Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

A HIGH-FIDELITY, MACHINE-LEARNING ENHANCED QUEUEING NETWORK SIMULATION MODEL FOR HOSPITAL ULTRASOUND OPERATIONS

Yihan Pan
Zhenghang Xu
Jin Guang
Xinyun Chen
J.G. Dai

School of Data Science
The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong 518172, PRC

Chengwenjian Wang
Xuanming Zhang

School of Mathematical Sciences
Fudan University
Shanghai 200433, PRC

Jingjing Sun

School of Information Management and Engineering
Shanghai University of Finance and Economics
Shanghai 200433, PRC

Pengyi Shi

Krannert School of Management
Purdue University
West Lafayette, IN 47907, USA

Yichuan Ding

Desautels Faculty of Management
McGill University
Montreal, QC H3A 1G5, CAN

Song Wu
Kai Yang

Institute of Urology
Luohu Hospital Group
Shenzhen, Guangdong 518000, PRC
Department of Urology
South China Hospital, Health Science Center
Shenzhen, Guangdong 518116, PRC

Hongxin Pan

Department of Gynaecology
Luohu Hospital Group
Shenzhen, Guangdong 518000, PRC

ABSTRACT

We collaborate with a large teaching hospital in Shenzhen, China and build a high-fidelity simulation model for its ultrasound center to predict key performance metrics, including the distributions of queue length, waiting time and sojourn time, with high accuracy. The key challenge to build an accurate simulation model is to understand the complicated patient routing at the ultrasound center. To address the issue, we propose a novel two-level routing component to the queueing network model and use machine learning tools to calibrate the routing components from data. Our empirical results show that the calibrated model is of high fidelity and yields accurate prediction results for the performance metrics.

1 INTRODUCTION

In this research, we collaborate with a large teaching hospital in Shenzhen, China, *Luohu Hospital*, to develop a high-fidelity simulation model to predict key performance metrics in its ultrasound center. Founded in April 1957, Luohu Hospital is a modern medical treatment center that integrates preventive health care, rehabilitation, research, and teaching. Like many hospitals in Asia, the hospital has a large campus with several buildings that provide primary care, medical examinations, emergency care, and inpatient care. This model is similar to the integrated practice unit (IPU) adopted by several major hospital systems in the US, including Mayo Clinic and Cleveland Clinic (Swensen et al. 2009; Patrnchak 2016). In an IPU, most patients' outpatient *itinerary*, including the primary visit and examinations, can be finished in one day given that primary doctors and examination services are co-located in the same place.

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

Among the examination services that Luohu Hospital provides, the ultrasound examination center is of critical importance (referred to as the *ultrasound center* in the rest of this paper). The demand for its ultrasound center has been rapidly growing in the past few years since its strategic focus shifted to strengthening its obstetrics and gynecology (OB/GYN) specialty. Most OB/GYN patients need an ultrasound exam after the initial consultation with the primary doctor. The ultrasound exam can typically be scheduled and finished within the same day. After getting the exam results, patients return to the primary doctor for the second consultation and get a final diagnosis, concluding their itinerary (a small proportion of them may need a second set of exams before the final diagnosis can be achieved). Preliminary analysis shows that the sojourn time in an ultrasound center is the most time-consuming part of a patient's itinerary. Thus, it is important to understand the operations of the ultrasound center and develop accurate performance prediction on key metrics including the waiting time. Such performance prediction would then provide the basis to identify operational strategies that could reduce the total amount of time these patients need to spend in the hospital for their outpatient itinerary, improving overall patient experience.

In this paper, we develop a high-fidelity simulation model that integrates predictive tools and queueing models to capture the operations in the ultrasound center. The model is calibrated with detailed patient-level data and can provide accurate prediction on the key performance metrics, such as distributions of *time-dependent* queue length, waiting time, and sojourn time. In the following two subsections, we first describe the challenges we met in building the simulation model. Then, we highlight the contribution we made in incorporating salient features in the simulation model that are crucial for accurate prediction.

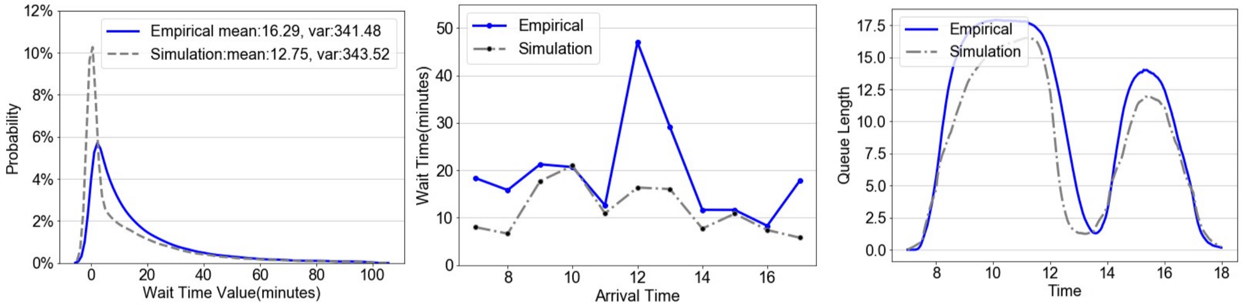
1.1 Challenges

We model the ultrasound center as a multi-class, multi-pool parallel-server queue. Each class corresponds to one type of patient, and each pool of servers corresponds to one or a few examination rooms. Compared with the standard multi-class, multi-pool queue, operations in the ultrasound center have several features, including *patient and server heterogeneity, complicated patient routing*, and *time-dependency* on the server side. We specify each as follows.

First, there is large heterogeneity in the patients and exam rooms (servers). The center serves patients from more than 20 medical specialties, who come with very distinctive test items, often more than one item. This results in a vast combination of service requests and high variances in the service times. Meanwhile, the servers are highly different. Most rooms are capable of performing common tests, but some tests have to be done in rooms with certain equipment. The majority of technicians are known as "general practitioners," who are qualified to perform most of the common tests. Several technicians, known as the "specialists," are specialized to handle certain specific test items. Furthermore, different technicians vary in their skill levels and effectiveness, presenting a large difference in their service speed.

Second, the routing from different patients to different rooms is highly complex. Modeling routing properly turns out to be one of the most important, yet challenging, steps for us to calibrate the simulation model with empirical performance. One source of the challenges comes from the great heterogeneity in patients and servers. Additionally, (i) there is no specific rule for the routing, and nurses who are in charge of the routing use their discretion to assign patients to different rooms; (ii) rooms are not necessarily open for the entire day due to different staffing levels and technicians/doctors taking breaks occasionally (resulting in the room being unavailable). Indeed, existing stylized routing policies such as first-come-first-serve (FCFS) or join-shortest-queue (JSQ) all fail to produce accurate predictions. Figure 1 compares the empirical performance and the simulation output from a conventional multi-class queue and JSQ routing. The simulation output fails to capture the empirical performance, in particular, the waiting time distribution is far off from the empirical one though the average is reasonably close.

Lastly, as in many healthcare systems, there is a strong time-dependency in the arrival rate. Our descriptive analysis also reveals a strong time-dependency in the service speed and server unavailability. We find that the service rate tends to be faster in the late morning than early morning, presenting additional heterogeneity in service speed even within the same server. There are two sources for server unavailability.



(a) Wait time distribution (b) Average wait time (c) Average queue length

Figure 1: The performance of conventional multi-class model with join-shortest-queue (JSQ) routing.

First, not all rooms are open during the entire day. More rooms are open during the peak time (mid-morning) than in the afternoon. Second, some technicians may take a random "break" after serving a patient, resulting in random server unavailability that further complicates the simulation.

In summary, the unique features in the ultrasound center operations present the following challenges to construct a high-fidelity simulation model that can accurately capture empirical performance: (1) the main challenge is to "learn" a proper routing model from the actual practice for the simulation model; and (2) to build a proper routing model, we further need: (2a) patient and server classifications, and (2b) estimating server unavailability. For (2a), we need to strike a balance between having sufficient classes to capture the heterogeneity (in patients and servers) and avoiding too many classes that can lead to estimation inaccuracy. For (2b), if we do not properly account for server unavailability, direct estimation of the routing policy from data may end up with counter-intuitive results. For example, without modeling room unavailability, direct estimation shows that more patients are assigned to rooms with a long queue. However, this result is biased because a room that is unavailable for the next half-hour will have a much longer queue.

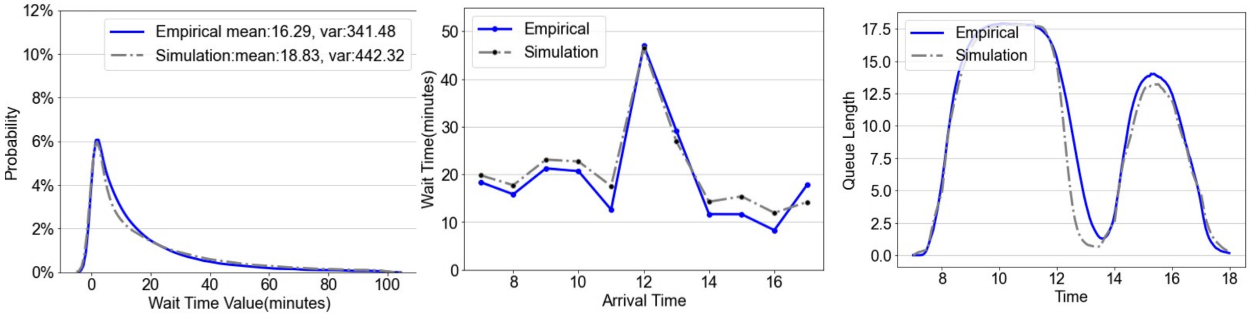
1.2 Previous Works

There is a rich literature in adopting discrete event simulation to investigate the effectiveness of service policies in hospitals (Martin et al. 2020; Feng et al. 2020; Swan et al. 2019; Zhang 2018). In these studies, researchers usually build simplified models to extract and analyze the key factors causing congestion. Many works in this area were based on flexible simulation models. For example, Guo et al. (2004) found that large variations in patients' service type and service duration could greatly affect the simulation performance. Harper and Gamlin (2003) modeled the detailed dependency between oncologists and chemotherapy appointments to optimize the efficiency of patient flow. Lin (2013) presented an analysis on an eye clinic located in Singapore, carefully considering patient behavior in the appointment system (e.g., unpunctuality). These works dealt with a certain class of patients and an appointment mode, which are not in line with our circumstances, where there are high variations in both patient (customer) and doctor (server) sides. From an operations perspective, the ultrasound center is a queueing network of multi-class customers and servers. Among the numerous papers on simulation models in healthcare settings, our work is most relevant to those studying emergency department (ED) and inpatient department for hospitals in North America or others (Wang et al. 2011; Medeiros et al. 2008; Swan et al. 2019). Wang et al. (2011) developed an open queueing network with blocking to capture the situation when a hospital is full and patients have to be sent to another hospital. Medeiros et al. (2008) considered an ED with a short-stay unit that provides an observation place to determine whether a patient will be sent to an inpatient unit or not. Comparing to these models, our ultrasound center presents several unique features as discussed above, which prevent us from deploying off-the-shelf simulation models directly.

Directly predicting waiting times with machine learning tools is an alternative to our model-based approach. For example, Hermanto et al. (2018), Limlawan and Anussornnitisarn (2020) apply artificial neural networks to predict waiting time using patient-level features. Our approach is different from this line of research, as it is a “white box” model and captures the underlying queueing dynamics. Capturing the queueing dynamics is important for counterfactual study, where one needs to evaluate the performance changes under different operational strategies (e.g., capacity changes). Black-box machine learning models often lack the ability to capture the complex dependence between performance metrics and system parameters.

1.3 Main Contribution

We develop a high-fidelity simulation model that integrates machine learning tools and queueing models in a novel manner, instead of using either alone. We leverage machine learning as predictive tools to calibrate key modeling components in our stochastic queueing network. This integrated model produces accurate performance prediction enhanced by machine learning, while the queueing-based model allows us to retain desirable properties such as being structural and interpretable. Figure 2 demonstrates the output from our high-fidelity simulation model, which can accurately capture various key performance metrics, including the waiting time distribution that was far off from the conventional model as shown in Figure 1. More validation results are in Section 4.



(a) Wait time distribution

(b) Average wait time

(c) Average queue length

Figure 2: The performance of our high-fidelity simulation model. The horizontal axis of (b) (c) ranges from 7 to 17, which points to the hourly index in one day. For the other figures in our paper, the time axis has the same meaning as we show here.

We highlight our contribution to the literature as follows:

Two-level patient routing component. To address the main challenge of building a proper routing component, we combine queueing structures with data to analyze routing behaviors in practice, rather than imposing stylized routing policies. Specifically, we develop a two-level hierarchical routing policy. At the first level, we decided which group of exam rooms to route a patient. Then, at the second level, we decide which room, within the selected group from the first level, to route the patient. At each level, we apply machine learning methods to learn the routing rule from data. This two-level structure allows us to separate the key determinants in the routing decisions: the patient’s examination items (medical factor) in the first level, and the queue length and the room availability (operational factor) in the second level.

Clustering-based patient and room classification. We leverage clustering algorithms to classify patients and servers. On the patient side, the clustering is based on the number and duration of examination items that a patient has to do and whether these items need be performed in a specific server. On the server side, the clustering is based on the empirical routing probability, opening time, and service time distributions. These clustering algorithms provide us the flexibility to choose a proper number of classes and the desired proximity on key features for calibrating the model (e.g., service time).

Estimating server availability. As discussed, the server availability is affected by (i) the staffing (room open/close) policy and (ii) the technician taking a random break after serving the last patient. For (i), we calibrate a time-varying room open policy from data. For (ii), we analyze the sequence of patients, who received exams in the same room, on their arrival, service start, and service end timestamps. We identify two types of gap times. The first type is the break time, defined as the duration between the service-end time of the previous patient and the service-start time of the next patient in a busy period. The second type is on the patient side – walking time, defined as the duration between the arrival time and service-start time of the next patient after the previous patient has finished service. It turns out incorporating these two gap times in the simulation can further improve the prediction accuracy. See more details in Section 3.4.

We conclude this section by discussing the relevance in decision-making and the outline of this paper. **Relevance of our simulation model in decision making.** The simulation model we build in this research provides an accurate prediction on key performance metrics, which can be used to provide patients with an estimated waiting time on finishing the ultrasound exam. Such "delay announcement" is often used in call center operations and has been shown to greatly improve customer experience. Moreover, this simulation model provides a high-fidelity platform to evaluate the impact of different operational changes. For example, in (Chen et al. 2021), the authors develop a patient scheduling policy that explicitly accounts for patient revisits in a day. The solution algorithm relies on the estimation of the sojourn time distribution at the examination center (ultrasound for OB/GYN patients) as a key input to search for the optimal schedule. The empirical estimates of the sojourn time can be possibly used as an input for the initial iteration of the solution algorithm. However, in later iterations, the sojourn time distribution needs to be re-evaluated when all primary doctors' offices use the new schedule. This will cause a major shift to the arrival pattern to the ultrasound center and hence, change the sojourn time. In this case, one cannot use empirical estimates anymore. Having a high-fidelity simulation model that allows the evaluation of the sojourn time distribution under the new arrival pattern is critical. Compared with pure black-box machine learning tools, our queueing-based model can better capture the complicated dependence between arrival patterns and system congestion.

Paper outline. In the rest of the paper, we first give an overview of the ultrasound center and present descriptive statistics in Section 2. We also describe the classification of patients and rooms in Section 2. Then we specify the details of our simulation model in Section 3. We present the simulation results that validate our model in Section 4. We conclude the paper in Section 5.

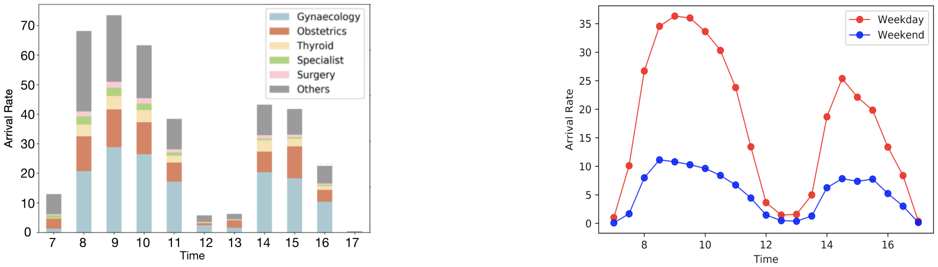
2 DATA AND DESCRIPTIVE ANALYSIS

Patients who request tests from the ultrasound center come from various departments (medical specialties) and their request examination items vary greatly. There are 189 different ultrasound examination items out of the 175702 total records in our 1-year data in 2018. The top 10 examination items account for over 75% of the tests with more than 2000 occurrences in a year. Table 1 lists the top 10 examination items. Section A in the online appendix (Pan et al. 2021) details our process of merging different data sources and the key attributes/timestamps in the data.

Table 1: Top 10 ultrasound examination items.

Index	Inspection Items	Proportion
A	Transabdominal and transvaginal color Doppler ultrasound	32.3%
B	Color Doppler ultrasound examination of uterine attachment	8.79%
C	Fetal Level 2 Examination	7.07%
D	Superficial color Doppler ultrasound: breast and axillary lymph nodes	6.32%
E	Color Doppler ultrasound: liver, gallbladder, spleen, pancreas and portal vein system	5.26%
F	Intracavitary three-dimensional color ultrasound imaging	3.96%
G	Four-dimensional Level 2 fetal examination	3.39%
H	Nuchal Translucency Screening	3.25%
I	Superficial color Doppler ultrasound: thyroid and lymph nodes	3.20%
J	Cardiac Color Doppler Examination Package 2	3.02%

Figure 3 plots the hourly arrival rates to the ultrasound center, which shows a time-dependent pattern. Figure 3a decomposes the hourly arrival rates into the corresponding rate from each department such as OB/GYN and Surgery. The peak arrival occurs in the morning (8-11 AM) each day. In addition to the time-of-day pattern, the arrivals also demonstrate a day-of-week pattern. The two curves in Figure 3b show the hourly arrival rates averaged over all weekdays and weekends, respectively. Next, we discuss how to classify the patients and the servers (examination rooms).



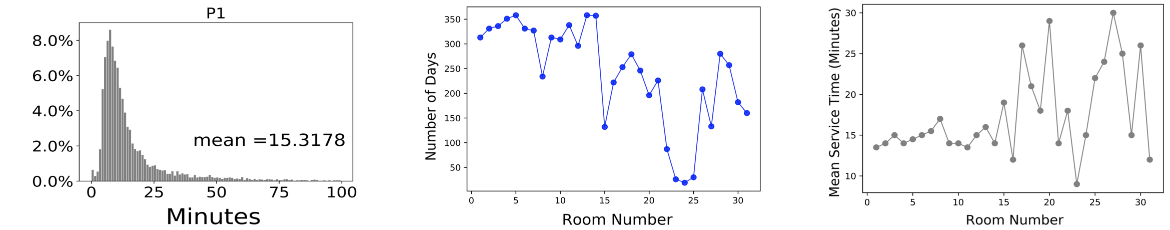
(a) Arrival rate by department (b) Arrival rate on half hour by weekday/weekend

Figure 3: Arrival rate for ultrasound center

2.1 Classification of Examination Item and Room Types

To classify patients, we note that, in addition to a large number of examination items, a significant number of patients come to the ultrasound center with more than two examination items. Thus, if we were to classify patients according to their required examination items, it would lead to too many classes and cause estimation errors due to the small sample size in each combination. To avoid this issue, we leverage clustering algorithms to classify patients, based on the mean and standard deviation of their service time and the frequency of going to different rooms. We further separate the 17.48% of patients who need multiple examination items into a different category. Section D.1 in the online appendix (Pan et al. 2021) details the total six groups of examination items classified from the clustering algorithm (labeled as P1 through P6 in the rest of the paper). Figure 4a plots the empirical service time distribution for the first group (P1). The empirical service time distributions for other groups are provided in Section D.1 in the online appendix.

For the server-side, the ultrasound center at Luohu Hospital has 32 rooms that are equipped with different kinds of exam machines. Additionally, not every room is open during the entire working hour in a day. Figure 4b shows the number of open days, out of 365 days in a year, for each of the 32 rooms. Figure 4c shows the mean service time for each of the 32 rooms. We can see that both the number of open days and service time vary greatly among different rooms. Again, we leverage clustering algorithms to classify the rooms into a handful of types. For this part, we use the Hierarchical Clustering Algorithm base on the following room features: (i) number of open days in one year, (ii) mean service time, which is related to service capacity, and (iii) proportion of examination items performed, which is related to routing. Table 2 shows the four room types classified from the clustering algorithm, along with the room ID each type contains and the major types of examination items performed. Figure 5 plots the number of patient records from different departments in our 1-year data that are routed to each of the four room types, with the width of the line scaled proportionally to reflected the different volumes.



(a) Service time distribution for group 1 (P1) exam type (b) Number of opening days in a year (c) Mean service time for each room

Figure 4: Exam item and room features for ultrasound center

Table 2: Classification of patients by examination items.

Room Type	Room Number	Patient Type
R1	1,2,3,4,5,6,7,8,10,12	P2,P3,P5,P6
R2	9,11,13,14,29	P1,P2,P3,P5,P6
R3	17,18,20,26,27,28,30	P2,P3,P4,P5,P6
R4	15,16,19,21,22,23,24,25,31	P2,P3,P5,P6

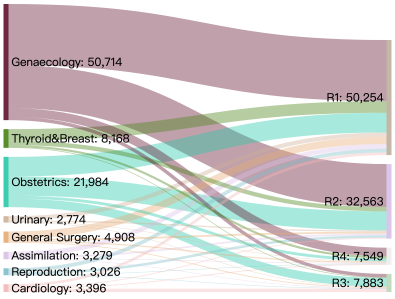


Figure 5: Patient routing process

3 SIMULATION MODEL

In this section, we introduce our high-fidelity simulation model to capture the patient flow and operations of the ultrasound center. This model is a multi-class, multi-pool parallel-server queueing system. The model contains four major components: (1) patients arriving, (2) checking room availability, (3) routing patients to the specific room, and (4) patients receiving service and leaving the system after service completion. Figure 6 shows the flowchart of the simulation process. The main novelties of our simulation model include developing a two-level, machine-learning-based routing policy in component (3), and adding random break time to the patient service process in (4). Thus, we will briefly describe components (1) and (2) first, and then specify components (3) and (4) in the following subsections.

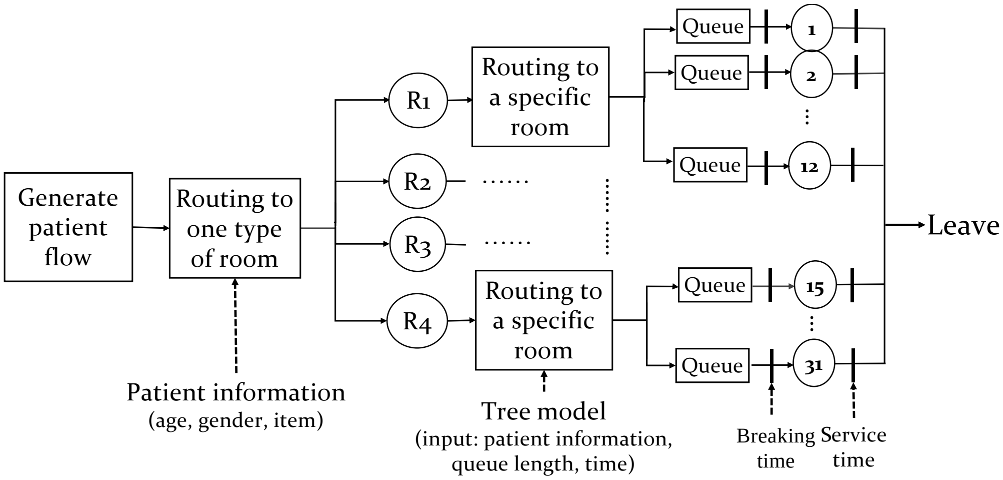


Figure 6: Flowchart of queueing simulation at ultrasound station.

3.1 Arrivals

Based on the examination item types introduced in Section 2.1, we classify patients according to a combination of three attributes: item type, age group, and gender. The classification gives us in total 75 distinct patient classes (see details in Section D in the online appendix (Pan et al. 2021)). We model the arrival process for each patient class as an independent time-nonhomogeneous Poisson process, with the hourly arrival rates for each estimated empirically from data.

3.2 Room Availability

Our initial model that uses the average proportion of opening hours/days to sample the room availability status performs poorly. Further analysis shows that the room opening pattern strongly correlates with individual days. For our final model, we estimate the patterns of room opening hours for each day. In the simulation, we first sample a day from our dataset at random at the beginning of each simulated day. We then use the estimated opening pattern of that day to determine the room availability in the corresponding simulated day.

3.3 Routing

As discussed in Section 1.1, models with stylized routing policies perform poorly in capturing the empirical performance, e.g., see Figure 1a. We develop a novel two-level hierarchical routing policy, with each level being estimated from data directly. In the first level, we decide on which one of four types of rooms to route the patient to. Then at the second level, we decide on which room within that type to route the patient. Correspondingly, we constructed two prediction models to learn the routing decisions at two levels. Our analysis shows that adopting the two-level structure is important in calibrating the simulation model, specifically, separating the second level from the first level. This is because assigning patients to different types of rooms (captured by the first level) is mainly from medical considerations. Within a given type of room, the specific room to assign the patients will then largely depend on operational features such as queue length and room availability. With the two-level structure, we are able to differentiate the two types of considerations and obtain an accurate prediction of the routing decisions.

3.3.1 First Level: Routing to One Type of Rooms

The first level in our routing component is to assign an incoming patient to one of four room types (R1, R2, R3, R4). In this part, we formulate it as a classification problem, with the true class label as the actual room type from the data. We then fit different machine learning models to predict this label, including multinomial logit model, random forest, gradient boosting tree, and neural network. We compare their prediction accuracy performance and pick the one with the best out-of-sample prediction power. The random forest model is found to be the best among all models we test. Features used in the random forest model are summarized in Table 3.

Table 3: Table of features for routing model: first level in the routing component.

feature name	explanation
age	age of patient
item type	classified examination item types, encoded as P1 to P6 in Sec 2.1
arrival hour	hour of the day when patient arrives
weekday	day of the week when patient arrives
queue length	the number of patient in queues when patient arrives
numServer	number of open test rooms when patient arrives

Notice that in this first level, the queue length is generated by summing up all queues in each type of room. Similarly, “numServer” – the number of open rooms, is also computed among all rooms contained in each type. The prediction accuracy of this random forest model is provided in Section B of the online appendix (Pan et al. 2021).

3.3.2 Second Level: Routing to a Specific Room

After selecting a room type for an arriving patient, we decide on which room within the type to route the patient to. Similarly, we consider it as a supervised learning problem, but now the classification labels are changed to individual rooms. We fit one routing model for each room type, which leads to a total of four models. Similar to the first level, we test different classification methods for the second level. Again, random forest is shown to be the best. The input features to this second-level random forest model are mostly the same as Table 3, except following two differences: (1) *queue length*: queue length refers to the number of waiting patients at an individual test room, instead of the entire room type. (2) *numServer*: in the first level we adopt numServers to represent room open status within one type. Here we use a binary indicator that shows whether an individual room is open or not in the current hour. Section B of the online appendix summarizes the hyperparameters chosen in the two levels.

3.4 Service Process

The service time of each patient is sampled empirically from historical patient records with the same room, hour-of-day, and patient type combination. A patient leaves the system after getting service in the system. We illustrate two additional features in generating service times in our simulation model: random *server break time* and *patient walking time*.

From both the data and our field observation in the hospital, we find that patients are not always called into the room as soon as the last patient in the queue leaves. It is possible that technicians need to take time to organize materials, adjust the machines, take a short break, or on some occasions, leave the room to deal with urgent situations. To model this server break time, we analyze a sequence of patients served in the busy period of each room, and take the gap between two consecutive patients as the break time. That is, the duration between the service-end time of the previous patient and the service-start time of the next patient. We then fit distributions of the break time for each of the four room types and for each hour of the day. Details of the break time distributions are in Section E of the online appendix (Pan et al. 2021). Similarly, patients need to spend some time walking to the exam room after being routed there. To estimate this walking time, we use the duration between a new patient's arrival time and his/her service-start time in an idle period. We also fit this distribution for each of the four room types and each hour of the day; see Section E in the online appendix.

4 MODEL VALIDATION

4.1 Validation on Routing

We first discuss the validation of the two-level routing component. We compare the average number of patients routed to each room type (first level) or each room (second level) with the empirical counterparts. Figure 7 compares for the first level. We can see that the predicted number of patients routed to each room type matches closely with the empirical value in each hour. The absolute difference in the hourly arrival rate is between 0.04 and 2.08 across different hours and room types. The relative difference is between 0.60% and 22.55%, with the relative difference defined as the absolute difference between simulation and empirical values divided by the empirical value.

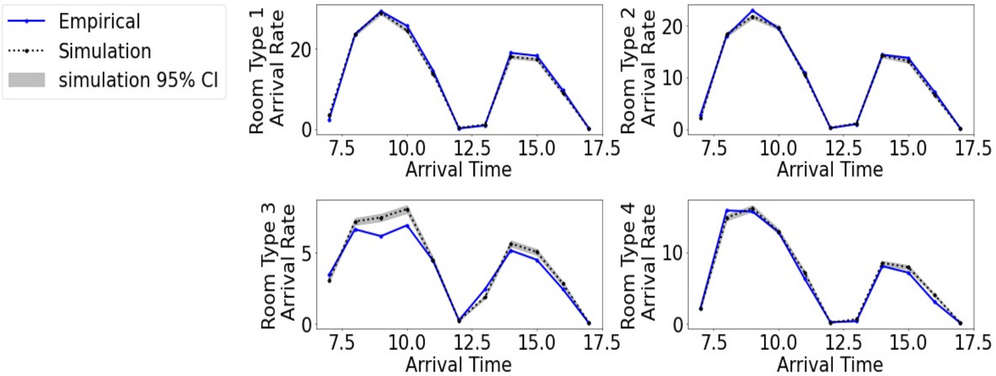


Figure 7: First level: number of routed patients from four types.

For the second level, Figure 8 compares the empirical and predicted hourly number of patients routed to each Type 1 room. The figures for the other three room types are available in Section B of the online appendix (Pan et al. 2021). Again, we can see the predicted values match well with the empirical values. For room type 1, the absolute difference in the hourly arrival rates is between 0 and 0.53 across different hours, with the relative difference between 0% and 15.31%.

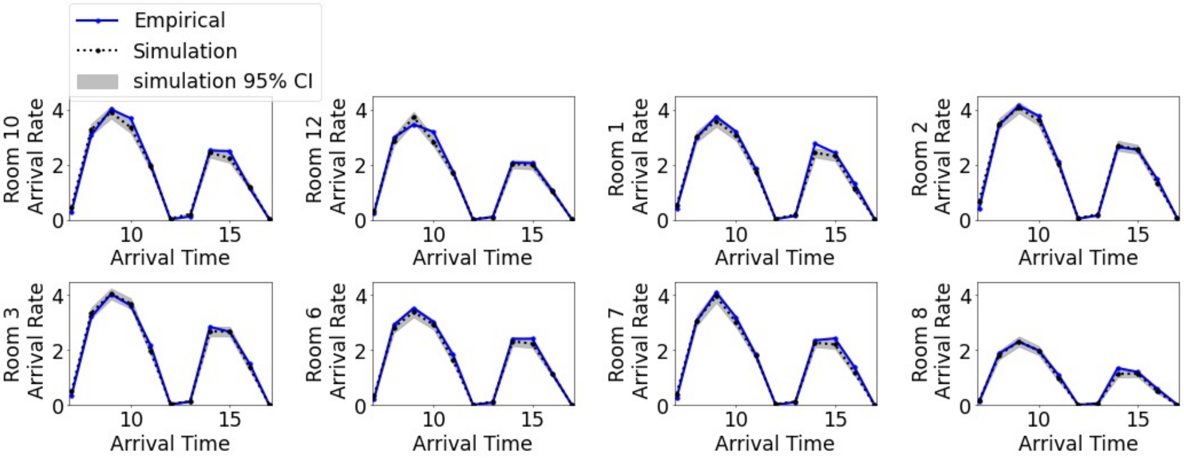


Figure 8: Second level: number of routed patients from each room of room type 1.

4.2 Validation on Key Performance Metrics

Figure 1 in the introduction shows the performance comparison between the empirical performance and the simulation output from the conventional multi-class queue with JSQ routing. The middle column in Table 4 summarizes statistics on the differences in the key performance metrics. It is worth noting that while the differences (absolute and relative) in the mean waiting time are moderate, the differences in the mean queue length and waiting time distributions are significant, as observed in Figure 1 and Table 4. The largest relative difference in the hourly mean queue length can be as high as 32%.

For our high-fidelity simulation model, Figure 2 shows a similar set of performance comparisons between the simulation output and the empirical performance. The last column in Table 4 summarizes statistics on the differences in the key performance metrics in this high-fidelity model. The improvement on the waiting time distribution is the largest: the Kolmogorov-Smirnov difference (KS diff) is reduced from 0.371 to 0.058. The difference in the mean queue length also shows a significant reduction (from 2.23 to 0.82), with the largest relative difference in the hourly queue length reduced to 20% (from 32% previously). The mean waiting time also shows some improvement, with the difference reduced from 3.11 minutes to 2.40 minutes. Overall, our high-fidelity model produces much more accurate performance predictions compared to the conventional model.

Table 4: Comparison on simulation output. The numbers in the parenthesis for the first two rows are the corresponding relative differences.

	Conventional model	High-fidelity model
Diff in mean queue length (patient)	2.23 (16.49%)	0.82 (6.73%)
Diff in mean waiting time (min)	3.11 (21.25%)	2.40 (19.34%)
KS diff in waiting time distribution	0.371	0.058

5 CONCLUSION

For many health service systems, conventional modeling-based analysis has its limitation as being inflexible in capturing complicated behavioral features, e.g., the routing decisions in our studied setting. This calls for the need of developing high-fidelity simulation models that can capture these complicated features and make more accurate predictions. Taking our healthcare partner as an example, the salient features in the operations of their ultrasound center would make a conventional modeling-based analysis infeasible. These features include patient and server heterogeneity, time-dependency, complicated patient routing, and server vacation. Thus, it is imperative for us to develop a high-fidelity model, instead of a stylized model, to predict the performance of the ultrasound service system in case the hospital manage wish to change some operations component in the system. In particular, our simulation model can provide direct help for the hospital manager to predict system performance under different scheduling and staffing policies and thus make better decisions. For instance, our partner hospital is facing a chaotic situation in the obstetrics and gynecology clinics due to the lack of scheduling for those revisit patients who have gone through the ultrasound center for examination. Our simulation model provides a highly accurate generator of sojourn times for each patient in the ultrasound center, which is subjective to the scheduling decision, and hence enables us to develop simulation optimization methods to improve the scheduling policy for both appointment patients and revisit patients (Chen et al. 2021). Moreover, many of the features that we identified and modeled in this paper, such as time-varying arrival and service rates, patient and server heterogeneity, server vacation, etc., are likely to exist in other hospitals. Thus, the methods we proposed in developing the simulation model have the potentials to be generalized to study other ultrasound centers and service systems.

One future research direction is to develop an analytically tractable model that still captures some of the critical features in our simulation model. This would allow us to further understand the underlying driving forces of the ultrasound system and develop generalizable managerial insights. Another potential research direction is to develop a decision support system, based on this simulation model, that could facilitate the hospital manager to make real-time scheduling decisions such as the one in (Chen et al. 2021).

REFERENCES

Chen, X., J. Dai, Y. Ding, P. Shi, and L. Sun. 2021. "Joint Appointment and Reentry Scheduling: Mitigating Onsite Overcrowding in Outpatient Services". *working paper*.

Feng, H., Z. Li, M. M. Alvarado, C. M, and C. M. Colón-Morales. 2020. "A simulation study of outpatient surgery clinic with stochastic patient re-entrance". In *Proceedings of the 2020 Winter Simulation Conference, edited by Bae, K.; Feng, B.; Kim, S.; Lazarova-Molnar, S.; Zheng, Z.; Roeder, T.; and Thiesing, R., editor(s)*, pages 910–921. Piscataway New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Guo, M., M. Wagner, and C. West. 2004. "Outpatient clinic scheduling-a simulation approach". In *Proceedings of the 2004 Winter Simulation Conference, edited by Smith, J.; Peters, B.; Ingalls, R.; and Rossetti, M., editor(s)*, Volume 2, pages 1981–1987. Piscataway New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Harper, P. R., and H. Gamlin. 2003. "Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach". *Or Spectrum* 25(2):207–222.

Hermanto, R. P. S., A. Nugroho et al. 2018. "Waiting-time estimation in bank customer queues using RPROP neural networks". *Procedia Computer Science* 135:35–42.

Limlawan, V., and P. Anussornnitisarn. 2020. "Development of waiting time predictor based Artificial Neural Network". In *IOP Conference Series: Materials Science and Engineering*, Volume 847, 012026. IOP Publishing.

Lin, C. 2013. "A decision-support simulator for improving patient flow and increasing capacity at an eye outpatient department". In *2013 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, 21–26. IEEE.

Martin, J., J. Kiel-Locey, K. Shehadeh, S. D. Saini, and J. E. Kurlander. 2020. "Integrated simulation tool to analyze patient access to and flow during colonoscopy appointments". In *Proceedings of the 2020 Winter Simulation Conference, edited by Bae, K.; Feng, B.; Kim, S.; Lazarova-Molnar, S.; Zheng, Z.; Roeder, T.; and Thiesing, R., editor(s)*, pages 934–943. Piscataway New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Medeiros, D. J., E. Swenson, and C. DeFlitch. 2008. "Improving patient flow in a hospital emergency department". In *Proceedings of the 2008 Winter Simulation Conference, edited by Jefferson, T.; Fowler, J.; Mason, S.; Hill, R.; Moench, L.; and Rose, O., editor(s)*, pages 1526–1531. Piscataway New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Pan, Y., Z. Xu, J. Guang, C. Wang, X. Zhang, J. Sun, X. Chen, J. Dai, Y. Ding, P. Shi, H. Pan, K. Yang, and S. Wu. 2021. "ONLINE APPENDIX to A High-fidelity, Machine-learning Enhanced Queueing Network Simulation Model for Hospital Ultrasound Operations". Technical report, Chinese University of Hong Kong, Shenzhen. available at https://github.com/ZhenghangXu-CUHKSZ/Luohu-Ultrasound-Simulation/blob/main/simulation%20report.pdf.

Patrnchak, J. M. 2016. "Implementing servant leadership at cleveland clinic: A case study in organizational change". *Servant Leadership: Theory & Practice* 2(1):3.

Swan, B., O. Ozaltin, S. Hilburn, E. Gignac, and G. McCammon. 2019. "Evaluating an emergency department care redesign: a simulation approach". In *Proceedings of 2019 Winter Simulation Conference, edited by Son, Y.; Hass, P.; Mustafee, N.; Rabe, M.; Bae, K.; Szabo, C.; and Lazarova-Molnar, S.,editor(s)*, pages 1137–1147. Piscataway New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Swensen, S. J., J. A. Dilling, D. S. Milliner, R. S. Zimmerman, W. J. Maples, M. E. Lindsay, and G. B. Bartley. 2009. "Quality: the Mayo Clinic approach". *American Journal of Medical Quality* 24(5):428–440.

Wang, X., X. Shen, and X. Liu. 2011. “Improving patient flow at hospital emergency services—A simulation study”. In *Icsssm11*, 1–6. IEEE.

Zhang, X. 2018. “Application of discrete event simulation in health care: a systematic review”. *BMC health services research* 18(1):1–11.

AUTHOR BIOGRAPHIES

YIHAN PAN is a senior student in the School of Data Science at The Chinese University of Hong Kong, Shenzhen. Her email address is yihanpan@link.cuhk.edu.cn.

ZHENGHANG XU is a senior student of Statistical Science in the School of Data Science at The Chinese University of Hong Kong, Shenzhen. His email address is zhenghangxu@link.cuhk.edu.cn.

JIN GUANG is a first-year Ph.D. candidate in the School of Data Science at The Chinese University of Hong Kong, Shenzhen. His email address is jinguang@link.cuhk.edu.cn.

JINGJING SUN is a senior student from the Department of Information Management and Information System at the Shanghai University of Finance and Economics. Her email address is 2017111642@live.sufe.edu.cn.

CHENGWENJIAN WANG is a senior student of Mathematics and Applied Mathematics in the School of Mathematical Sciences at the Fudan University. His email address is cwjwang17@fudan.edu.cn.

XUANMING ZHANG is a senior student of Mathematics and Applied Mathematics in the School of Mathematical Sciences at the Fudan University. His email address is 17307130065@fudan.edu.cn.

XINYUN CHEN is an Assistant Professor in the School of Data Science at The Chinese University of Hong Kong, Shenzhen. Her research interests include applied probability, stochastic simulation, queueing theory, and reinforcement learning. Her email address is chenxinyun@cuhk.edu.cn.

JIANGANG DAI is a Presidential Chair Professor in the School of Data Science at The Chinese University of Hong Kong, Shenzhen. Professor Dai received his Ph.D. in Mathematics from Stanford University. He is also on the faculty of Cornell University, where he is the Leon C. Welch Professor of Engineering in the School of Operations Research and Information Engineering. His main research direction is on stochastic processing networks. His email address is jimdai@cuhk.edu.cn.

YICHUAN DING is an Assistant Professor in the Desautels Faculty of Management, McGill University. His research interests include optimization, queueing, and statistics, as well as their applications in public sectors. His email address is daniel.ding@mcgill.ca.

PENGYI SHI is an Assistant Professor of Operations Management in the Krannert School of Management, Purdue University. Her research focuses on building data-driven, high fidelity models to support decisions making under uncertainty in healthcare and service systems. Her email address is shi178@purdue.edu.

HONGXIN PAN is a gynecologist at the Third Affiliated Hospital of Shenzhen University (Luohu Hospital Group). His email address is panhongxin@cuhk.edu.cn.

KAI YANG is in charge of the Information System at the Third Affiliated Hospital of Shenzhen University (Luohu Hospital Group) and the South China Hospital, Health Science Center. His email address is yangkai@zxbiomed.org.

SONG WU is a urologist at the Institute of Urology of the Third Affiliated Hospital of Shenzhen University (Luohu Hospital Group) and the department of urology of the South China Hospital, Health Science Center. His email address is wusong@szu.edu.cn.