Noname manuscript No.
(will be inserted by the editor)

Queueing Theoretic Analysis of Labor and Delivery

Understanding Management Styles and C-Section Rates

Matthew Gombolay · Toni Golen · Neel Shah · Julie Shah

Received: 14 February 2017 / Accepted: 25 August 2017

Abstract Childbirth is a complex clinical service requiring the coordinated support of highly trained healthcare professionals as well as management of a finite set of critical resources (such as staff and beds) to provide safe care. The mode of delivery (vaginal delivery or cesarean section) has a significant effect on labor and delivery resource needs. Further, resource management decisions may impact the amount of time a physician or nurse is able to spend with any given patient. In this work, we employ queueing theory to model one year of transactional patient information at a tertiary care center in Boston, Massachusetts. First, we observe that the M/G/∞ model effectively predicts patient flow in an obstetrics department. This model captures the dynamics of labor and delivery where patients arrive randomly during the day, the duration of their stay is based on their individual acuity, and their labor progresses at some rate irrespective of whether they are given a bed. Second, using our queueing theoretic model, we show that reducing the rate of cesarean section – a current quality improvement goal in American obstetrics – may have important consequences with regard to the resource needs of a hospital. We also estimate the potential financial impact of these resource needs from the hospital perspective. Third, we report that application of our model to an analysis of potential patient coverage strategies supports the adoption of team-based care, in which attending physicians share responsibilities for patients.

Keywords Obstetrics · Labor and Delivery · Hospital Management · Queueing Theory · Hypercube Model · C-Section Rate · Healthcare Cost

1 Introduction

Healthcare systems are challenged by the need to provide high-quality healthcare to a growing population with finite resources [3,4,31]. The resource management challenges of the labor and delivery (L & D) floor, where 99% of American babies are born, are uniquely complex: the floor must be staffed and equipped for triage, emergency surgery, and close surveillance of labor progress, as well as standard inpatient care for both adults and newborn infants.

A current concern in American obstetrics is the high rate of cesarean delivery (C-section) and its effect on hospital resources, such as rooms and staff. While spontaneous vaginal births require protracted periods of clinical attention prior to delivery, C-sections are more expeditious. On the other hand, C-sections require a significantly longer hospital stay. In 1965, the national C-section rate was 4.5% [52]; by 2009, this rate skyrocketed to 32% [10]. The magnitude of this shift is not well-explained by shifts in patients' risks or preferences, nor by medical malpractice or professional reimbursement [15,14,17,55]. Regardless, there is wide agreement that current C-section rates are too high by a large margin [43]. Overuse of C-sections may increase the risk

This work was supported by the National Science Foundation Graduate Research Fellowship Program under grant number 2388357.

M. Gombolay
Georgia Institute of Technology
North Ave NW, Atlanta, GA 30332, USA
E-mail: gombolay@mit.edu

T. Golen
Beth Israel Deaconess Medical Center
330 Brookline Avenue, Boston, MA 02215, U.S.A.
E-mail: tgolen@bidmc.harvard.edu

N. Shah
Beth Israel Deaconess Medical Center
330 Brookline Avenue, Boston, MA 02215, U.S.A.
E-mail: ntshah@bidmc.harvard.edu

J. Shah
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, MA 02139, U.S.A.
E-mail: gombolay@mit.edu

of surgical complications and other adverse events for both mothers and infants [1,9,16,39].

Reducing the incidence of avoidable C-section in the United States could improve the safety, cost, and experience of care for millions of mothers and newborn infants annually [54]. However, doing so is likely to require a shift in the resource composition of the hospital (e.g., the number of beds and staff members required in each ward within the obstetrics unit). A better understanding of how to optimally reallocate hospital resources can help ensure that patients receive safe and appropriate care. Prior work has incorporated queuing theoretic models to investigate resource use on labor and delivery units at steady state but does not account for differences in patient flow over the course of a day [53].

A related concern for optimizing staff resources is the policy attending physicians maintain with regard to sharing the responsibility of patient care. It is a common practice in many tertiary care centers (i.e., large hospitals staffed by specialists) for each attending physician to be the primary provider for patients assigned to his or her care. If this primary provider is occupied with one patient, other staff (residents or staff nurses) may temporarily assume care for the primary provider's other patients, but will typically defer consequential decisions until the primary provider is available again. However, in some hospitals, physicians share their responsibilities, and the assignment of patients to physicians is more flexible and dynamic. Prior work within the obstetrics community has not used a queueing theory-based approach to assess the merits of these management styles for a given obstetric unit's needs.

We provide three novel contributions: First, we confirm the applicability of the M/G/∞ queueing model to study operations in labor and delivery, not just at a specific time when the number of patients is at a steady state, as in the work by Takagi et al. [53], but for patient flow throughout an entire 24-hour day. Second, we employ the M/G/∞ model to quantify how bed utilization would change if the C-section rate were reduced. We find that hospitals would need to substantially increase the number of beds on the labor and delivery floor, while decreasing the number of beds in the postpartum ward. Applying data from prior work by Shah et al. [50], we estimate the impact that changing the C-section rate would have on operating costs for the obstetrics department. We find that decreasing the C-section rate may, in fact, not increase overall operating costs. Third, using a novel application of the hypercube queueing model [34], we quantitatively demonstrate the benefits of the two aforementioned staffing strategies (i.e., whether or not attending physicians share responsibility for laboring patients) using the M/G/∞ model. We measure the amount of increased time that patients receive care from an attending physician when attending physicians share the responsibility of patient care.

The paper is structured as follows: First, we briefly review important work related to the modeling and optimization of healthcare processes (Section 2). Next, we provide an overview of labor and delivery operations (Section 3), as well as a description of the data set we obtained for our analysis (Section 4). In Section 5, we confirm the applicability of the M/G/∞ queueing model and discuss the nuances that necessitate its use over the M/M/m model. We then address how resource needs would change as a function of a change to C-section rate (Section 6). In Section 7, we present our investigation into the relative merits of two common management styles for patients in obstetrics. Finally, we present our conclusion in Section 8.

2 Background

Healthcare operations have received much attention from researchers attempting to improve the efficiency and quality of hospital care [49,27,19,46]. In this section, we briefly review related work in applied statistical modeling and discrete event simulation (DES) focused on modeling and improving hospital operations. Next, we discuss related work from the complementary perspective of queueing theory. We conclude the section by outlining the novel contributions of our work relative to these prior studies.

Applied statistical modeling and DES are ubiquitous techniques for understanding and improving healthcare operations [2,6,12,13,18,23,26,38,40,48,49,51,59]. For example, Hall et al. [26] presented modeling tools (e.g., process maps and task analysis) for understanding a healthcare system, measuring that system's performance, and resolving delays in interfaces between units in hospitals. In their paper, Hall et al. [26] presented a case study of a Los Angeles County/University of Southern California Hospital in which they demonstrated the use of their tools. Hall et al. [26] showed that the studied hospital should increase the size of the ward responsible for discharging patients, optimize the assignment of personnel to logistical tasks (i.e., transporting patients), improve scheduling and forecasting of non-urgent procedures, and implement bed and personnel tracking systems to reduce operator workload.

Marmor [40] first developed a simulation-based operation to identify bottlenecks and improve the studied hospital's operations. Marmor [40] also considered staff scheduling, as poor scheduling often limits the ability of the hospital to operate efficiently. Zeltyn et al. [59] used a simulation-based technique that employs the concept of "offered load" to understand staffing problems arising in hospital operations. In essence, offered load accounts for the time required by a provider to care for an individual patient, as well as a correction factor to address inefficiencies resulting from high workload [5,25]. Zeltyn et al. [59] showed that incorporating offered load into DES improved the ability of the simulation to provide real-time control of emergency department (ED) operations.

Armony et al. [2] performed an exploratory data analysis to help answer questions such as whether simple queueing models adequately capture hospital operations and how established patient flow processes affect delays. Their results indicated that hospital events (e.g., patient arrival rates) can be modeled with relatively simple probability distributions, and provided a set of challenges for the research community to develop more effective queueing theoretic models in order to understand and improve hospital resource management. Day et al. [12] developed a DES to predict whether the addition of an additional triage nurse would decrease the proportion of patients who remain in triage for longer than 6 hours. Based upon a positive result in simulation, the authors implemented this change at their hospital and observed a similar positive result in practice, thus demonstrating the power of DES.

De Bruin et al. [13] investigated how inpatient bed availability affects admission rates among cardiac patients; they found that limited bed availability increased the rate at which cardiac patients were turned away at the point of entry. Similarly, Litvak et al. [38] studied how scheduling of non-emergent surgical operations affected patient flow in the ED. They found that variation in the utilization of surgery resources – which is partially controllable by the hospital resource managers – contributed directly to delays in ED operations, and concluded that scheduled surgeries should be better balanced between days to decrease the variation and magnitude of ED delays. Shi et al. [51] investigated transfers from the ED to inpatient wards and developed a data-driven model to provide insight for managerial decisions.

With a focus on labor and delivery, Cochran and Bharti [11] developed a DES to better understand how to allocate beds across various care centers (e.g., whether a bed should be designated for triage versus post-operative recovery) within the obstetrics department. They found that increasing the number of beds in their department by 15% would increase the number of patients the hospital could care for by 38%. The authors noted that they have implemented the results of the study by increasing the number of beds in their department [11]. Ferraro et al. [18] developed a DES to aid in capacity planning for maternal/fetal medicine. They found that the addition of three beds to their existing center (the Children's Hospital of Philadelphia) nearly tripled the amount of time before the hospital would reach capacity and need to refuse admission.

Kwak and Lee [33] developed a multi-criteria decision-making model to determine the staffing needs at a healthcare organization in the midwestern United States. The authors used goal programming (a linear program with a multi-criteria objective function and associated constraints) to determine the optimal staffing levels for nurses, physicians, and technicians for each of six, 4-hour shifts over the course of a 24-hour day. Kwak and Lee emphasized that an added benefit of their investigation is the value it provided for hospital managers in terms of increased awareness of and insight into the multi-criteria goals and constraints inherent in hospital resource management.

While much of this prior work focused on discrete event models and descriptive statistical analysis, there have been important applications of theoretical modeling to the problem of improving healthcare resource management [23,24, 22,21,28,33,41,53,57,58]. For example, Yom-Tov and Mandelbaum [58] developed a model based on the Erlang distribution, called Erlang-R, to model patients who return multiple times during their need for hospital services. Yom-Tov and Mandelbaum [58] used this analysis to determine how many doctors and nurses were required to care for patients. Further, Huang [28] incorporated a day-of-the-week component into a queueing model to evaluate the need for emergency room beds and showed that occupancy on any given day follows a Poisson distribution.

McManus et al. [41] studied the problem of maintaining patient flow within an intensive care unit (ICU). The authors fit the M/M/m queue to their data, and showed that it was able to accurately capture the probability distributions governing bed occupancy. Based on their exploration of the fitted M/M/m queueing model and their experience in practice, McManus et al. posited a set of practical implications for the community. At the core of these implications is the need to consider entire probability distributions, rather than mere averages or point estimates. For example, many hospitals base their utilization estimates on nightly census data, which do not capture the transient impact of patients flowing in or out of the ICU [41].

Similarly, Green et al. [24] used the M/M/m queue to develop an understanding of the staffing hours required to maintain an acceptably low balking rate (i.e., the rate of patients who leave without being seen by a medical professional). Using the M/M/m queue, the authors showed that increasing the total provider hours by merely 3.1% would decrease the balk rate by 22.9%.

Using related techniques, Gerchak et al. [21] employed stochastic dynamic programming (DP) to improve the scheduling of elective surgeries given the uncertainty of emergent surgical needs, as well as limited surgical capacity. DP, as with many queueing theoretic models, defines the environment in question as a Markov decision process [44]. Gerchak et al. then developed interpretable bounds on the system's performance for various distributions and parameters defining the operating environment.

Some important works focus on obstetrics as well: Green and Nguyen [23] analyzed data from a hospital based in Boston (Beth Israel Deaconess Medical Center) to determine how well the patient discharge process and duration of patient stays at an obstetrics department could be modeled using existing queueing theory models. The authors found that the M/M/m queueing model can accurately predict the likelihood that a patient's service delay will be of a given duration, as a function of the number of beds in the obstetrics department and patient arrival rate.

Takagi et al. [53] applied M/G/∞ and M/M/m queues to represent the flow of patients within an obstetrics ward at the University of Tsukuba Hospital in Japan. The authors began by confirming the applicability of Little's Law of queueing theory [37] for patient flow in each ward within the department. Next, they tuned their queueing models to predict the probability distribution of the number of patients in each ward at the time of the nightly census. However, as noted by McManus et al. [41], simply modeling nightly census data fails to capture transient, flow-related stressors that often result in patient service denial.

In our work, we take a queueing theoretic perspective. While DES and related techniques can readily identify bottlenecks and performance limitations, they also require extensive, hospital-specific modeling. Development of queueing theoretic models allows for a more broadly applicable understanding, although these models often lack the detail required to more fully predict phenomena at a specific individual hospital. Our goal is to develop a broadly applicable understanding of how C-section rates and management styles affect obstetrics care; as such, we base our analysis on a queueing theoretic model (the M/G/∞ queue), which yields equations that can be easily tuned to answer questions about a specific hospital of interest.

We provide three novel contributions to management science vis-à-vis obstetrics: First, we demonstrate the validity of the M/G/∞ queue. While Takagi et al. [53] applied the same theoretical model as that we use in our work (the M/G/∞ queue), Takagi et al. only considered bed occupancy during the scheduled nightly census. In our investigation, we consider the viability of the M/G/∞ queue for modeling data throughout the entire day. Showing that the M/G/∞ queue can accurately capture behavior within the obstetrics department based on events occurring over the course of a day represents a novel advance because the department is in flux – not in a steady state. Specifically, patients are discharged at both random (i.e., patients discharged from triage after being deemed healthy) and non-random times (i.e., physicians typically discharge patients from postpartum in the afternoons, as discussed further in Section 5). Relative to the work by Green and Nguyen [23], we show that the more general M/G/∞ queue can accurately model processes in obstetrics.

Second, we use the validated M/G/∞ queue to model how a change in C-section rates would affect resource utilization within an obstetrics department. We are unaware of any prior application of a queueing theoretic model – including those by Green and Nguyen [23] and Takagi et al. [53] – to assess the impact of altering C-section rates on resource utilization.

Third, we investigate how patient management styles affect the time a physician spends caring for an actively laboring patient. Using the M/G/∞ queue with a hypercube modeling framework [34], we find that a paradigm in which physicians share the responsibility of caring for laboring patients increases the time those patients are cared for by an attending physician (as opposed to a resident) relative to a scenario in which physicians only care for their own, preallocated primary care patients. Again, we are unaware of any prior application of a queueing theoretic model to study these management styles in obstetrics care, or of any work translating the Larson model for the purpose of analyzing the queueing behavior of physicians caring for patients.

3 Labor and Delivery: An Overview

Resource management in labor and delivery (L & D) is complex, as patients may take multiple possible routes through various care centers before giving birth. Here, we first describe the various steps in the overall L & D care process. Next, we discuss how labor and delivery staff are assigned to patients.

Figure 1 depicts a simplified process map that describes how patients (pregnant women) move through the hospital to receive care. To describe this model, we consider three types of patients: scheduled inductions, scheduled cesarean sections, and unscheduled.

3.1 Scheduled Inductions

At the recommendation of her obstetrician/gynecologist (OB/GYN) or midwife, a patient may be scheduled for induction of labor. Typically, these are cases in which patients are not yet in labor but the risk of remaining pregnant is higher than the risk associated with delivery for either the mother or infant. These patients will commonly arrive at the labor and delivery floor and be directly admitted to a labor bed without first being seen in L & D triage. These patients generally experience the longest service times on the labor and delivery floor because the medicine required to induce labor takes longer to bring a woman to active labor than if the woman's body stimulates labor on its own. After delivery, induced patients are moved to the postpartum floor. Under extreme situations, when the labor and delivery floor may be too full to admit such patients, a scheduled induction may first be moved to the antepartum ward to begin the procedure, or the procedure may be postponed until a later date.

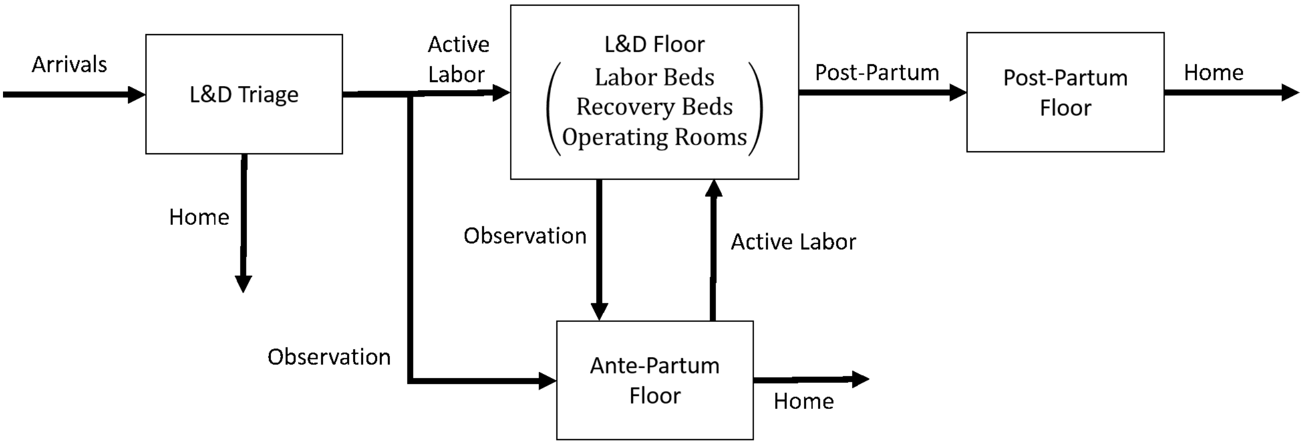

Fig. 1 A simplified process map for labor and delivery operations.

3.2 Scheduled Cesarean Sections

As with scheduled inductions, a patient may be scheduled for a cesarean section at the recommendation of her OB/GYN. Such a situation could arise if vaginal delivery is deemed unsafe for the mother or infant; however, scheduled cesarean sections can also be elective. C-section patients are scheduled to arrive at the labor and delivery floor approximately 2 hours before their procedures and are admitted directly to recovery room beds on the labor and delivery floor. The patients are then prepped for surgery and moved to an operating room. After the C-section, patients are returned to their recovery room beds. Following a monitoring period, these patients are then moved to the postpartum floor and will remain there for approximately 4 days to ensure proper recovery.

3.3 Unscheduled Patients

The majority of patients on the labor and delivery floor are women who have not arrived for scheduled procedures but are either in spontaneous labor or responding to concerns about their pregnancies. For example, a woman may come to the floor if she has a headache or high blood pressure (signs of preeclampsia), has fallen and is concerned about the baby, or if the baby has decreased fetal movement. A woman may call her obstetrician's office first, which may recommend that she be evaluated in triage or be admitted to labor and delivery. A patient who either does not consult her obstetrician from home or whose obstetrician recommends an evaluation in triage will be seen by the triage nurse upon arrival at the labor and delivery floor. The triage nurse will then admit the patient to the labor and delivery or antepartum floor or send her home, depending on the needs of the patient (i.e., the severity of the patient's condition).

4 Data Set

For our analysis, we collected data from Beth Israel Deaconess Medical Center (BIDMC), a tertiary care medical center in Boston, for the 2014 calendar year. This data set includes timestamps for bed occupancy in all of the care centers of the unit: triage, the labor and delivery floor, the antepartum floor, and the postpartum floor. In total, the data includes 34,937 individual records of patient encounters. During 2014, this hospital treated 7,486 patients in labor and delivery, 6,060 (80.95%) of whom delivered babies. Of these 6,060 patients, 3,947 delivered vaginally (65.13%) and 2,113 (34.87%) delivered via C-section. These patients generated a total of 6,778 visits to L & D Triage, 6,361 visits to the L & D Floor, 943 visits to Antepartum, and 5,072 visits to Postpartum. Further, 1,286 (21.22%) of the 6,060 patients were scheduled for an induction. These numbers are similar to other tertiary care centers throughout the United States of America. We note that our data comes from raw, experimentally uncontrolled data, entered manually (rather than through RFID tracking) into an electronic database.

5 Modeling

In this section, we develop an accurate theoretical model of operations in labor and delivery. This model enables us to assess the hypothetical performance of the floor as a function of key model parameters. Specifically, we use our model to assess the impact of varying the C-section rate on the number of beds required to adequately care for patients on the L & D Floor and the postpartum ward (Section 6), and we evaluate the efficacy of various care paradigms as a function of the C-section rate (Section 7). In the following sections, we develop our theoretical model (Section 5.1) and validate its attributes (Sections 5.2).

5.1 The M/G/∞ Queueing Model

In outlining the development of our model, we first review the fundamental M/M/m queueing theoretic model. Second, we relax two aspects of this model to better reflect labor and delivery processes, which, in turn, gives us the M/G/∞ queueing model. For more background on the model, see Gautam [20] or Larson and Odoni [36].

5.1.1 Preliminaries: The M/M/m Queueing Model

The M/M/m queue has three components. First, the model has a set of m servers (beds), each of which can process customers (patients), as denoted by the "m" in "M/M/m." Second, patients arrive according to Poisson ("Markovian") process, as denoted by the first "M" in "M/M/m." The Markovian aspect implies that the time at which one patient arrives at the hospital does not depend upon when the previous patient arrived. This time between two patient arrivals ("inter-arrival time") is exponentially distributed according to a Poisson process. Third, customers are served (i.e., spend time using the server) according to an exponentially distributed ("memoryless") process, as denoted by the second "M" in "M/M/m." The term "memoryless" reflects that the time one patient will spend in a bed is independent of how long that patient has spent in a bed thus far.

Figure 2 depicts a graphical description of the M/M/m queueing model. The nodes of the graph represent the system state (i.e., the number of patients in the queueing system). The directed edges between nodes represent transitions from one state to another (i.e., the arrival of a new patient or the discharge of an existing patient). The weights of the edges represent the relative likelihood of transitioning from one state to the next. The edge weight for the addition of a patient to the system is typically denoted, λ, representing the average number of patients arriving per hour. The edge weight for the discharging of a patient from the system is typically denoted, μ, representing the average number of discharges per hour.

As the number of patients in the system increases, the probability of discharging any one of those patients increases up until the number of patients equals the number of servers (e.g., beds), as depicted by states 0 through m. When there are more patients than servers, the excess patients are processed sequentially as servers become available. As such, the weight for transitioning from a state with more than m patients to one with fewer than m patients is a constant $m\mu$. Patients waiting for a server are admitted in a first-come-first-serve priority.

The average time a patient spends waiting for a server is given by Equation 1, where $C(.,.)$ is Erlang's C formula [45], the arrival rate, λ (corresponding to the first "M in M/M/m), the rate of service, μ (corresponding to the second "M in M/M/m), and the number of servers, m (corresponding to the "m in M/M/m). Intuitively, this equation shows that as the number of servers or the service rate increase, the patient waiting time decreases. However, as the patient arrival rate increases, the patient waiting time increases. Finally, the system is considered stable (i.e., the average wait time is finite) as long as $m\mu < \lambda$ [30].

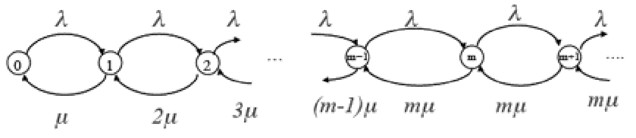

Fig. 2 An M/M/m transition diagram.

$$E[W] = \frac{C(m, \lambda/\mu)}{m\mu - \lambda} \tag{1}$$

5.1.2 The M/G/∞ Queue: Relaxing the M/M/m Model

While the M/M/m queueing model captures a variety of processes, there are two aspects of L & D that are not well described by this model: 1) exponentially distributed service times and 2) the m-server capacity. First, patients waiting to be admitted to a bed are not simply waiting – their labor is not arrested while waiting unattended in the waiting area or elsewhere on the labor floor. Instead, each woman's body is "processing the pregnancy" in parallel. As such, the number of servers is equal to the number of patients. This type of phenomena is typically modeled as an ∞-server queueing system (e.g., the M/M/∞ queue).

Second, service times are not exponentially distributed. Rather, the service times are a function of the individual acuity of the patient. For example, a patient requiring a C/S will typically deliver her baby more quickly than a patient delivering vaginally. As such, the service times for patients in L & D are better captured by a "general," denoted by replacing the second "M" with a "G" (e.g., the M/G/m queue).

Combining these two aspects yields the M/G/∞ queueing model. The M/G/m and M/G/∞ models are strictly more general than M/M/m and M/M/∞ queues, respectively. Any system that can be modeled as an M/M/m or M/M/∞ queue can also be modeled as an M/G/m or M/G/∞ queue, respectively. Lastly, we note that we maintained the Markovian patient arrival process from the original M/M/m queue, as we found that it accurately modeled the arrival processes at our hospital of interest.

5.1.3 Validation Metrics

To assess the applicability of this model, we validated three key attributes: patient inter-arrival times, patient service times, and the queue size (i.e., bed occupancy). We validated these

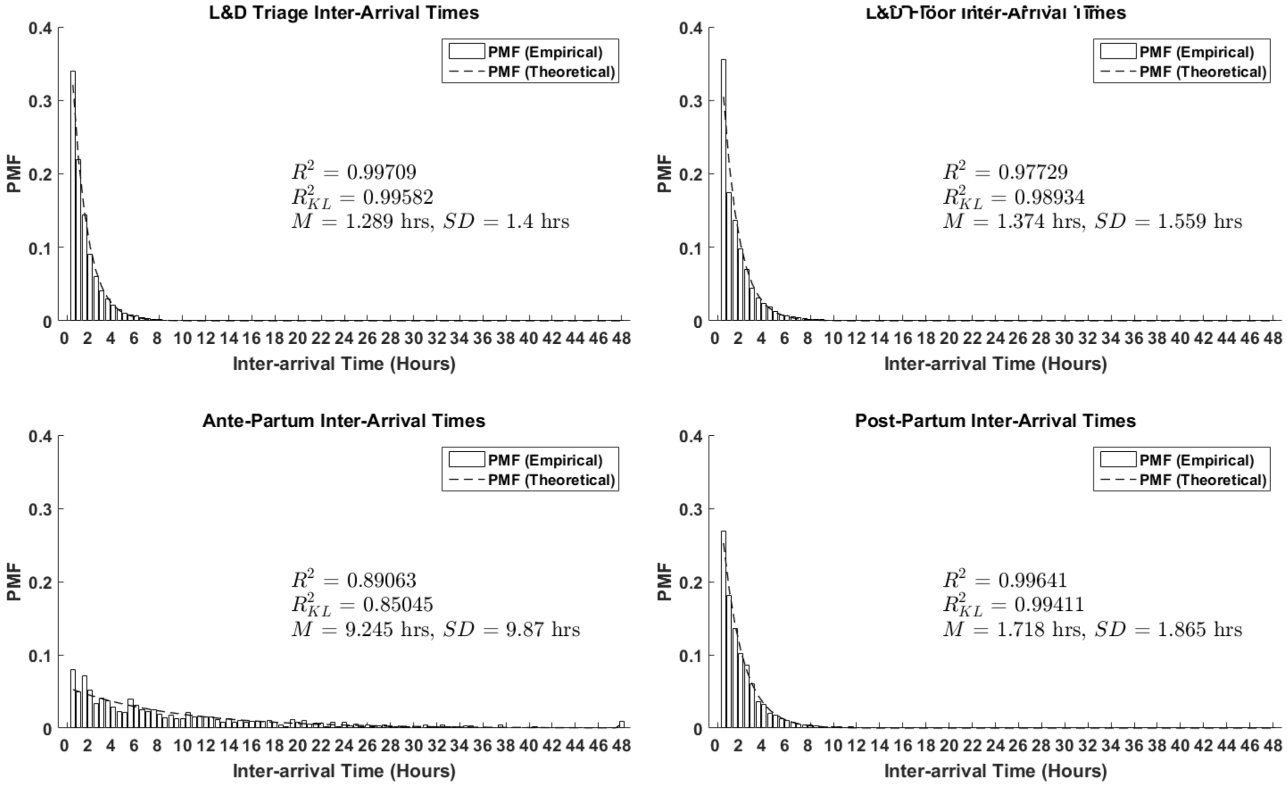

Fig. 3 A histogram (normalized) of the inter-arrival times of mothers admitted to triage (upper left), the labor and delivery floor (upper right), the antepartum floor (lower left), and the postpartum floor (lower right). The y-axis depicts the probability of a given inter-arrival time.

attributes across the four primary care centers: L & D triage, the L & D floor, the postpartum ward, and the antepartum ward. As our key metric, we report the R^2 value for how well the queueing theoretic model predicts the actual distribution of patient inter-arrival times, patient service times, and queue size. Note that the R^2 value represents the proportion of the variance explained by the model. Because the R^2 statistic can be less helpful for nonlinear regression [7], we also report a pseudo-R^2 statistic, denoted R^2_{KL}, which is based on the KL-divergence [32]. R^2_{KL} is computed through Equation 2, where $KL(y,\hat{y})$ is the KL-divergence between the data, y, and the fitted values, $\hat{y}$ (i.e., as predicted by the exponential distribution); likewise, $KL(y,\bar{y})$ is the KL-divergence between the data and the mean value, $\bar{y}$, of the data [8].

$$R^2_{KL} = 1 - \frac{KL(y,\hat{y})}{KL(y,\bar{y})} \tag{2}$$

The M/G/∞ queue has two specific limitations when applied to L & D operations: First, transient phenomena inherent in L & D operations are not modeled by the M/G/∞ queue. For example, physicians typically arrive at the hospital during the early morning, make “rounds” on their patients before lunch, and discharge patients in the early afternoon. Further, appointment times for scheduled C-sections and inductions are not random: BIDMC and many other hospitals maintain predefined times for such procedures. These phenomena violate the memoryless assumption of the M/G/∞ queue. Second, because of these and other transient phenomena – for example, pregnancy rates may be higher in the winter than summer – the flow of patients on the labor and delivery floor is never truly at steady-state. This violates an assumption made when characterizing the performance of a queueing theoretic model: that the system is in equilibrium. Nonetheless, we validate in Sections 5.2 through 5.4 that the M/G/∞ model is capable of modeling patient inter-arrival times, patient service times, and queue size across the four primary care centers in L & D.

5.2 Inter-arrival Times

In an M/G/∞ queue, patient arrivals are governed by a Poisson arrival process. In such a process, the time between two patient arrivals is exponentially distributed, with an average arrival rate of λ patients per hour. In turn, the average and standard deviation of the inter-arrival times is given by $\frac{1}{\lambda}$. Thus, we would expect the mean (M) and standard deviation (SD) to be equal if the inter-arrival times at an L & D care center were exponentially distributed. Further, we would expect the exponential distribution to well approximate patient inter-arrival times.

Table 1 Inter-arrival times at L & D triage, the L & D floor, and the inpatient floors.

	Mean (hrs)	SD (hrs)	Diff. (%)	R^2	R^2_{KL}
L &D Triage	1.289	1.400	7.93%	0.997	0.996
L & D Floor	1.374	1.559	18.28%	0.977	0.989
Antepartum	9.245	9.870	6.33%	0.891	0.850
Postpartum	1.718	1.865	7.88%	0.996	0.994

Table 2 Results of χ^2 tests for independence are reported for consecutive Inter-arrival times at L & D triage, the L & D floor, and the inpatient floors.

	χ^2 value	χ^2 critical value	p-value
L & D Triage	$\chi^2(121) = 56.364$	147.674	≈ 1
L & D Floor	$\chi^2(169) = 70.361$	200.334	≈ 1
Antepartum	$\chi^2(2,209) = 668.9$	2,319.455	≈ 1
Postpartum	$\chi^2(196) = 86.8502$	228.663	≈ 1

To determine the overall arrival process, we constructed histograms for the inter-arrival times among mothers arriving in L & D triage and the L & D, antepartum, and postpartum floors (Figure 3). Table 1 depicts the mean and standard deviation of the inter-arrival times, as well as the R^2 values for the exponential curves for each ward, with the corresponding rate parameter λ (patients per hour) set to the empirical average. The data indicate that the arrival process via the inter-arrival times is well-approximated as an exponential distribution. Specifically, the R^2 values for the inter-arrival times of L & D triage, the L & D floor, antepartum, and postpartum are 0.997, 0.977, 0.891, and 0.996, respectively. The corresponding R^2_{KL} values for the inter-arrival times of L & D triage, the L & D floor, antepartum, and postpartum are 0.996, 0.989, 0.850, and 0.994, respectively, denoting a strong fit.

To provide further evidence of the validity of modeling patient arrivals as a Poisson arrival process, we investigate the relationship between consecutive inter-arrival times. For the model to be valid, consecutive inter-arrival times should be independent. Figure 4 depicts a scatter plot of a given inter-arrival time (i.e., inter-arrival time $i+1$) versus the preceding arrival time (i.e., inter-arrival time i). From these figures, as well as the results of χ^2 tests for independence which are reported in Table 2, we demonstrate that consecutive inter-arrival times are independent. Specifically, the χ^2 tests – with a bin size of 1-hour intervals and applying the Yates' correction for continuity [56] – show that the probability that consecutive inter-arrival times are independent is approximately 1; thus, we do not reject the null hypothesis that consecutive inter-arrival times are independent.

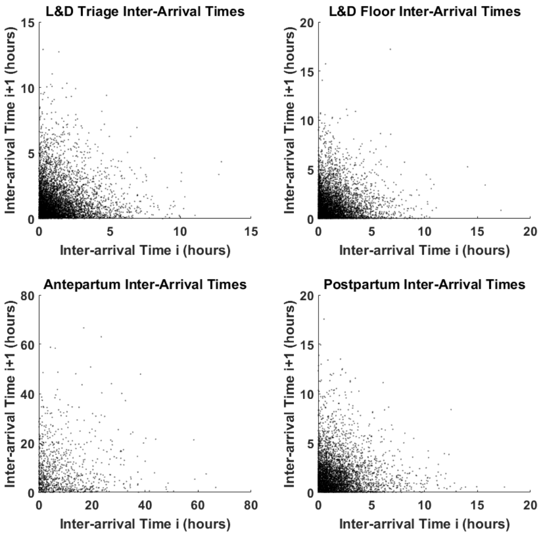

Fig. 4 A graphical depiction of the relationship between consecutive inter-arrival times for mothers admitted to triage (upper left), the labor and delivery floor (upper right), the antepartum floor (lower left), and the postpartum floor (lower right).

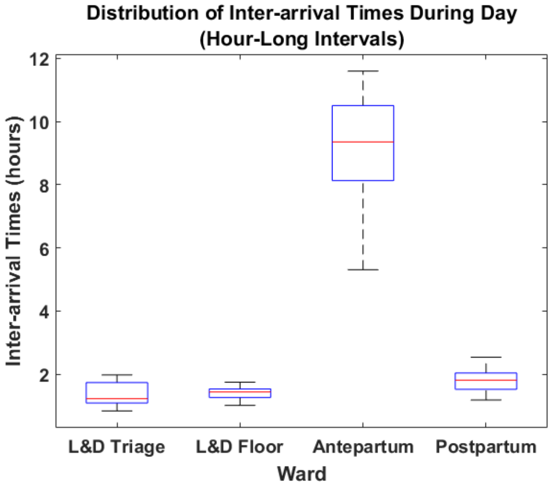

Fig. 5 This figure depicts the distribution of inter-arrival times as a function of the time of day for each ward.

Finally, we also consider whether the arrival process is time-varying. We expect there to be periodic fluctuations in the arrival rate based upon controllable (e.g., scheduled inductions start at either 8 a.m. or 8 p.m.) and uncontrollable factors (e.g., patients may be less likely to come to the hospital during their sleeping hours). To determine the degree to which the inter-arrival time, $1/\lambda$, varies during the day, we compute the average inter-arrival time for each 1-hour window during a 24-hour day (e.g., midnight to 1 a.m., 1 a.m. to 2 a.m., etc.) for each ward. Figure 5 depicts the distribution of these average inter-arrival times. The mean and standard deviation (hours) for L & D Triage, the L &D Floor, Antepartum and Postpartum are 1.378 ± 0.370, 1.391 ± 0.177, 9.218 ± 1.568, and 1.7851 ± 0.375, respectively. The distributions for L & D Triage and Postpartum are relatively narrow considering the wide range of controllable and uncontrollable factors that alter the arrival process.

We note, however, that the Antepartum floor has a larger variance for two reasons. First, OBGYN out-patient clinics are typically open during normal business hours (e.g., 9 a.m. - 5 p.m., Monday through Friday). Conditions for which an admission to Antepartum are warranted are more likely to be detected during an out-patient visit, and, in turn, be admitted to Antepartum during normal business hours. Second, patients are less likely to become aware of a pregnancy complication while they are sleeping, thus increasing the proportion of admission to Antepartum during the day. While the above is true for the other wards (e.g., L & D), both the type and sheer volume of patients visiting the other wards better reflects normal physiology, which is largely random.

Despite this variance, we show in Section 5.4 that the M/G/∞ queueing model very tightly approximates the figurative circadian rhythm of the wards involved in labor and delivery (See Figure 7). We believe this close approximation shows that the M/G/∞ model has value in guiding the management of resource needs in labor and delivery.

5.3 Service Times

The service time (length of stay) at care centers is influenced by a number of factors. In triage, mothers who require admission to the labor and delivery floor may be delayed while the L & D floor prepares to receive them; at the same time, mothers who can be safely discharged may experience an expedited service time due to the reduced acuity of their condition. Women on the antepartum or postpartum floors are typically discharged once or twice per day in batches when doctors make their rounds. Further, women in active labor have distinct modes of service time: nulliparous patients (i.e., ones who have not birthed any children) typically experience significantly longer labor than multiparous patients (i.e., ones who have birthed at least one child). Furthermore, doctors may intervene in the normal course of labor via cesarean section if the mother's or baby's health is at serious risk. The duration of a cesarean section is typically much shorter than that of a spontaneous vaginal delivery, increasing the complexity of the model.

While the inter-arrival times in our data (Section 5.2) are accurately modeled with an exponential distribution, the service times are not. In this section, we justify the use of the M/G/∞ queue over the M/M/∞ queueing model by inspecting two key discrepancies between the data and the corresponding behavior predicted by the M/M/∞ queue, as shown in Figure 6 and Table 3.

First, service times on the L & D floor and the postpartum ward are dependent upon the mode of delivery. Service times for patients on the L & D floor who deliver vaginally (M = 13.016 hrs, SD = 4.246 hrs) are longer than those among patients who deliver via C-section (M = 9.012, SD = 5.723), t(2264) = -146.715, p $<$ 0.001. Service times for patients on the postpartum ward who deliver vaginally (M = 46.943, SD = 6.705) are shorter compared with those who deliver via C-section (M = 87.243, SD = 9.963), t(4089) = 29.820, p $<$ 0.001. While not ideal, these results are to be expected. Intervention via cesarean section curtails the time a patient would taken to deliver the baby via normal vaginal delivery, but a C-section also requires the patient to be monitored in the postpartum ward for 4 days, as opposed to 2 days among patients who deliver vaginally.

Second, the exponential distribution does not accurately capture empirical service times. If the data were exponentially distributed, the mean and standard deviation would have to be equivalent; however, these statistics do not correspond: The difference between the mean and standard deviation of the service times in our data set ranges from 36% to 89%.

Due to these problem characteristics, we adopted the M/G/∞ model, which allows inclusion of any distribution that can be parameterized by a mean and variance, as opposed to M/M/∞, which would require that these service times be exponentially distributed.

We note that the mean and standard deviation depicted in Table 3 are empirically derived from our data set, which is described in Section 4. Further, since service time data for patients on the L & D Floor and Postpartum were heavily dependent on the method of delivery necessitating the separate analysis in Table 3. Specifically, the number of visits to the L & D Floor for patients delivering vaginally and via C-Section were 3,924 (61.7%) and 2,437 (38.3%), respectively, and to Postpartum for patients likewise delivering vaginally and via C-Section were 3,485 (68.7%) and 1,587 (31.3%), respectively. One can use these statistics to back out the overall, average service time $^1/_\mu$ for a given ward using Equation 3, where $^1/_{\mu_{vaginal}}$ and $^1/_{\mu_{c/s}}$ are the average service times for patients delivering vaginally and via C-Section, respectively, and $p_{vaginal}$ and $p_{c/s}$ are the proportion of patients delivering vaginally and via C-Section, respectively.

$$\frac{1}{\mu} = \frac{p_{vaginal}}{\mu_{vaginal}} + \frac{p_{c/s}}{\mu_{c/s}} \tag{3}$$

We observe that the proportion of C-Section patient visits on Postpartum is slightly higher than for the L & D Floor. C-Section patients are generally more acute, requiring more frequent monitoring. This monitoring would occur on the L & D floor. As such, C-Section patients may visit the L & D floor multiple times prior to delivery. On the other hand, they would only visit postpartum once, which would occur after delivery.

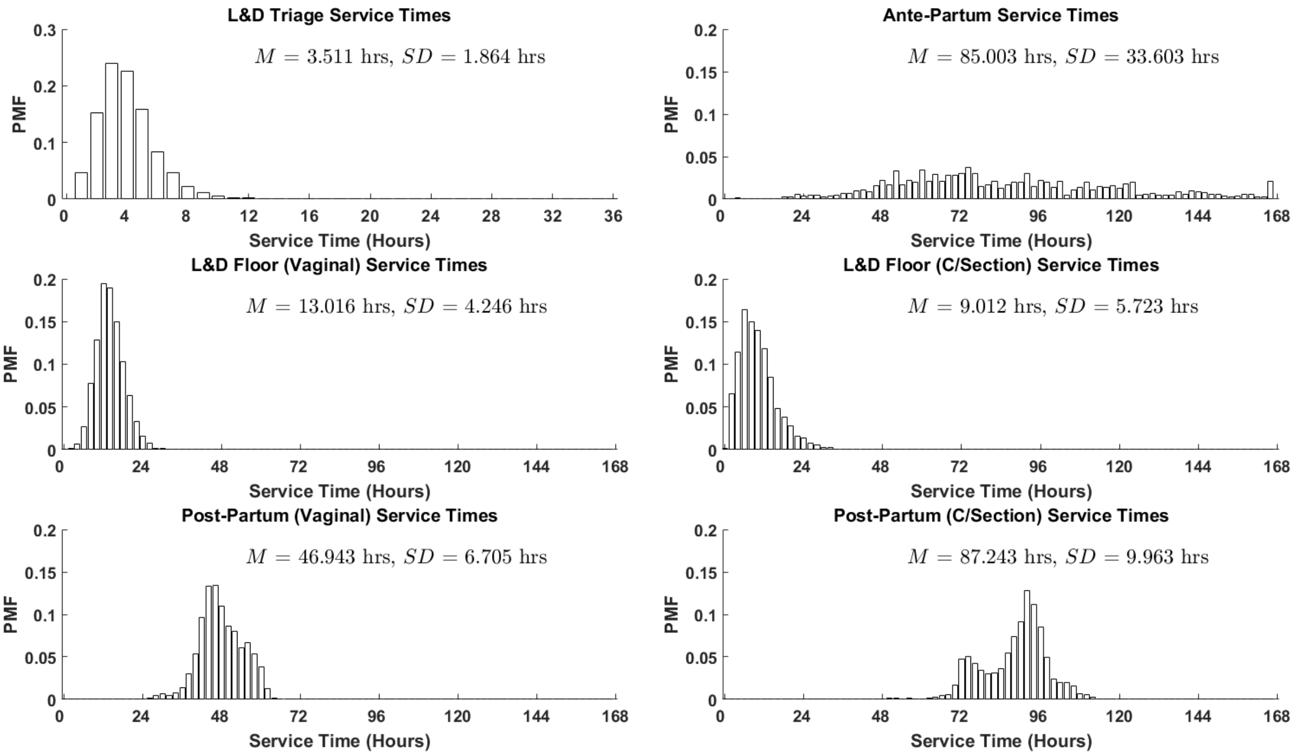

Fig. 6 A histogram (normalized) of the service times for mothers in triage (upper left), labor and delivery floor vaginal deliveries (middle left), labor and delivery floor cesarean section deliveries (lower left), mothers in antepartum (upper right), postpartum vaginal deliveries (middle right), and postpartum cesarean section deliveries (lower right).

Table 3 Mean and Standard Deviation Service times on L & Triage, the L & D floor, and the inpatient floors. For the L & D Floor and Postpartum, the statistics are reported based on the method of delivery; the proportion of patient visits for each method is also reported.

Ward	Delivery Method	Mean (hrs)	SD (hrs)	Diff.
L & D Triage	Both	3.511	1.864	46.9%
L & D Floor	Vag. (61.7%)	13.016	4.246	67.4%
	C-S (38.3%)	9.012	5.723	36.4%
Antepartum	Both	85.003	33.603	60.5%
Postpartum	Vag. (68.7%)	46.943	6.705	85.7%
	C-S (31.3%)	87.243	9.963	88.6%

5.4 Bed Occupancy

A salient measure of system performance is the number of beds occupied in a care center; beds and associated resources are among the primary drivers of care costs. Hospitals often employ a systems-level analysis to estimate the correct number of beds necessary to provide in order to handle the patient population. In order to validate the applicability of the M/G/∞ model, we compared the proportion of time m beds are occupied to the probability of m beds being occupied as indicated by the model. The probability of m beds being occupied at any one time according to an M/G/∞ is given by Equation 4 [42], where λ is the patient arrival rate, $\frac{1}{\mu}$ is the mean service time, and $\gamma(t)$ is the number of patients in the system at time t:

$$\lim_{t\to\infty} Pr\{\gamma(t) = m\} = \frac{\left(\frac{\lambda}{\mu}\right)^m e^{-\frac{\lambda}{\mu}}}{m!} \tag{4}$$

Figure 7 depicts the actual and theoretical proportion of time during which m beds are occupied at any given moment. Likewise, Table 4 reports the empirical and expected mean and standard deviation of the bed occupancy, the model error, the R^2 value, and the R^2_{KL} value of the M/G/∞ queueing model for each ward. Specifically, we found that the R^2 values for bed occupancy in L & D triage, the L & D floor, antepartum, and postpartum were 0.985, 0.928, 0.834, and 0.936, respectively. The corresponding R^2_{KL} values are 0.987, 0.883, 0.867, and 0.867. Note that the hospital from which this data was collected has a total of six triage beds and 13 beds on the L & D floor (including three operating room beds and six recovery room beds). While the number of labor beds on the L & D floor is only 13, it is possible that more than 13 patients may temporarily be on the floor at the same time; in times of high demand, recovery room beds may act as overflow. Further, on occasion, patients will be placed double-booked in rooms or temporarily moved to a hallway in times of extreme overflow. Finally, due to imperfect, manual data entry, there may be a delay between when a patient is moved out of a location and the recording of that information.

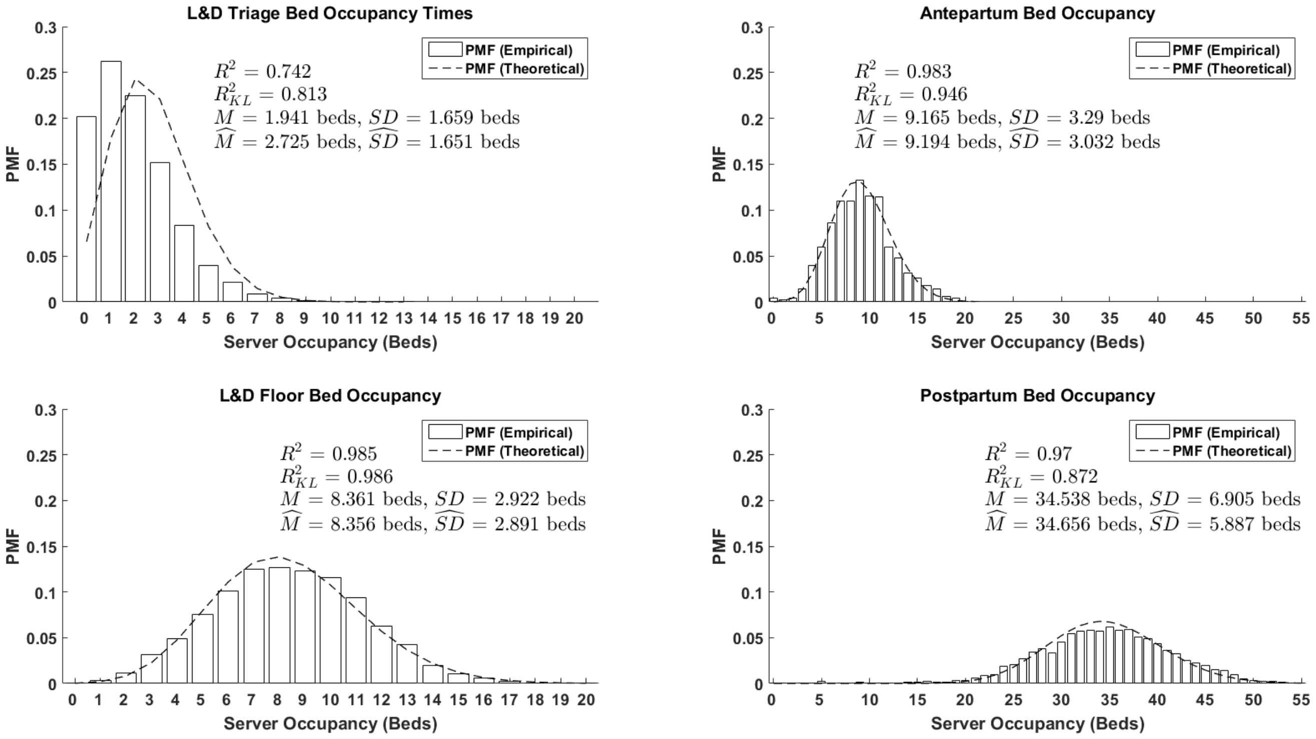

Fig. 7 A histogram of server occupancy (i.e., the number of occupied beds) in triage (upper left), the L & D floor (lower left), antepartum (upper right), and postpartum (bottom right).

Table 4 Mean and Standard Deviation of Bed occupancy (empirical, theoretical, and the error) for L & D triage, the L & D floor, antepartum, and postpartum.

	Mean (beds occupied)			SD (beds occupied)			R^2	R^2_{KL}
	Empirical (M)	Theoretical ($\widehat{M}$)	Diff.	Empirical (SD)	Theoretical ($\widehat{SD}$)	Diff.		
L & D Triage	1.941	2.725	28.771%	1.659	1.651	0.482%	0.742	0.813
L & D Floor	8.361	8.356	0.060%	2.922	2.891	1.061%	0.985	0.986
Antepartum	9.165	9.194	0.315%	3.290	3.032	7.842%	0.983	0.946
Postpartum	34.538	34.656	0.340%	6.905	5.887	14.743%	0.970	0.872

6 Resource Requirements as a Function of the Cesarean Section Rate

One core theme of research and improvement efforts in obstetrics is attempting to understand why the cesarean section rate is 35% when the optimal rate suggested by the World Health Organization is closer to 15%. Rates above 15-19%, on average, do not appear to improve maternal or fetal outcomes; furthermore, there are significant near- and long-term risks associated with cesarean sections [43]. As such, researchers have proposed lowering the C-section rate. However, we are unaware of a prior queueing theoretic investigation showing the consequences of such a rate reduction on the logistics of labor floor operations.

Utilizing our M/G/∞ queueing model, we are able to predict the effects of decreasing the cesarean section rate. We independently computed the service times for vaginal and cesarean deliveries and determined a new aggregate expected service time via a weighted combination of these specific times. In other words, the aggregate expected service time for a patient is equal to the sum of the product of p (the proportion of cesarean deliveries), and the expected service time for a cesarean delivery and the product of 1-p and the expected service time for a vaginal delivery. In order to provide a helpful analysis for researchers at hospitals with higher or lower arrival rates, we also varied arrival rates among patients.

As Figure 8 indicates, changing the C-section rate can have significant consequences for the resources required to care for patients on the labor and delivery floor and in postpartum. In this figure, arrival rates at BIDMC are depicted as a function of the arrival rate, λ_o, Vertical bars denote cesarean section rates of 15% and 35%.

The number of beds required to accommodate the average and 99th-percentile cases on the labor and delivery floor increases as the rate of cesarean sections decreases:

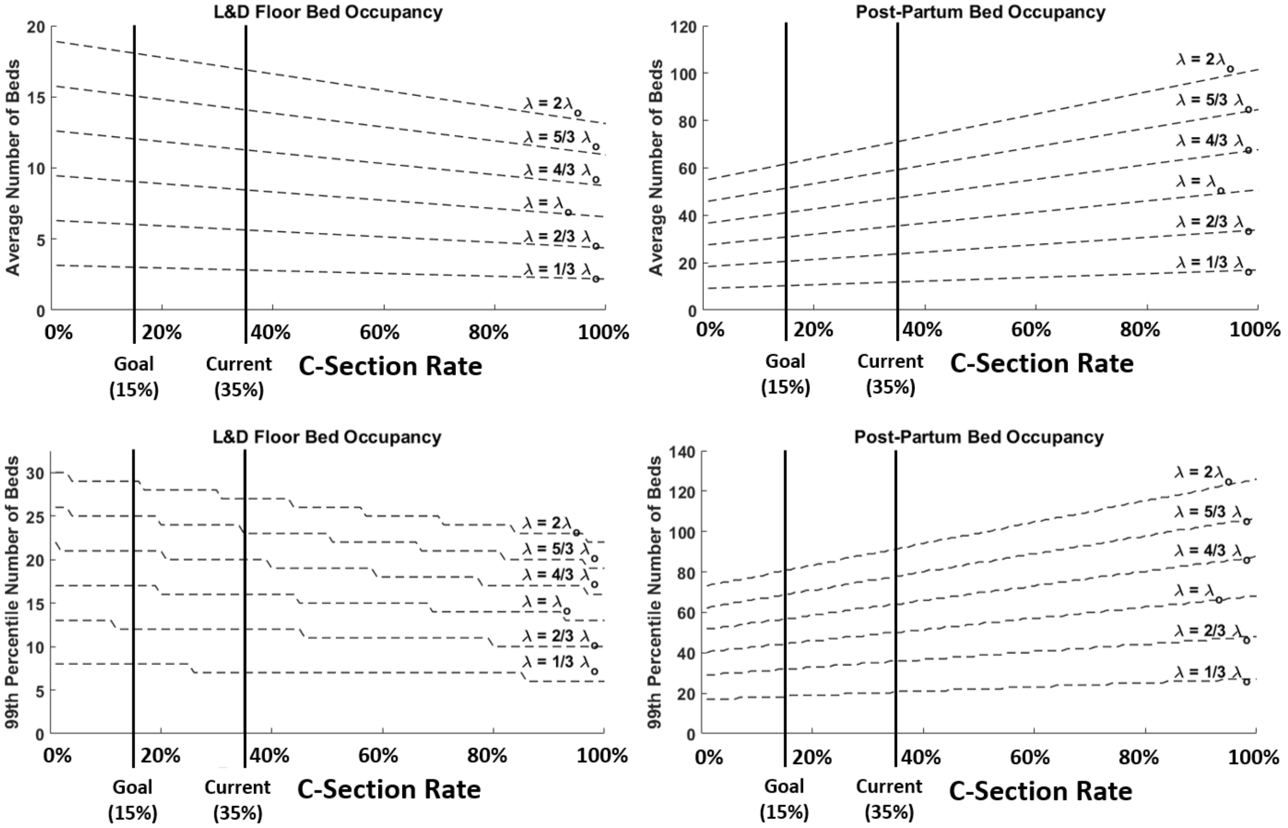

Fig. 8 This figure depicts the average (top) and 99% percentile (bottom) bed occupancy as a function of the C-section rate and the arrival rate of patients to the labor and delivery floor (left) and to postpartum (right). The current and ideal C-section rates of 15% and 35% are represented by vertical bars.

patients who deliver vaginally generally require longer delivery times than patients who undergo C-section. To cover the 99th percentile of patient occupancy, it would be necessary to increase the number of beds on the L & D floor at BIDMC by 1, which translates to an 8% increase.

Moreover, as cesarean section rates increase, the number of beds required to accommodate the average and 99th-percentile cases increases in postpartum: patients require a significantly longer amount of time (approximately twice as long) to recover in postpartum after undergoing a cesarean section than patients who delivered vaginally. Specifically, the number of beds in postpartum could be reduced by 6.7 – a 12% decrease – to cover the 99th percentile of patient occupancy. Because the typical length of stay in postpartum is substantially longer than on the labor floor, the greater magnitude of change for postpartum relative to the L & D floor is to be expected.

6.1 Financial Impact

In prior work, researchers at BIDMC conducted a financial investigation into the cost of care at BIDMC [50]. These researchers used the time-driven, activity-based costing method by Kaplan and Anderson [29]. Based on these cost rates, we can employ our M/G/∞ queue model to translate the change in occupancy requirements following a change in the C-section rate into a financial cost for our care center.

Shah et al. [50] showed that the average cost of personnel caring for an occupied L & D bed is $2.11 per minute, while the average cost of personnel caring for an occupied postpartum bed is $0.32 per minute. The average cost of maintaining an L & D bed and associated equipment is $0.22 per minute; the average cost of maintaining a postpartum bed and associated equipment is $0.04 per minute. Finally, the average cost of a C-section is $12.47 per minute.

In order to compute a cost, we assumed that the hospital maintains sufficient beds on the L & D floor and in postpartum to account for the 99th percentile of bed occupancy. Given this assumption, we estimated the total cost to the obstetrics unit, as shown in Equation 5. We assumed a baseline of 1,112 C-sections per year given a 35% C-section rate, which we based on data from BIDMC.

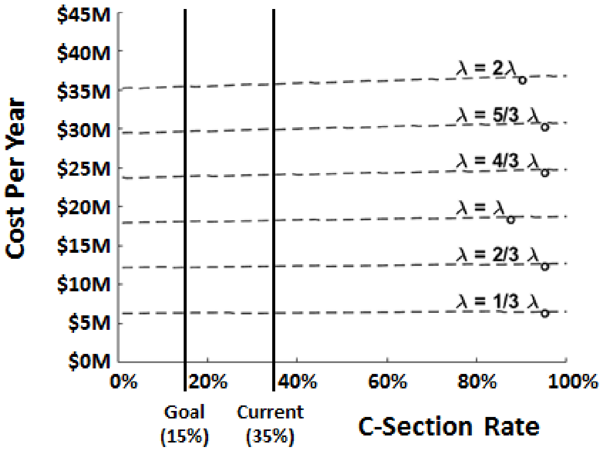

Fig. 9 This figure depicts the total cost per year, given by Equation 5, as a function of C-section and patient arrival rates.

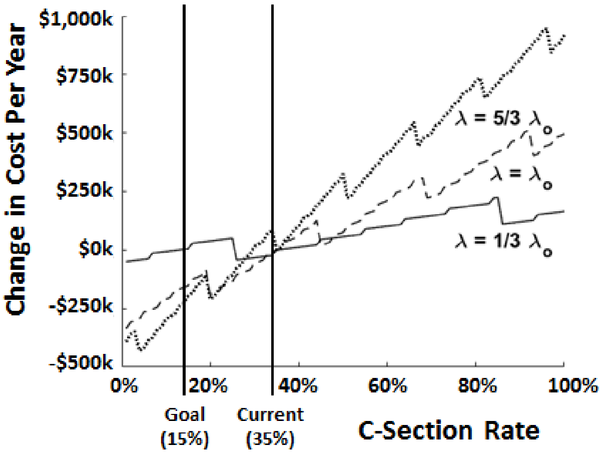

Fig. 10 This figure depicts the change in total cost per year, derived from Equation 5, relative to the baseline cost at a 35% C-section rate, as a function of C-section and patient arrival rates.

$$
\begin{aligned}
&\text{Total Cost Per Minute}\\
&= \$2.11 * (\text{average occupancy on L \& D})\\
&+ \$0.32 * (\text{average occupancy on Postpartum})\\
&+ \$0.22 * (\text{\# L \& D beds for 99th percentile occupancy})\\
&+ \$0.04 * (\text{\# Postpartum beds for 99th percentile occupancy})\\
&+ \$12.47 * (\text{\# C-sections per minute}) \qquad (5)
\end{aligned}
$$

Given this analysis, we were able to compute the total cost per year for the entire L & D ward, as well as the change in cost relative to the 35% C-section rate baseline, as depicted in Figure 9 and 10, respectively. Surprisingly, the cost savings observed in the reduction of relatively less expensive postpartum beds outweighed the cost increase resulting from the greater number of relatively expensive L & D beds necessary to accommodate a reduction in the C-section rate. Decreasing the C-section rate from 35% to 15% would decrease the cost of the entire operating unit by $153k. However, note that the change as a percentage of the total cost for the operating unit at 35% is less than 1%.

6.2 Recommendation

Hospital administrators seeking to reduce cesarean section rates should be prepared to increase the number of beds available for patients on labor and delivery floors. Further, hospital management should also either anticipate more unoccupied beds on postpartum floors or reduce the number of beds in order to decrease cost. Specifically, our hospital of interest would need to increase the number of beds on the L & D floor by 8% and could decrease the number of beds in postpartum by 12% if there was a reduction in C-section rates from 35% to 15%. Because L & D often has a narrow financial operating margin, this aggregate change in resource requirements could have important implications for hospital operations.

At our hospital of interest, a decrease in the C-section rate does not appear to significantly impact operational costs. The cost savings from reducing the capacity of the relatively inexpensive postpartum ward marginally outweighs the increase in cost for the more-expensive L & D ward.

7 Inter/Intra-Team Deliveries

During pregnancy, women are typically monitored on an outpatient basis through regular prenatal visits. The obstetricians they select for their care are each a member of a team, with team members “taking call” in turns on the labor and delivery floor. While a member of the team is taking call, he or she is directly responsible for managing the care of any women seen by his or her team. Ideally, a woman’s own obstetrician will deliver her baby; however, due to the uncertain duration of gestation, a team member other than the woman’s primary attending physician may perform the delivery. Also, the L & D floor in a hospital may support multiple teams who concurrently share the hospital’s resources to care for their respective patients.

In one conceivable scenario, an obstetrician is delivering a baby in one room when a second woman under that obstetricians care enters the second stage of labor (i.e., the cervix is fully dilated) and has begun pushing. Since the same person cannot be in two places at once, hospitals use different strategies to ensure patients have access to clinical staff. At our hospital of interest, an obstetrician from a different team would be responsible for covering a patient until the occupied doctor is free. We call this model, in which doctor-to-patient care is flexible, the “team” model. However, there is an alternative, "individual," model, in which another staff member (resident or nurse) would be responsible for caring for and potentially delivering the second baby. Under this model, another available doctor of equal experience to the patient's primary doctor would not assist, because he or she is part of a different and unrelated team.

It is natural for a mother to prefer to have her baby delivered by a physician with whom she is already familiar. If the obstetrician from the patient's usual practice is occupied delivering another baby, then a mother would most likely then prefer that an obstetrician from a different team deliver her baby, as opposed to an unsupervised resident or nurse. Assuming this tiered set of preferences, a next logical question is what the benefits and detriments of the team and individual models are. Specifically, we wanted to develop a model capable of identifying what proportions of women have their deliveries conducted exclusively by an obstetrician from their usual practice, by attending obstetricians (regardless of practice affiliation), or by residents (at least for a portion of the delivery).

In order to determine the benefits of the team and individual models, we can employ the hypercube queueing model [34], as depicted in Figure 11. This model was developed to aid city planners in understanding how many ambulances (or similar emergency response units) a city should maintain given its population, as well as where to station those ambulances throughout the city's boroughs in order to ensure sufficient quality of service. This model represents queueing system states as nodes and transition probabilities between those states as weighted arcs connecting the corresponding nodes. However, we developed a novel analogy relating the hypercube model to physician practices so that we might model the time patients spend with their primary attending, any attending, and residents.

Consider a system under the team model consisting of two obstetrics practices: Patients in each practice arrive at the second stage of labor (i.e., fully dilated and ready to push) at rates of λ_1 and λ_2, respectively. (Note that this is not the arrival rate of women from each practice arriving at the hospital, but the rate at which a physician sees his or her patients who are entering the second stage of labor.) $S_{0,0}$ then represents a state wherein no women are in the second stage of labor for either practice. $S_{1,0}$ represents the state at which the obstetrician from Practice 1 is attending to a mother in the second stage of labor, with no other mothers at this labor stage (vice versa for $S_{0,1}$). $S_{1,1}$ refers to the state in which obstetricians from both Practice 1 and Practice 2 are attending to mothers in the second stage of labor in the absence of any other mothers at this stage.

Under these circumstances, a patient could be matched with an obstetrician from a different team via two routes: First, given that the system begins in state $S_{0,0}$, the system could experience two consecutive arrivals of women from the same practice at the second stage of labor, with such a brief inter-arrival time that the first woman had yet to finish delivering her baby before the second woman arrived. The first woman would then be cared for by the OB from her usual practice, while the second woman would be cared for by an OB from a different practice. The second route begins in state $S_{1,1}$. Given this initial state, two women are currently in the second stage of labor with both obstetricians occupied, and one of these obstetricians will finish first. A third woman at the second stage of labor might arrive after the first obstetrician completes delivery, but before the second obstetrician does (i.e., the system is in state $S_{1,0}$ or $S_{0,1}$). If this third woman is from a practice other than that of the obstetrician who finished delivery first, then that third woman would be seen by that obstetrician. If both obstetricians remain occupied upon the arrival of a third woman, that patient would be seen by a resident.

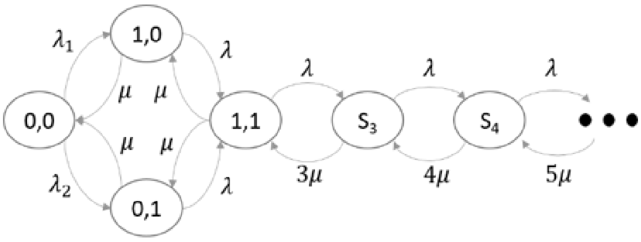

Fig. 11 An infinite-server hypercube state space representation with two primary servers.

In our analysis, we assume that women entering the second stage of labor arrive according to an exponential distribution with rate λ_o. Given that there are n teams on the labor and delivery floor, we also assume that a woman has an equally likely chance of being cared for by any one of the teams. As such, the arrival rate of women for any given team is $\lambda = \frac{\lambda_o}{n}$.

We further assume patients receive care with an average service time $\frac{1}{\mu}$. Here, the duration of service is equal to the duration of the second stage of labor, when the mother is actively pushing. We seek to determine the likelihood that an obstetrician from the desired practice is present during this stage. While we cannot estimate μ directly from our data, we can use data taken from a cohort of 4,126 mothers in a prior study by Rouse et al. [47]. Of the 3,152 women who delivered, the duration of the second stage of labor was between 0-1 hours for 1,901 mothers, 1-2 hours for 1,251 mothers, 2-3 hours for 217 mothers, 4-5 hours for 97 mothers, and longer than 5 hours for 46 mothers. If we assume the average duration of the second stage of labor is equal to the weighted sum of the middle of the range of the bin (e.g., 0.5 hours for the 0-1 hour bin and 5.5 hours for the >5 hours bin), then the average duration of the second stage of labor is 1.41 hours ± 1.10 hours. This distribution is approximately

Table 5 The steady-state probabilities of being in state S_i for the M/M/∞ queueing model.

	S_0	S_1	S_2	S_Q
$P(S_i)$	0.360	0.368	0.189	0.085

exponentially distributed; thus, we estimate the service rate to be μ=0.709 mothers per hour.

The first step for constructing the system as a hypercube model is to determine the steady-state probabilities for a collapsed version of the model: a simple M/G/∞ queue with arrival rate λ_o and average service time $\frac{1}{\mu}$. We can readily employ Equation 4 to calculate these steady-state probabilities, which are depicted for a two-practice model in Table 5. Here, S_i represents a state (i.e., the temporary condition of the labor floor) where i obstetricians are occupied, and S_Q represents a state in which there are more women in the second stage of labor than there are obstetricians (i.e., residents are caring for patients).

Note that $P(S_i)$ in the simplified M/G/∞ queue is equal to the sum of the probabilities of the associated hypercube states in which there are i women in the second stage of labor; in other words, $P(S_i) = \sum_{j,k|j+k=i} P(S_{j,k})$. Because we assume that μ and λ are not a function of which care team is caring for the patient, we can make the simplifying assumption that $P(S_{j,k}) = P(S_i)/n$ for all $j+k=i$ where n is the number of care teams, thus circumventing the assumption of Markovian service times in the traditional hypercube model. We note that if the arrival rates or service times of patients were a function of the patient's care time, a non-Markovian extension of the hypercube model could be employed [35]. Nonetheless, for our purposes, we can calculate the steady-state probabilities for an example two-server hypercube system in Equations 6 through 8.

$$P(S_{0,0}) = P(S_0) = 0.358 \tag{6}$$

$$P(S_{1,0}) = P(S_{0,1}) = \frac{P(S_1)}{2} = 0.184 \tag{7}$$

$$P(S_{1,1}) = P(S_2) = 0.189 \tag{8}$$

Next, we determine the fraction of all dispatches (an attending physician or resident responding to a mother entering the second stage of labor) for which the mother is treated by an obstetrician from her practice and experiences no queue delay. This fraction of dispatches is presented in Equation 9 for the same two-server (two-practice) system:

$$\begin{aligned} f_{1,1} = f_{2,2} &= \frac{\lambda_1}{\lambda}\left(P(S_{0,0}) + P(S_{0,1})\right) \\ &= \frac{1}{2}(0.358 + 0.184) = 0.271 \end{aligned} \tag{9}$$

The fraction of total dispatches for which an obstetrician from the patient's usual practice is present for the entire duration of the second labor stage is presented in Equation 10:

$$F_I = f_{1,1} + f_{2,2} = 0.271 + 0.271 = 0.542 \tag{10}$$

Thus, for the team model, during the second stage of labor, 54.2% of patients are cared for exclusively by an obstetrician from their usual obstetrician's practice, while 18.4% of patients receive care from an obstetrician not from their obstetrician's practice, and 27.4% of patients are cared for in part by a resident.

For the individual model, we can again model each obstetrician as an M/G/∞ queueing system with arrival rate $\frac{\lambda_o}{n}$. There are two outcomes in this model: either a patient receives care entirely from an obstetrician from her usual practice, or the patient is cared for (partially or entirely) by a resident. (Note that residents are not assigned to a particular practice.) The proportion of patients seen exclusively by an obstetrician is $P(S_0)$, and the proportion of patients who receive care from residents is $1 - P(S_o)$. In our two-practice example, the probabilities for these outcomes are 35.8% and 64.2%, respectively.

We investigated the benefits of the team and individual models for hospitals hosting between one to four teams and for patients arriving at their second stage of labor at rates ranging from $\frac{i\lambda_o}{3}, i \in \{1, 2, \ldots, 6\}$, as shown in Figure 12. Figure 12 (left-hand side) depicts the fraction of patients under the team model who would receive care from obstetricians from their usual practices, from practices other than their usual practice, and from residents. Figure 12 (right-hand side) shows the fraction of patients under the individual model who receive care from the attending obstetrician from their usual practices, as well as from residents.

Figure 13 presents this information in the form of the probability of patients receiving care exclusively from their primary attending physician (i.e., the attending physician from her usual obstetrician's practice), any attending physician regardless of practice affiliation (including the primary obstetrician), or by a resident, for both the team and individual models.

These figures indicate that the individual model yields better performance if the goal is to maximize the likelihood that a patient is exclusively cared for by an attending physician from her usual obstetrician's practice. For example, as the rate of patients entering labor increases to 1 pt/hour, a patient has an ∼10% increased chance of receiving care only from her primary attending physician. However, this occurs along with a decrease in the probability that the patient is seen by any attending physician: as the rate of patients entering labor increases to 1 pt/hour, the chance of being cared for by a resident instead of a primary attending physician increases by 50%.

As shown in the bottom chart of Figure 13, the relative advantage of the team model decreases as the rate at which women enter labor increases. Upon inspection, however, this

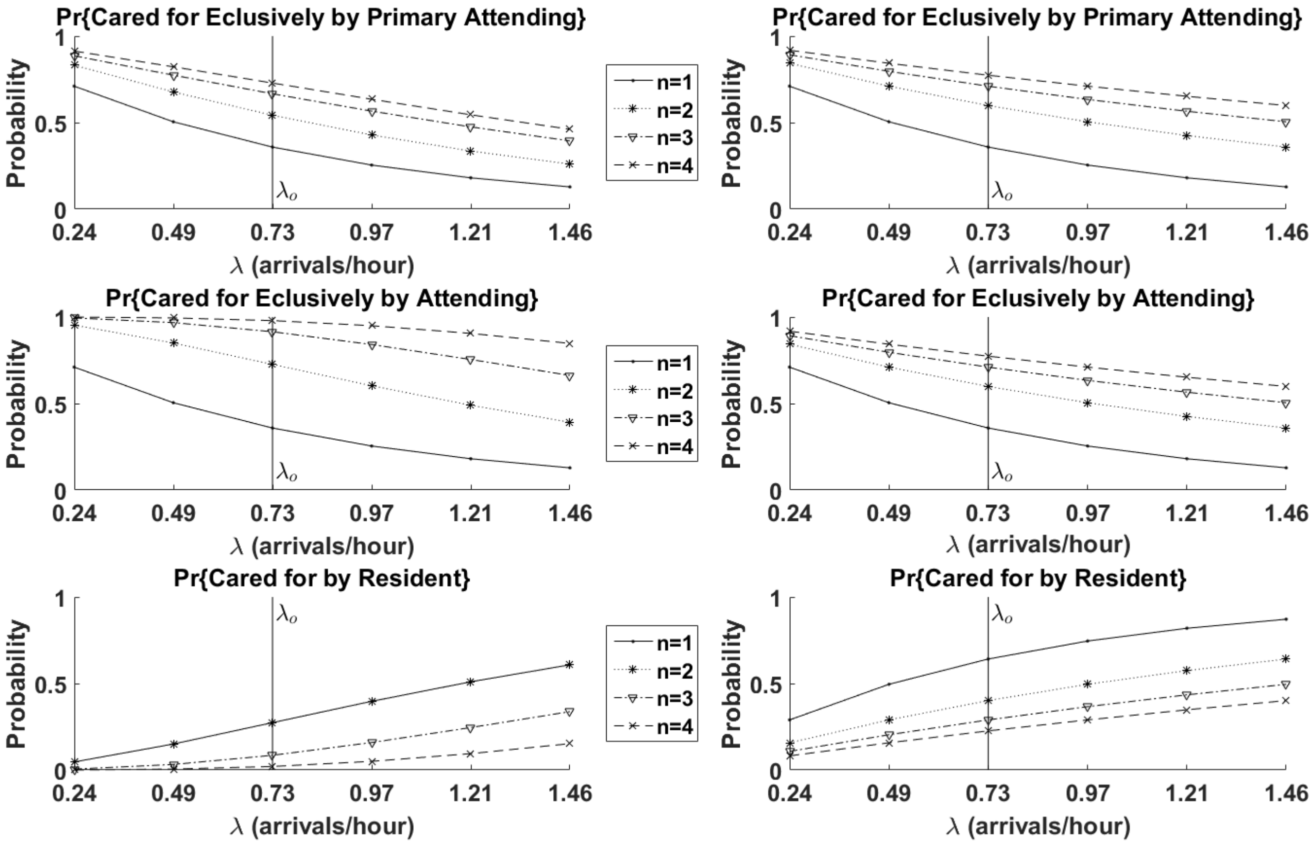

Fig. 12 This figure depicts the probability of a woman being cared for by her primary attending physician (top), any attending physician (middle), and by a resident (bottom), depending on whether the hospital operates under a team model (left) or individual model (right). Note that the top-right plot is exactly equivalent to the middle-right plot because only the primary attending physician is available to care for the patient under the individual model.

behavior is interpretable. Consider a scenario in which two patients are in labor at the same time at a hospital staffed by two teams. There is an $\sim 50\%$ chance that the two patients are assigned to the same physician. Under the team model, both patients could be seen by an attending physician; however, under the individual model, the second attending physician would remain idle while a resident cared for the second patient. Now, imagine a scenario involving 20 patients: under the individual model, it is more likely that both attending physicians would have patients from their respective practices present, and would thus be more fully utilized. Thus, the relative advantage of the individual model tends to disappear as the number of patients in the hospital relative to the number of attending physicians increases. Nonetheless, even at $\sim$1.5 labor arrivals per hour, the relative advantage of the team model ranges from 30% to 90% based on the number of practices represented by attending physicians.

7.1 Recommendation

We believe that supervision by at least one attending physician – regardless of whether that physician is a member of a patient's primary care group – is preferable for patient care compared to a delivery lacking full supervision. As such, the team model for patient care appears to be optimal. Of course, there are patient-specific concerns that might warrant the use of an individual model in select cases: For example, a certain attending and resident may have intimate knowledge of a particularly complex patient, and it might be dangerous for an attending lacking such knowledge to perform a delivery in such a case.

We provide this recommendation based upon our sensitivity analysis (Figures 12 and 13) of the probability an attending is available during the entire delivery. This probability is fundamentally a function of the patient arrival rate, λ, the number of care teams, n, the average duration of the final stage of labor, and whether the hospital is operating under an intra- or inter-team delivery scheme. In our sensitivity analysis, we varied $\lambda \in \{\frac{i}{3}\lambda_o | i \in \{1,2,\ldots,6\}\}$, $n \in \{1,2,3,4\}$, and the delivery model. We did not vary the average duration

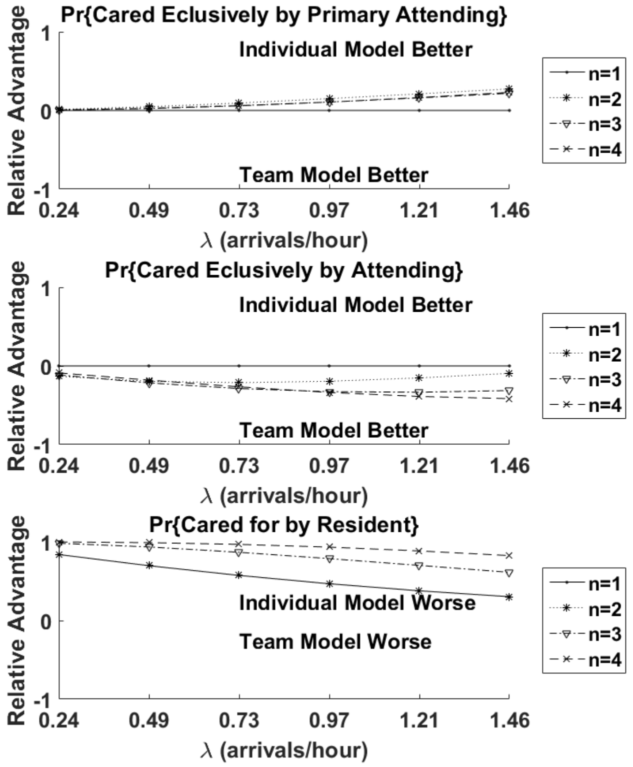

Fig. 13 This figure depicts the one-minus-ratio of the likelihood of a woman being cared for by her primary attending physician (top), any attending physician (middle), and by a resident (bottom), depending upon whether the hospital operates under the team or individual models.

of the final stage of labor, $\frac{1}{\mu}$, as this is a biological process that is patient-specific rather than hospital-specific [47].

8 Conclusion

Labor and delivery is a complex clinical service requiring the orchestration of a diverse set of critical resources in order to provide proper care for mothers and their babies. In this work, we analyzed data from the Department of Obstetrics and Gynecology at a Boston-area hospital, and found that the activity of the L & D floor can be well-approximated using an M/G/∞ queueing model. This model is able to accurately predict the expected number of beds occupied in the various care centers associated with labor and delivery. We also investigated the potential effects of lowering the cesarean section rate on resource utilization across the obstetrics department, and found that hospital managers must prepare for changes to resource needs when altering the C-section rate: specifically, an increase of bed and staffing availability on the labor floor, as well as a decrease in capacity in postpartum. Finally, we compared the benefits and detriments of two common care models (team-based and independent) as they relate to the amount of time labor and delivery patients spend with their physicians, and found that adopting a team-based style can greatly increase the time a labor patient has with a certified attending physician.

References

1. American College of Obstetricians and Gynecologists et al.: Acog committee opinion no. 529: Placenta accreta. Obstetrics & Gynecology **120**(1), 207–211 (2012)
2. Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y.N., Tseytlin, Y., Yom-Tov, G.B., et al.: On patient flow in hospitals: A data-based queueing-science perspective. Stochastic Systems **5**(1), 146–194 (2015)
3. Baum, R., Bertsimas, D., Kallus, N.: Scheduling, revenue management, and fairness in an academic-hospital radiology division. Academic radiology **21**(10), 1322–1330 (2014)
4. Ben-Tal, A., Do Chung, B., Mandala, S.R., Yao, T.: Robust optimization for emergency logistics planning: Risk mitigation in humanitarian relief supply chains. Transportation research part B: methodological **45**(8), 1177–1189 (2011)
5. Borst, S., Mandelbaum, A., Reiman, M.I.: Dimensioning large call centers. Operations research **52**(1), 17–34 (2004)
6. Brandeau, M.L., Sainfort, F., Pierskalla, W.P.: Operations research and health care: a handbook of methods and applications, vol. 70. Springer Science & Business Media (2004)
7. Cameron, A.C., Windmeijer, F.A.: R-squared measures for count data regression models with applications to health-care utilization. Journal of Business & Economic Statistics **14**(2), 209–220 (1996)
8. Cameron, A.C., Windmeijer, F.A.: An r-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics **77**(2), 329–342 (1997)
9. Caughey, A.B., Cahill, A.G., Guise, J.M., Rouse, D.J., of Obstetricians, A.C., Gynecologists, et al.: Safe prevention of the primary cesarean delivery. American journal of obstetrics and gynecology **210**(3), 179–193 (2014)
10. Centers for Disease Control and Prevention et al.: National vital statistics system, birth data (2009)
11. Cochran, J.K., Bharti, A.: Stochastic bed balancing of an obstetrics hospital. Health Care Management Science **9**, 31–45 (2006)
12. Day, T.E., Al-Roubaie, A.R., Goldlust, E.J.: Decreased length of stay after addition of healthcare provider in emergency department triage a comparison between computer-simulated and real-world interventions. Emergency Medicine Journal **30**(2), 134–138 (2013)
13. De Bruin, A.M., Van Rossum, A., Visser, M., Koole, G.: Modeling the emergency cardiac in-patient flow: an application of queuing theory. Health Care Management Science **10**(2), 125–137 (2007)
14. Declercq, E., Menacker, F., MacDorman, M.: Rise in no indicated risk primary caesareans in the united states, 1991-2001: cross sectional analysis. Bmj **330**(7482), 71–72 (2005)
15. Declercq, E.R., Sakala, C., Corry, M.P., Applebaum, S., Herrlich, A.: Listening to mothers iii: Pregnancy and birth. New York: Childbirth Connection (2013)
16. Deneux-Tharaux, C., Carmona, E., Bouvier-Colle, M.H., Bréart, G.: Postpartum maternal mortality and cesarean delivery. Obstetrics & Gynecology **108**(3, Part 1), 541–548 (2006)
17. Ecker, J.L., Frigoletto Jr, F.D.: Cesarean delivery and the risk–benefit calculus. New England Journal of Medicine **356**(9), 885–888 (2007)
18. Ferraro, N.M., Reamer, C.B., Reynolds, T.A., Howell, L.J., Moldenhauer, J.S., Day, T.E.: Capacity planning for maternal–fetal medicine using discrete event simulation. American journal of perinatology **32**(08), 761–770 (2015)
19. Fomundam, S., Herrmann, J.W.: A survey of queuing theory applications in healthcare (2007)

20. Gautam, N.: Analysis of queues: methods and applications. CRC Press (2012)
21. Gerchak, Y., Gupta, D., Henig, M.: Reservation planning for elective surgery under uncertain demand for emergency surgery. Management Science **42**(3), 321–334 (1996)
22. Green, L., Yih, Y.: Queueing theory and modeling. Handbook of healthcare delivery systems pp. 1–22 (2011)
23. Green, L.V., Nguyen, V.: Strategies for cutting hospital beds: the impact on patient service. Health services research **36**(2), 421 (2001)
24. Green, L.V., Soares, J., Giglio, J.F., Green, R.A.: Using queueing theory to increase the effectiveness of emergency department provider staffing. Academic Emergency Medicine **13**(1), 61–68 (2006)
25. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Operations research **29**(3), 567–588 (1981)
26. Hall, R., Belson, D., Murali, P., Dessouky, M.: Modeling patient flows through the healthcare system. In: Patient flow: Reducing delay in healthcare delivery, pp. 1–44. Springer (2006)
27. Hall, R.W.: Handbook of healthcare system scheduling. Springer (2012)
28. Huang, X.M.: A planning model for requirement of emergency beds. Mathematical Medicine and Biology **12**(3-4), 345–353 (1995)
29. Kaplan, R.S., Anderson, S.R.: Time-driven activity-based costing. Available at SSRN 485443 (2003)
30. Kleinrock, L.: Queueing systems, Volume 2: Computer applications, vol. 66. Wiley (1976)
31. Konrad, R., DeSotto, K., Grocela, A., McAuley, P., Wang, J., Lyons, J., Bruin, M.: Modeling the impact of changing patient flow processes in an emergency department: Insights from a computer simulation study. Operations Research for Health Care **2**(4), 66–74 (2013)
32. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1), 79–86 (1951)
33. Kwak, N., Lee, C.: A linear goal programming model for human resource allocation in a health-care organization. Journal of Medical Systems **21**(3), 129–140 (1997)
34. Larson, R.C.: A hypercube queuing model for facility location and redistricting in urban emergency services. Computers & Operations Research **1**(1), 67–95 (1974)
35. Larson, R.C.: Approximating the performance of urban emergency service systems. Operations Research **23**(5), 845–868 (1975)
36. Larson, R.C., Odoni, A.R.: Urban operations research. Prentice-Hall (1981)
37. Little, J.D., Graves, S.C.: Little's law. In: Building intuition, pp. 81–100. Springer (2008)
38. Litvak, E., Long, M.C., Cooper, A.B., McManus, M.L.: Emergency department diversion: causes and solutions. Academic emergency medicine: official journal of the Society for Academic Emergency Medicine **8**(11), 1108–1110 (2001)
39. Liu, S., Liston, R.M., Joseph, K., Heaman, M., Sauve, R., Kramer, M.S., of the Canadian Perinatal Surveillance System, M.H.S.G., et al.: Maternal mortality and severe morbidity associated with low-risk planned cesarean delivery versus planned vaginal delivery at term. Canadian medical association journal **176**(4), 455–460 (2007)
40. Marmor, Y.: Developing a simulation tool for analyzing emergency department performance. Msc Thesis, Technion (2003)
41. McManus, M.L., Long, M.C., Cooper, A., Litvak, E.: Queuing theory accurately models the need for critical care resources. The Journal of the American Society of Anesthesiologists **100**(5), 1271–1276 (2004)
42. Mirasol, N.M.: Letter to the editorthe output of an m/g/∞ queuing system is poisson. Operations Research **11**(2), 282–284 (1963)
43. Molina, G., Weiser, T.G., Lipsitz, S.R., Esquivel, M.M., Uribe-Leitz, T., Azad, T., Shah, N., Semrau, K., Berry, W.R., Gawande, A.A., et al.: Relationship between cesarean delivery rate and maternal and neonatal mortality. JAMA **314**(21), 2263–2270 (2015)
44. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming, 1st edn. John Wiley & Sons, Inc., New York, NY, USA (1994)
45. Rappaport, T.S.: Wireless communications: principles and practice, vol. 2. Prentice Hall PTR, New Jersey (1996)
46. Reid, P.P., Compton, W.D., Grossman, J.H., Fanjiang, G., et al.: Building a better delivery system: a new engineering/health care partnership. National Academies Press (2005)
47. Rouse, D.J., Owen, J., Savage, K.G., Hauth, J.C.: Active phase labor arrest revisiting the 2-hour minimum. Obstetrics & Gynecology **98**(4), 550–554 (2001)
48. Rutberg, M.H., Wenczel, S., Devaney, J., Goldlust, E.J., Day, T.E.: Incorporating discrete event simulation into quality improvement efforts in health care systems. American Journal of Medical Quality **30**(1), 31–35 (2015)
49. Saghafian, S., Austin, G., Traub, S.J.: Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. IIE Transactions on Healthcare Systems Engineering **5**(2), 101–123 (2015)
50. Shah, N.T., Golen, T.H., Kim, J.G., Mistry, B., Kaplan, R., Gawande, A.: A cost analysis of hospitalization for vaginal and cesarean deliveries [282]. Obstetrics & Gynecology **125**, 91S (2015)
51. Shi, P., Chou, M.C., Dai, J., Ding, D., Sim, J.: Models and insights for hospital inpatient operations: Time-dependent ed boarding time. Management Science **62**(1), 1–28 (2015)
52. Taffel, S.M., Placek, P.J., Liss, T.: Trends in the united states cesarean section rate and reasons for the 1980-85 rise. American journal of public health **77**(8), 955–959 (1987)
53. Takagi, H., Kanai, Y., Misue, K.: Queueing network model for obstetric patient flow in a hospital. Health care management science pp. 1–19 (2016)
54. Truven Health Analytics: The cost of having a baby in the united states. Tech. rep., Prepared for Childbirth Connection (2013)
55. Yang, Y.T., Mello, M.M., Subramanian, S., Studdert, D.M.: Relationship between malpractice litigation pressure and rates of cesarean section and vaginal birth after cesarean section. Medical Care **47**(2), 234 (2009)
56. Yates, F.: Contingency tables involving small numbers and the χ 2 test. Supplement to the Journal of the Royal Statistical Society **1**(2), 217–235 (1934)
57. Yom-Tov, G.: Queues in hospitals: Queueing networks with reentering customers in the qed regime. The Technion-Israel Institute of Technology PhD Thesis (2010)
58. Yom-Tov, G.B., Mandelbaum, A.: Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. Manufacturing & Service Operations Management **16**(2), 283–299 (2014)
59. Zeltyn, S., Marmor, Y.N., Mandelbaum, A., Carmeli, B., Greenshpan, O., Mesika, Y., Wasserkrug, S., Vortman, P., Shtub, A., Lauterman, T., et al.: Simulation-based models of emergency departments:: Operational, tactical, and strategic staffing. ACM Transactions on Modeling and Computer Simulation (TOMACS) **21**(4), 24 (2011)