12-2013

Improving patient length-of-stay in emergency department through dynamic queue management

Kar Way TAN
Singapore Management University, karway.tan.2007@smu.edu.sg

Hoong Chuin LAU
Singapore Management University, hclau@smu.edu.sg

Francis Chun Yue LEE
Khoo Teck Puat Hospital

Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

IMPROVING PATIENT LENGTH-OF-STAY IN EMERGENCY DEPARTMENT THROUGH DYNAMIC QUEUE MANAGEMENT

Kar Way Tan
Hoong Chuin Lau

Singapore Management University
80 Stamford Road
Singapore 178902, SINGAPORE

Francis Chun Yue Lee

Khoo Teck Puat Hospital
90 Yishun Central
Singapore 768828, SINGAPORE

ABSTRACT

Addressing issue of crowding in an Emergency Department (ED) typically takes the form of process engineering or single-faceted queue management strategies such as demand restriction, queue prioritization or staffing the ED. This work provides an integrated framework to manage queue *dynamically* from both demand and supply perspectives. More precisely, we introduce intelligent dynamic patient prioritization strategies to manage the demand concurrently with dynamic resource adjustment policies to manage supply. Our framework allows decision-makers to select both the demand-side and supply-side strategies to suit the needs of their ED. We verify through a simulation that such a framework improves the patients' length-of-stay in the ED without restricting the demand.

1 INTRODUCTION

Experiencing long wait and hence patient's length-of-stay (LOS) at an Emergency Department (ED) is a common problem faced by many hospitals around the world. Managing wait times at ED is generally challenging because ED deals with patients without appointment and with a large variety of illnesses and large variance in the time required to diagnose and treat. Such characteristics make managing the patient queue and planning for resources at the ED complex. Having a fixed queueing policy such as FIFO and static resource schedule is inflexible and unable meet the varying demand or adapt to operational deviation from expected service time of a patient (a patient requires more attention than expected).

There is a rich literature related to the problem managing wait-time or reducing patient's LOS, but most approaches are typically single-faceted: One either attempts to solve the problem from the demand end by managing the patients arriving at the ED or from the supply of resources (e.g., through staffing of doctors and nurses, bed allocation). Managing demand comes in the form of either restricting arrival of patients such as ambulance diversion (Pham et al. 2006) or in managing the patient flow (King, Ben-Tovim, and Bassham 2006; Chakravarthy 1992) in the ED process. Managing the patient flow includes use of priority queue and fast-track lane. The supply perspective focuses on staffing problems. Some approaches use simulation to design proactive staffing policy (Cabrera et al. 2012; Samaha, Armel, and Starks 2003); some use analytical methods with considerations of time-varying arrival (Green, Kolesar, and Whitt 2009; Yom-Tov 2010; Feldman et al. 2008; Jennings et al. 1996). Others such as Marmor et al. (2009) and Thorwarth and Arisha (2012) use simulation as a backend engine to provide decisions to real-time (dynamic) staffing. Queues analysis in healthcare has also been studied extensively. We refer our reader to comprehensive survey on queuing analysis from Green (2006). Other examples to improve the ED via simulation can be found in Komashie and Mousavi (2005) and Gunal and Pidd. (2006). However, to our best knowledge, there is no study that integrates queue management from both demand and supply perspectives.

In this work, we study the ED process as a complex system. We draw inspiration from Traffic/Civil Engineering and Economics domains where the performance of the system is managed via control from both demand and supply perspectives. A key model in Traffic Engineering is the DynaMIT system http://mit.edu/its/dynamit.html (Ben-Akiva et al. 1998). It is a real-time model of traffic condition on roads (supply) and driver's travel requirements (demand) to predict and generate strategies to guide drivers towards optimal route decisions. In Economics, an example of managing queue by adjusting demand and supply is shown in Mendoza, Sedaghat, and Yoon (2011). The methods used in both of these domains are not directly applicable to the context of ED. The demand for ED services is unpredictable and uncontrollable as patients arrive without scheduled appointments. The ED resources (manpower, physical space) are limited to cope with the surges and variations of patient arrival. Measures such as scheduling of manpower to complement the peaks and troughs of patient volume have limited success to deal with variability. As such, we need to manage the demand and supply in ED in a more robust manner.

To this end, we propose a framework which include two major queue management components: (i) demand-side queue management via dynamic priority queue (Tan, Wang, and Lau 2012) and (ii) supply-side queue management via dynamic resource adjustment strategies (Tan, Tan, and Lau 2013). Our key contribution is the ability to seamlessly integrate strategies from both perspectives. We show in an experimental simulation prototype that this integrated queue management framework allows the strategies from both perspectives to work together and complement each other.

2 SCOPE OF STUDY

A real life study is conducted in an ED of a selected hospital. This ED, based on a national guidelines, patients are classified (during triage) into four acuity categories, namely P1, P2, P3 and P4. The physical layout of the department segregates the patient care work into 2 areas; the *critical care area* manages P1 and P2 patients while the *ambulatory area* (clinic rooms) manages P3 and P4 patients. P3 and P4 are minor emergencies and non emergencies respectively, and represents 70% of this ED's workload. While P3 and P4 are considered lesser emergencies in comparison to P1 and P2, the relatively straight forward nature of their conditions allows the opportunity for workflow improvement and maximising efficiency. The hospital management has empowered the department to strive for a specified desired service level ambulatory area, for example, to serve patients with a length-of-stay (LOS) of 60 mins.

The patient's LOS is the time between the start of registration till the end of the case. It consists for several time segments within, starting with registration and followed by triage, consultation, investigations, observations and treatment till case ends, where a patient is either discharged or admitted to the hospital (Figure 1). The investigations, observation and treatment steps are highly variable and differ greatly between patients. The more consistent steps which all patients experience include registration, triage, consultation and case end. We limit our scope of study to the consultation process in the ambulatory area, applying demand-side queue management on the patient queue to doctors versus supply-side queue management on scheduling the doctors for consultation.

The existing patient queue is managed as first-in-first-out (FIFO) for new patients (with the same acuity) but ad-hoc for re-entrants. The doctor's schedule is static. The static doctor's schedules is planned manually based on perceived understanding about the demand at the ED at various hour of a day, over the entire week or month. Such an ad-hoc policy may not be optimal in minimizing the LOS for all patients. In Tan, Wang, and Lau (2012), we showed that an intelligent patient dispatch (to doctor) policy based on one of the dynamic priority queue strategies enables the hospital to reduce the average patients' length-of-stay. There are three strategies, namely, shortest-consultation-time-first (SCON), shortest-remaining-time-first (SREM) and finally, a mixed strategy (MIXED) based on combination of SCON and SREM. The experimental results showed that SCON has the best performance but present a challenge in implementation readiness as it require the hospital to accurately estimate consultation time of each patient and re-entrant. SREM provides a less significant improvement over FIFO but it is readily implementable. MIXED strategy provides an ability to hedge the risk of inaccurate prediction of consultation time and have a reasonable improvement

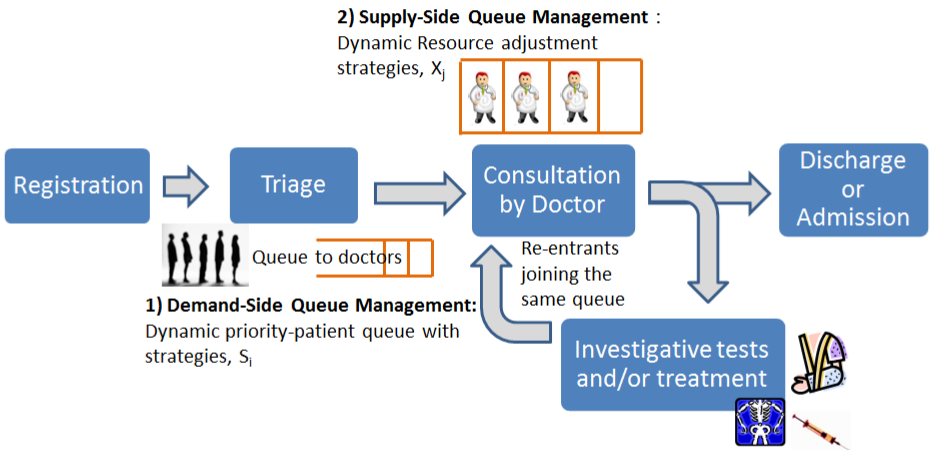


Figure 1: The process with demand and supply perspectives in the ambulatory area.

over FIFO. We presented the analysis for healthcare decision makers to select a strategy that is most appropriate based on their implementation readiness.

In the supply aspect, a static doctor's schedule generally does not react well to uncertainties such as surges in patient arrivals. In Tan, Tan, and Lau (2013), we presented 4 staffing strategies, namely, historical (HIST), dynamic (DYN), historical with optimization (HIST-OPT) and dynamic with optimization (DYN-OPT). The *historical* methods use only historical data and the *dynamic* methods use both historical and real-time data. DYN-OPT also uses symbiotic simulation (Low et al. 2007) model to generate short-termed schedule for the next couple of hours. The DYN-OPT schedule considers both existing queue conditions as well as make projections of future doctors' requirements based on historical understanding about patient's arrival. Our results showed that the optimized strategies are the good performers but incurs higher cost of deployment of doctors. DYN is an easy-to-implement strategy (as it does not require a symbiotic system) with the ability to adapt to actual arrivals of patients. However, DYN can be too reactive. Static strategies cannot react to surges. Again, we presented the analysis for healthcare decision makers to select a strategy that is most appropriate based on their quality improvement appetite and implementation readiness.

3 THE INTEGRATED DYNAMIC QUEUE MANAGEMENT FRAMEWORK

In this section, we present an integrated framework on how both demand and supply strategies can jointly work together in a single framework for healthcare decision makers to make appropriate selection of strategies that best suit their needs. There are three main components in the framework, namely System/data, Analytical Model and Decision-Support Model. The pictorial representation and content of the three components are shown in Figure 2.

3.1 Live Systems and Data

This consists of the various Information Technology (IT) systems and databases that support the live operations of the ED and its supporting departments. For example, the ED processes interacts with processes in laboratory (e.g., blood test) and X-ray departments. The processes in laboratory and X-ray departments use Blood System and X-ray System respectively. The Blood System contains the blood test results and X-ray System contains X-ray test results of a patient who has done both tests. These systems are the physical systems supporting the ED process. The live systems provide data that serve as the input to the analytical model and the decision-support model.

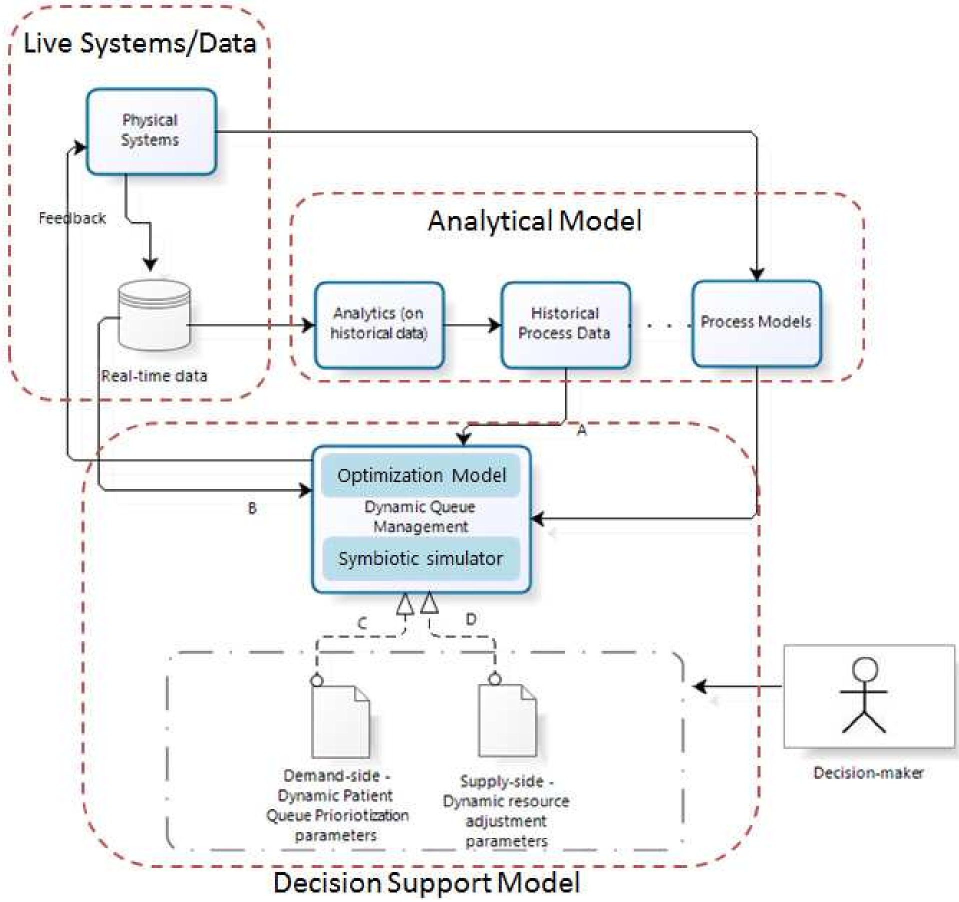


Figure 2: The framework.

3.2 Analytical Model

We build our initial process models and queueing models based on the historical data from the live systems. The components in the Analytical Model are

- Analytics (on historical data): It is a set of activities done on the historical data using a commercial business analytics software (such as SAS®). Historical data is a snapshot of the live data at the point when it is taken. Output is the *Historical Process Data*.
- Historical Process Data: It consists of ED process parameters such as time-varying arrival rates, service rates of triage, service rates of doctors, service rates of treatment/tests, probability of re-entrance.
- Process Models: It is a set of process flow that depicts the real-life process of the ED. An example of a simplified process flow is as in Figure 1.

The Analytical Model provides a set of output that is required by the Dynamic Queue Module (DQM) (more details in 3.3) in order to carry out decision-support functions. The set of output (reflected as input set A in Figure 2 contains:

- $\lambda_f(t)$ - Time-varying arrival rates of new patients at the ambulatory area. Each $\lambda_f(t)$ is defined hourly over a week's horizon. We observe that Sundays and Mondays have a higher volume of patients. In each day, the time-varying pattern is fairly similar. The low peak period of a day is between 1am to 8am daily. The peak period is between 9am to midnight. Midnight to 1am and 8am to 9am periods have moderate demand.

- μ_n - Service rate of the doctors if the patient is a new patient. We assume a homogeneous service rate for all doctors.
- μ_r - Service rate of the doctors if the patient is a re-entrant. The review consultation time is represented by a set of 4 exponential distributions with corresponding probability of occurrence of each patient type. A patient with clean test results usually take a shorter review duration compared to a patient with complex test results.
- δ - Service rate for investigative tests or treatment, which is assumed to be exponential.
- b - Probability of re-entrance. Patients who require test(s) and/or treatment are required to be reviewed by a doctor, hence re-entering the queue.

3.3 Decision-Support Model

The main component of the decision-support model is the Dynamic Queue Management (DQM). It is the engine that embeds the intelligent demand-side and supply-side strategies. In our prototype, we mimic the real world by having the DQM contain a discrete-event simulator that generates the events such as patient arrivals, distributions of registration, triage, consultation, treatment and investigative tests. The DQM also generates plans on the supply of doctors based on the selected supply-side strategy and the patient queue with selected demand-side prioritization strategy. We term this as the *DQM simulator*. In the DQM simulator, there is an *optimization model* and a *symbiotic simulator*. The optimization model performs optimization required for the supply-side HIST-OPT and DYN-OPT strategies. The decision-makers make decisions to select the dynamic queue prioritization parameters and select the dynamic resource adjustment parameters. Strategies selection, associated parameters and real-time data are fed into DQM collectively. DQM generates output (e.g., which patient to serve next, how many doctors are required) and the information is fed back to the live systems for real-time execution. If the supply-side strategy DYN-OPT is selected, symbiotic simulator is used in real-time to perform optimization for the specified planning horizon during the course of simulation. If HIST-OPT is selected, the symbiotic simulator is used as a pre-processing tool to obtain the optimized doctor's schedule. For the purpose of evaluating the HIST-OPT schedule, DQM is used for the second time to obtain the performance. Details on the need of symbiotic simulator and planning horizon are presented in Tan, Tan, and Lau (2013).

DQM receives real-time data (as reflected as input set B in Figure 2) from the live systems to support the demand and supply strategies. Input set B includes $\lambda'_f(t)$ (the real-time time-varying arrival rates of new patient at the ambulatory area), actual queue conditions (the patients in the queue) and the potential re-entrants who are currently receiving treatment or tests.

3.3.1 Demand-side Dynamic Queue Prioritization

We implemented 3 dynamic queue prioritization strategies in DQM. We provide a summary of each strategies here and details of the strategies can be found in Tan, Wang, and Lau (2012). The priorities of all patients in the queue are calculated each time a doctor is available. Note that there is a significant difference between our dynamic priority queue and standard priority queue. Standard priority queue gives a priority to a patient and it stays the same throughout the patient's life-time in the queue system. Our dynamic priority queue changes each time the priority is calculated (i.e., when a doctor is available).

In the Shortest-Consultation-Time-First (SCON) strategy S_1, we rank (give priority to) patients according to their estimated consultation times with the doctor. The intuition for this strategy is based on our observation that some re-entrants have very short estimated consultation time. Let c_k be the estimated consultation time of patient k based on μ_n and μ_r, we use an exponential function (with ρ_1 as the constant parameter to set the gradient of the exponential function) to determine the priority of a patient p_k under S_1, $p_k^{S_1} \leftarrow e^{\frac{\rho_1}{c_k}}$.

In the Shortest-Remaining-Time-First (SREM) strategy S_2, we rank patients according to their remaining times. This is based on the intuition to meet the hospital's desired service quality to serve its patients within a specified duration. When a patient k arrives at the ED, the patient has a common-to-all initial target LOS d_k and an individual elapsed time e_k (initialized to zero) that he has spent in the department. If the patient requires an investigation test and/or treatment, he incurs additional time t_k based on treatment service rate δ. The *remaining time* for the patient, r_k, is then defined as $d_k + t_k - e_k$. Note that we add t_k so as to allow time to be given to patients who genuinely require more time for quality treatment and/or investigation. We assign to each patient a priority value, which tends to a large number when r_k tends to zero, and tends to 1, when r_k is sufficiently large. When r_k is negative, priority of patient becomes linearly negative. Let ρ_1 be the constant parameter to set the gradient of the exponential function when remaining time is positive. Let c be a small value (e.g., 0.1) to ensure priority is very high (only few cases fall into such category). We let $f_c = e^{\frac{\rho_2}{c}}$ when $r_k = [-c, c]$. When $r_k < 0$, we use a negative linear function with a constant slope $m > 0$. Hence the priority function for patient p_k under S_2 is given by the following equation:

$$p_k^{S_2} \leftarrow \begin{cases} e^{\frac{\rho_2}{r_k}} & \text{if } r_k > c, \\ f_c & \text{if } r_k \in [-c, c], \\ f_c - m * r_k & \text{if } r_k < -c. \end{cases}$$

In the mixed (MIXED) strategy S_3, we consider multiple factors in determining the priorities of the patients. Suppose if we have n factors, each having a weight of a_n contributing to the MIXED strategy, then we have the priority of patient p_k under S_3 as $p_k^{S_3} \leftarrow \sum_n a_n . p_k^{S_n}$ where $\sum_n a_n = 1$ and $a_n \varepsilon [0, 1]$.

The decision maker must provide the following parameters (as reflected as input set C in Figure 2) to the demand-side Dynamic Patient Queue Prioritization configuration.

- S_i - Dynamic patient prioritization strategy where i is the selected strategy type.
- LOS_{max} - Hospital's desired service quality in terms of LOS.

3.3.2 Supply-side Dynamic Resource Adjustment

We implemented 4 supply strategies in DQM: 2 are static strategies (HIST, HIST-OPT) and 2 are dynamic strategies (DYN, DYN-OPT). HIST-OPT and DYN-OPT are optimization models. We briefly describe the strategies in this section; details can be found in Tan, Tan, and Lau (2013).

The HIST strategy X_1 uses the historical trends from *Historical Process Data*. In particular, we are referring to the historical arrival rates $\lambda_f(t)$, service rates μ_n, μ_r and δ. Using the data, we apply the Erlang-R method (Yom-Tov 2010) to obtain the time-varying offered load R_x at each service station x. The index $x = 1$ represents the doctor consultation service station and $x = 2$ represents the treatment/test service station.

$$R_x(t) = E[\lambda_x^+(t - S_{x,e})]E[S_x]. \tag{1}$$

λ_x^+ is the aggregated-arrival-rate function to node x, S_x represents the service time at node x and $S_{x,e}$ is a random variable representing the excess service time at node x.

The doctor's staffing at station 1 where patients await for consultation is then determined by substituting the time-varying offered load formula into the square-root staffing formula. The parameter β is chosen according to the steady-state Halfin-Whitt formula (Halfin and Whitt 1981).

$$s(t) = R_1(t) + \beta\sqrt{R_1(t)}, \quad \forall t > 0. \tag{2}$$

In addition, the supply-side model also ensures that the quality at critical care area is satisfied and the number of doctors in the ambulatory area does not exceed the physical constraint of the number of consultation rooms available in the ED. Supposed $S_b(t)$ is the time-varying doctor's requirement in the critical care area, $S_{max}(t)$ is the maximum number of doctors available at time t and $room_{max}$ is the physical constraint in ambulatory area, we have the constraints:

$$S_f(t) \leq S_{max}(t) - S_b(t); \quad (3)$$
$$S_f(t) \leq room_{max}. \quad (4)$$

An optimization variant of HIST is called the HIST-OPT strategy X_2. In HIST-OPT, we add an additional soft constraint:

$$\overline{LOS}(t) \leq LOS_{max} \quad (5)$$

where $\overline{LOS}(t)$ is the average LOS over the period of time t, and LOS_{max} is the hospital's desired service quality to serve the patients below this value.

The intuition behind HIST-OPT is to add a doctor before an LOS violation occurs. Let C_l be the cost of labour of deploying a doctor for a single unit time t and C_d be the cost if the number doctors in time t deviates from $t-1$. The deviation cost is included because it is not desirable to have a schedule where the number of doctors changes too frequently from hour to hour. The resulting objective function is $min \quad C_l \sum_t S_f(t) + C_d \sum_t [S_f(t) - S_f(t-1)]^2$, subject to constraints in Equations (3) to (5) of the model. If no feasible solution is found, the solution with the least violation is returned.

The DYN strategy X_3 is similar to HIST, except it uses the real-time arrivals (hence real-time demand requirements) instead. Let $R'_1(t)$ and $R'_2(t)$ be the real-time offered load based on *actual* arrival rates $\lambda'_f(t)$. The offered load for each service station is calculated as per Equations 1 and 2 and is subjected to constraints in Equations 3 and 4.

The DYN-OPT strategy X_4 is the optimized variant of the DYN strategy. Here, we define a vector (L,H), where L denotes the lead-time and H denotes the time horizon. Between the current time t and lead-time L, there will be no change in staffing. This is to cater for the case that the ED cannot add or remove a doctor instantaneously from the ED. The variable H defines the horizon of re-planning based on what is known currently, e.g., plan for the horizon of 8 hours. The next planning period is then $t+H$. The DYN-OPT objective function is $min \quad C_l \sum_{t=t+L}^{t+L+H} S(t) + C_d \sum_{t=t+L}^{t+L+H} [S(t) - S(t-1)]^2$.

A symbiotic simulator is used to evaluate which doctor's schedule is to be used for the planning horizon such that there is either no LOS violation or with the least violation. We use the concept of a snapshot. A snapshot contains the current queue conditions, doctors' availabilities, patients' statuses and the realised arrival rates. At the start of each planning period, a snapshot is taken and is used with the historical arrival rates for the planning horizon. A heuristic local search algorithm is then applied to find the best schedule. When the schedule has been found, the snapshot is restored and the best schedule is used in the DQM as the schedule for the next horizon.

The decision makers must provide the following parameters (as reflected as input set D in Figure 2) to the supply-side Dynamic Resource Adjustment configuration.

- X_j: Resource adjustment policy j.
- LOS_{max}: Hospital's desired service quality in terms of LOS. This should be consistent with the value set in input set C.
- $room_{max}$: Physical constraint in the ED at the ambulatory area, which corresponds to the maximum number of consultation rooms in the real-life set up of the ED.
- $S_{max}(t)$: Maximum number of doctors that can be deployed at ED (both areas of the ED combined) at time t.
- C_l: cost of labour of deploying a doctor for a single unit time t.
- C_d: cost if the number doctors in time t deviates from $t-1$.
- L: lead-time for dynamic planning.
- H: time horizon for dynamic planning.

4 EXPERIMENTAL EVALUATION

We developed a prototype of the DQM to evaluate its performance and vary the decision-maker's parameters in the data model. The DQM simulator takes in information about historical process data and generates the simulated real-time data (actual arrival, actual treatment / investigation time, actual consultation time). Both the DQM simulator and the symbiotic simulation system are implemented in Java^{TM}.

4.1 Experimental Setup

We set up our experiments using 6 months data from the studied hospital. Each experiment is run over 100 replications and the result is an average across the replications. As for HIST-OPT and DYN-OPT which require the symbiotic simulation system, each symbiotic simulation is run over 50 replications and average is taken over the symbiotic simulation replications. The maximum search iteration is set to 300. In DYN-OPT, the lead-time is set to 0 (plan for next hour) and the planning horizon is set to 8 hours.

In order to verify that our DQM simulator is sufficiently close to the real-world's performance, we use FIFO patient queue and static doctor's schedule in a verification experiment using the DQM simulator. The outcome of the experiment shows that the differences in mean and standard deviation of the actual hospital data and the results from the DQM simulator are observed to be less than 5% and 10% respectively. The ranges of LOS (i.e., minimum and maximum) are also consistent. Therefore, we conclude that the results from DQM simulator is representative of the performance of the ED process.

To compute the time-varying staffing requirements for HIST and DYN, approximation methods are used for the instance of the hospital data that we are using. Solving the Equation 1 is non-trivial. Based on Yom-Tov (2010), if the service rate is represented by an exponential distribution, then we can use numerical approximation to compute R_1 and R_2 by solving the following Ordinary Differential Equation (O.D.E):

$$\begin{aligned} \frac{d}{dt}R_1(t) &= \lambda_t + \delta R_2(t) - \mu R_1(t); \\ \frac{d}{dt}R_2(t) &= p\mu R_1(t) - \delta R_2(t). \end{aligned} \tag{6}$$

As mentioned in Section 3.2, the service rate of the consultation station is represented by a set of exponential distributions with associated probabilities according to patient review types, which yields a hyperexponential distribution. Statistically, we know that the mean of this distribution is equal to the weighted average of the means of the underlying set of exponential distributions. Experimentally, we found interestingly from our dataset that with inclusion of the service rate of new patient μ_n, the resulting hyperexponential distribution can be approximated quite closely by an exponential distribution whose mean is equal to the mean of the hyperexponential distribution. The evidence (see Figure 3) is derived from an experiment that simulates the service times provided by both the approximated exponential distribution and the service rates represented by the original set of exponential distributions. As such, the staffing requirements for HIST and DYN in our experiment can be computed by solving the O.D.E in Equation 6 using the mean of the resulting hyperexponential distribution as the single service rate μ. For each set of experiments (for all strategies), a simulation over 9 days is taken, and the first and last day are discarded in order to remove the inaccurate results from simulation start-up and winding down. The remaining 7 days represent the 7 days of the week with time-varying arrivals as observed in real-life. Through the Analytical Model, the probability of re-entrance b found to be 0.4. The average service rate of doctors for new patients, μ_n, is 4 per hour. The registration and triage service time is exponentially distributed with mean of 14.2 minutes. The hospital's desired service quality, LOS_{max} is set to 60 minutes. The rate of investigation and treatment is set to 2.3 per hour.

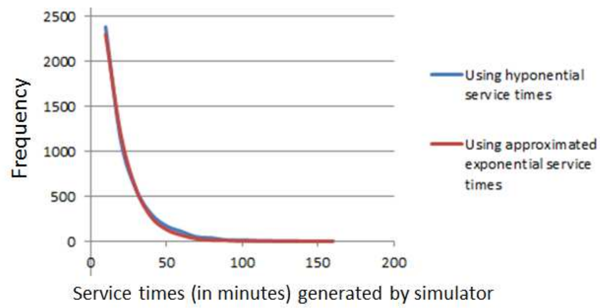
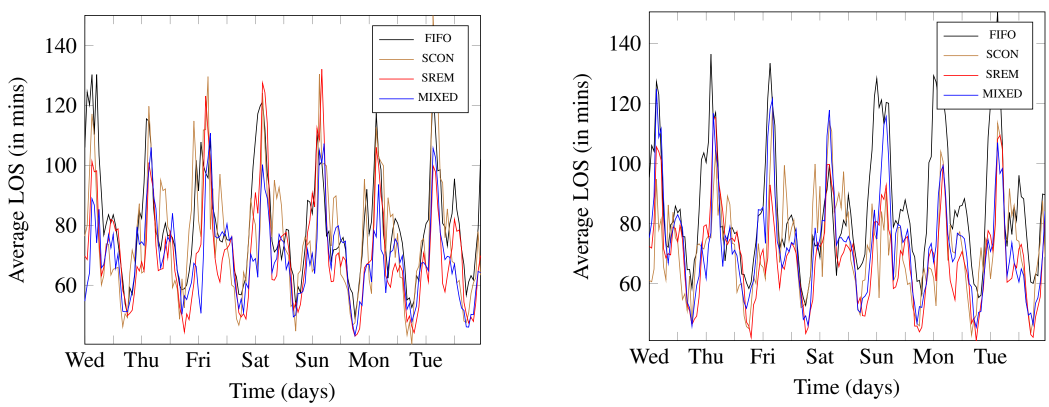


Figure 3: Approximation of Hyperexponential distribution with Exponential distribution.

4.2 Experimental Results

Two sets of experiments have been set up. First, we simulate the real-life situation when each hour has a maximum number of doctors that can be deployed. That means, $S_{max}(t)$ varies hourly. This experiment allows us to show the effects of the demand-side strategies on the supply-side strategies. We show in Figure 4, the results for HIST and DYN as representation of static and dynamic supply-side strategies. We



(a) Average LOS using HIST staffing

(b) Average LOS using DYN staffing

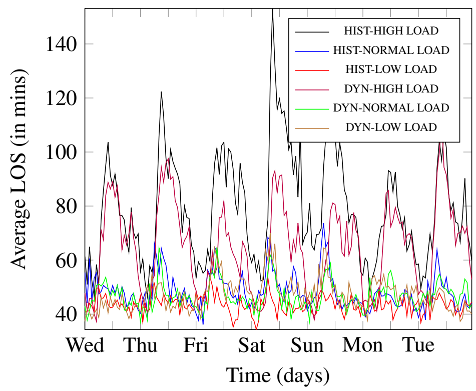
Figure 4: Results of using demand-side strategies with a selected supply-side strategy.

can see that demand-side strategies (SCON, SREM, MIXED) have similar performance as with FIFO in both HIST and DYN supply-side strategies. Although all the strategies (including FIFO) seem to perform well, we discover from Wilcoxon signed-rank tests that SCON, SREM and MIXED strategies still provide significant improvements over FIFO in both experiments of HIST and DYN. We present the outcome of Wilcoxon tests in Figure 5. In our Wilcoxon test, the null hypothesis H0 is ”Strategy S2 has no significant

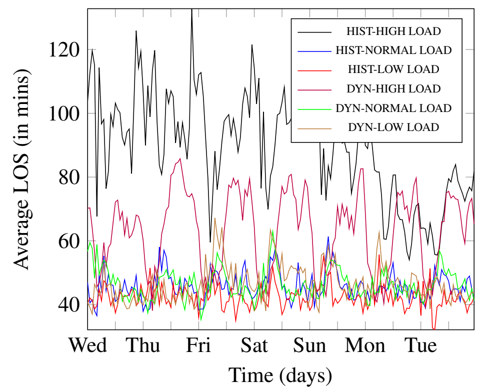
		S1							
		HIST (Cost = 524 doc hours/week)				DYN (Cost = 521 doc hours/week)			
		FIFO	SCON	SREM	MIXED	FIFO	SCON	SREM	MIXED
S2	FIFO	N.A	○	○	○	N.A	○	○	○
	SCON	✓	N.A	○	○	✓	N.A	○	✓
	SREM	✓	○	N.A	○	✓	✓	N.A	✓
	MIXED	✓	○	○	N.A	✓	○	○	N.A

Figure 5: The results of Wilcoxon signed rank test.

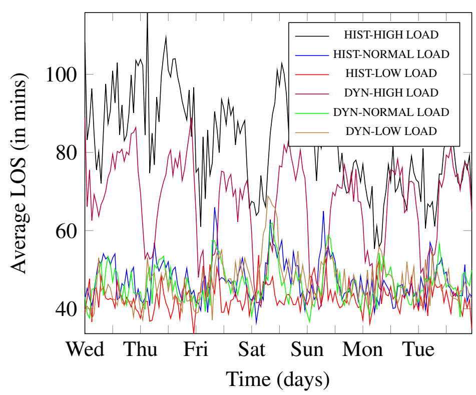
improvement over Strategy S1” and H1 is ”Strategy S2 has a significant improvement over S1”. We determine whether we reject H0 by the computation of the p-value. If the p-value is lower than 0.05, we can conclude that the null hypothesis does not hold (i.e. S2 is better than S1). For simplicity, we use a tick to represent values less than 0.05 and a circle otherwise.



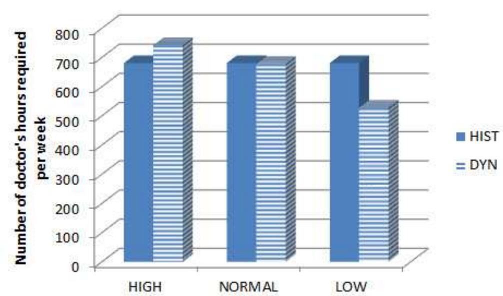
(a) Average LOS using SCON queueing policy



(b) Average LOS using SREM queueing policy



(c) Average LOS using MIXED queueing policy



(d) Cost of doctor's deployment

Figure 6: Results of using supply-side strategies with a selected demand-side strategy.

Our second set of experiments aims to determine how the supply-side strategies are useful in working together with the demand-side strategies. We now allow the ED to increase or decrease doctors to adapt to the demand changes as long as the number of doctors at any time t is still under the physical constraint of the ED as specified in Equation 4. To test the demand changes, we provided the simulator with three different types of arrival rates, high, normal and low. In the case of high load, the arrival rates are doubled for the days Thursdays, Fridays, Saturdays and Sundays. Likewise for the low load, the arrival rates are halved for the same days of the week. Using HIST and DYN as representation of static and dynamic strategies, Figure 6 shows how HIST and DYN perform under each demand-side strategy and with high, normal or low load conditions.

As we can see from Figure 6(a)-(c), the dynamic method DYN is coping better with demand surges as compared to HIST. This comes with a price of slight increase in the number of doctor's hours to be deployed at the ED (in a week) to handle the additional load. This is shown in Figure 6(d). The more interesting result is that the corresponding decrease in demand in fact yields a larger decrease in the number of doctor's hours to be deployed, with yet the performance is similar to that of HIST which is over-staffed under a normal and low-load conditions. Hence, we conclude that use of dynamic staffing method is effective in its ability to cope with demand surges and also cut cost when the demand is low.

5 CONCLUSION

In this work, we presented an integrated framework for dynamic queue management from both demand and supply perspectives. Our experimental analysis showed that the demand-side strategies work seamlessly with both static and dynamic strategies of the supply-side. Likewise, supply-side strategies are performing well with each demand-side strategies. In addition, the dynamic staffing (supply of doctors) can adapt to demand surges or cut cost when demand is reduced. The integrated framework allows healthcare decision makers to play a role in achieving the desired service quality and select from a list of possible strategies that suit the operation needs of the ED.

REFERENCES

Ben-Akiva, M., M. Bierlaire, H. Koutsopoulos, and R. Mishalani. 1998. "DynaMIT: A Simulation-Based System for Traffic Prediction". In *DACCORS Short Term Forecasting Workshop, The Netherlands*. Citeseer.

Cabrera, E., E. Luque, M. Taboada, F. Epelde, and M. L. Iglesias. 2012. "ABMS Optimization for Emergency Departments". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 89:1–89:12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Chakravarthy, S. 1992. "A Finite Capacity Dynamic Priority Queuing Model". *Computers & industrial engineering* 22 (4): 369–385.

Feldman, Z., A. Mandelbaum, W. A. Massey, and W. Whitt. 2008. "Staffing of Time-Varying Queues to Achieve Time-Stable Performance". *Management Science* 54 (2): 324–338.

Green, L. 2006. "Queueing Analysis in Healthcare". In *Patient flow: reducing delay in healthcare delivery*, 281–307. Springer.

Green, L. V., P. J. Kolesar, and W. Whitt. 2009. "Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System". *Production and Operations Management* 16 (1): 13–39.

Gunal, M. M., and M. Pidd.. 2006. "Understanding Accident and Emergency Department Performance using Simulation". In *Proceedings of the 2006 Winter Simulation Conference*, edited by L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 446–452. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Halfin, S., and W. Whitt. 1981. "Heavy-Traffic Limits for Queues with Many Exponential Servers". *Operations research*:567–588.

Jennings, O. B., A. Mandelbaum, W. A. Massey, and W. Whitt. 1996. "Server Staffing to Meet Time-Varying Demand". *Management Science* 42 (10): 1383–1394.

King, D. L., D. I. Ben-Tovim, and J. Bassham. 2006. "Redesigning Emergency Department Patient Flows: Application of Lean Thinking to Health Care". *Emergency Medicine Australasia* 18 (4): 391–397.

Komashie, A., and A. Mousavi. 2005. "Modeling Emergency Departments using Discrete Event Simulation Techniques". In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 2681–2685. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Low, M. Y., S. J. Turner, D. Ling, H. L. Peng, P. Lendermann, L. Chan, and S. Buckley. 2007. "Symbiotic Simulation for Business Process Re-engineering in High-Tech Manufacturing and Service Networks". In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 568–576. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Marmor, Y. N., S. Wasserkrug, S. Zeltyn, Y. Mesika, O. Greenshpan, B. Carmeli, A. Shtub, and A. Mandelbaum. 2009. "Toward Simulation-Based Real-Time Decision-Support Systems for Emergency Departments". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R.

Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 2042–2053. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Mendoza, G., M. Sedaghat, and K. P. Yoon. 2011. "Queuing Models to Balance Systems with Excess Supply". *International Business & Economics Research Journal (IBER)* 8 (1).

Pham, J. C., R. Patel, M. G. Millin, T. D. Kirsch, and A. Chanmugam. 2006. "The Effects of Ambulance Diversion: A Comprehensive Review". *Academic Emergency Medicine* 13 (11): 1220–1227.

Samaha, S., W. S. Armel, and D. W. Starks. 2003. "Emergency Departments I: The Use of Simulation to Reduce the Length of Stay in an Emergency Department". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P.J.Snchez, D. Ferrin, and D.J.Morrice, 1907–1911. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Tan, K. W., W. H. Tan, and H. C. Lau. 2013. "Improving Patient Length-of-Stay in Emergency Department Through Dynamic Resource Allocation Policies". In *Proceedings of the IEEE International Conference on Automation Science and Engineering (CASE)*.

Tan, K. W., C. Wang, and H. C. Lau. 2012. "Improving Patient Flow in Emergency Department Through Dynamic Priority Queue". In *Proceedings of the IEEE International Conference on Automation Science and Engineering (CASE)*, 125–130.

Thorwarth, M., and A. Arisha. 2012. "A Simulation-Based Decision Support System to Model Complex Demand Driven Healthcare Facilities". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Yom-Tov, G. B. 2010. "Queues in Hospitals: Queueing Networks with ReEntering Customers in the QED Regime". *The Technion - Israel Institute of Technology Ph.D. Thesis*.

AUTHOR BIOGRAPHIES

KAR WAY TAN is a PhD candidate at School of Information Systems, Singapore Management University. She received her BSc in Computer and Information Sciences (with First Class Honours) from National University of Singapore. Her research interests are in applied research spanning the cross-disciplinary domains of intelligent decision-support and business process management. She applies her research expertise in healthcare and logistics processes. Her email address is kwtan@smu.edu.sg.

HOONG CHUIN LAU is Associate Professor of Information Systems and Deputy Director of the Living Analytics Research Centre at the Singapore Management University. He received his Doctorate of Engineering degree in Computer Science from the Tokyo Institute of Technology (Japan), and BSc and MSc degrees in Computer Science from the University of Minnesota (Minneapolis). His research in the interface of Artificial Intelligence and Operations Research has been widely applied to decision support and optimization, and has contributed to advances of algorithms and applications to a variety of complex resource planning and scheduling problems in logistics, transportation, e-Commerce, health-care and tourism. He serves on the editorial board of the IEEE Transactions on Automation Science and Engineering. He is a chartered member and currently serves on the board of directors of the Chartered Institute of Logistics and Transportation (Singapore). His email address is hclau@smu.edu.sg.

FRANCIS CHUN YUE LEE is a Senior Consultant in Emergency Medicine and Head of Acute & Emergency Care Centre of Khoo Teck Puat Hospital. He is a fellow of the Royal College of Surgeon, Edinburgh and the Academy of Medicine, Singapore. He is a senior clinical lecturer of the Yong Loo Lin School of Medicine, National University Singapore. His subspecialty interest is in Disaster Medicine, Emergency Preparedness, Mass Event Medicine and Emergency Ultrasound. His email address is lee.francis@alexandrahealth.com.sg.