2.27,-1.07]
Num. of obs.	4817	4817	4807	4817	4807
R^2	0.431	0.219	0.241	0.220	0.237
Returns within 10 days					
ON (θ)	0.01	0.01	0.01	0.01	0.01
95% CI	[-0.00 , 0.03]	[-0.00 , 0.03]	[-0.00 ,0.03]	[-0.00 , 0.03]	[-0.00 , 0.03]
Num. of obs.	4817	4817	4807	4817	4807
R^2	0.049	0.049	0.049	0.049	0.049
Walkouts					
ON (θ)	-	0.03	0.02	0.05	0.01
95% CI	-	[-0.04 , 0.09]	[-0.04 ,0.09]	[-0.02 , 0.11]	[-0.05 , 0.08]
Num. of obs.	-	4817	4807	4817	4807
R^2	-	0.048	0.045	0.048	0.044

Note. All models include controls for age, gender, race, ethnicity, insurance, ESI, Elixhauser, shift, census, and number of nurses. In contrast to the other outcomes models, column (1) for Returns within 10 days *excludes* discharge disposition as the main model controlled for this. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.