Simulation-Based Models of Emergency Departments: Operational, Tactical and Strategic Staffing

S. ZELTYN, Y. N. MARMOR, A. MANDELBAUM, B. CARMELI, O. GREENSHPAN, Y. MESIKA, S. WASSERKRUG, P. VORTMAN, A. SHTUB, T. LAUTERMAN, D. SCHWARTZ, K. MOSKOVITCH, S. TZAFRIR, F. BASIS[1]

The Emergency Department (ED) of a modern hospital is a highly complex system that gives rise to numerous managerial challenges. It spans the full spectrum of operational, clinical and financial perspectives, over varying horizons: operational – few hours or days ahead; tactical – weeks or a few months ahead; and strategic - which involves planning on monthly and yearly scales. Simulation offers a natural framework within which to address these challenges, as realistic ED models are typically intractable analytically. Specifically, we apply a general and flexible ED simulator to address several significant problems that arose in a large Israeli hospital. The paper focuses mainly, but not exclusively, on workforce staffing problems over the above time horizons. First, we demonstrate that our simulation model can support real-time control, which enables short-term prediction and operational planning (physician and nurse staffing) for several hours or days ahead. To this end, we implement a novel simulation-based technique that implements the concept of offered-load and discover that it performs better than a common alternative. Then we evaluate ED staff scheduling that adjusts for midterm changes (tactical horizon, several weeks or months ahead). Finally, we analyze the design and staffing problems that arose from physical relocation of the ED (strategic yearly horizon). Application of the simulation-based approach led to the implementation of our design and staffing recommendations.

Categories and Subject Descriptors: I 6.3 **[Simulation and Modeling]** – Applications – Health care.

General Terms: Simulation and modeling, health care, queues, queuing theory
Additional Key Words and Phrases: Emergency departments, offered-load, operational planning, tactical planning, strategic planning
ACM File Format:

1 INTRODUCTION

1.1 Operations Management in Emergency Departments: Main Challenges and Simulation-based Modeling

The rising cost of healthcare services has been a subject of mounting importance and much discussion worldwide. Ample reasons have been proposed, such as prolonged life spans and increasing accessibility to costly diagnostic and therapeutic modalities [Hall 2006]. These costs, which are reaching staggering levels, impose pressures on healthcare providers to improve the management of quality, efficiency and economics in their organizations.

A critical healthcare organization is the large hospital. Its complexity is well represented by the microcosm of its Emergency Department (ED), which is our focus here. EDs are widely recognized to be in need of urgent improvement. Indeed, the ED is the window through which a hospital is judged for better or worse, and it amplifies a variety of problems that intertwine clinical, operational and financial dimensions – though here we take an operations viewpoint.

Overcrowding and its consequent excessive delays are among the costliest ED operational problems [Sinreich and Marmor 2005], having clear interactions also with ED clinical and financial dimensions. To wit, Green [2008] notes that "arguably, the most critical delays for healthcare are the ones associated with healthcare emergencies". Overcrowding in the ED has many negative consequences. They include deteriorating clinical status of delayed patients; extended waiting times that inflate staff workload and distress patients and their families; ambulance diversion, which renders ED services temporarily inaccessible; patients who Leave Without Being Seen (LWBS), and who later return in worsened conditions; and so on [Derlet and Richards 2000].

[1] S. Zeltyn, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, IBM Research Lab, Mount Carmel, Haifa 31905, Israel. E-mails: sergeyz@il.ibm.com, boazc@il.ibm.com, ohadc@il.ibm.com, mesika@il.ibm.com, segevw@il.ibm.com, vortman@il.ibm.com; Y. N. Marmor, A. Mandelbaum, A. Shtub, T. Lauterman, Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa 32000, Israel. E-mails: myariv@tx.technion.ac.il, avim@tx.technion.ac.il, shtub@ie.technion.ac.il, tirza2502@gmail.com; D. Schwartz, K. Moskovitch, S. Tzafrir, F. Basis, Rambam Health Care Center, 6 Ha'Aliya Street, POB 9602, Haifa 31096, Israel. E-mails: d_schwartz@rambam.health.gov.il, j_moskovitz@rambam.health.gov.il, s_tzafrir@rambam.health.gov.il, f_basis@rambam.health.gov.il.

Our experience suggests that a key driver for overcrowding is inadequate staffing, but other causes have been identified as well. For example, Tseytlin [2009] studied problems in the process of hospitalizing ED patients, revealing a tradeoff between ED delays of patients vs. fair workloads on nurses and physicians. Tools and methods have been developed to alleviate overcrowding and excessive waiting times. These involve careful planning of the ED processes, in concert with appropriate staffing and scheduling techniques for ED personnel (nurses, physicians, X-Ray technicians and others). In the present research, we emphasize simulation-based solutions of staffing problems, over time horizons that vary from several hours to months and beyond.

The first staffing problem that we consider is short-term (operational) planning, over a horizon that spans several hours to a few days, which raised several challenges. For instance, accurate data on the current state of the ED is a prerequisite. Practically, however, a significant part of this data is inaccessible or unreliable, since hospital personnel often lack the time for online updates of IT systems. We address the need to infer an accurate "ED-state" through online simulation. Next, we develop a simple yet adequate forecasting model that predicts the number of future exogenous arrivals to the ED. Finally, a simulation model that combines the forecasts of these arrivals with the internal dynamics of the ED is developed and analyzed. This model, which can be integrated into an ED decision support system, produces short-term staffing schedules.

While short-term planning deals with scheduling changes over several hours or a shift ahead, midterm tactical planning is concerned with baseline schedules. Our midterm staffing thus accommodates seasonal effects of patient arrivals such as increases in arrival volume in the winter due to flu: a horizon that may span a week to several months and can be modeled off-line.

Long-term strategic planning considers staffing problems that arise due to planned or proposed major design changes. An example, considered in this paper, is the physical relocation of the ED within a hospital. Given such plans, our aim is to identify their effects on staff scheduling. These considerations must be accounted for when deciding on whether to implement proposed design changes and, if so, how to implement them in an acceptable way.

Both short- and mid-term staffing must account for time-varying conditions. To this end, we adopt the simulation-based approach of Feldman et al. [2008], who generate staffing schedules that stabilize operational performance of time-varying systems. For long-term staffing, on the other hand, steady-state modeling suffices. But all our staffing challenges require a reliable model of the ED. Analytical models have been found unable to capture the complexity of ED operations, over the wide spectrum that arises here. Hence, a major component of our solution is the ED simulation model of Sinreich and Marmor [2005, 2004], which is discussed in Section 4. Tailored to our needs, this simulation-based model is general and flexible enough to address all the above challenges.

1.2 Contribution and Structure of the Paper

In subsequent sections, we continue with a brief survey of related work (Section 2) and describe the ED of an Israeli hospital where our models have been applied (Section 3). Section 4 provides a detailed discussion of our simulation model. Then we proceed to the core of the paper, describing simulation-based techniques for workforce planning over various horizons. Section 5 introduces a new approach to staffing, based on the concept of offered-load, which is then compared to the well-known method of Rough Cut Capacity Planning (RCCP). In that section we also study the problem of inferring missing ED data through simulation. This treatment also extends that of Marmor et al. [2009]. Section 6 discusses midterm tactical planning, where the approaches of offered-load and RCCP are applied and again compared. In Section 7, we present the main research issues in a project dedicated to a transfer from an original ED location to a temporary one (so that a new ED could be built) and focus on the long-term staffing of nurses. Section 8 discusses the impact of our work. Finally, Section 9 lists the paper conclusions and promising directions of possible future research.

Our contribution is to demonstrate that a single well-designed simulation model of an ED can be instrumental in the solution of ED staff scheduling problems, from online decision support, through short-term operational planning, midterm tactical planning, and finally, to long-term scenario analysis. We are unaware of any uses of simulation in a hospital setting for online decision support or of any work in which simulation has been used to complete missing data regarding the current operational state. In addition, we introduce a new offered-load approach to staffing that combines simulation and analytical formulae, and yields promising results over varying time domains. The research recommendations were successfully applied when the ED was moved to a temporary location.

Our simulation model also enables improvement of ED processes that are not directly related to staffing. For example, it can support the optimal design of ED units in order to decrease walking distances of nurses and physicians. Finally, this simulation framework is flexible and, being based on field research, carried out in nine Israeli Emergency Departments, can thus be tuned to meet the needs of other EDs in Israel and around the world.

2 RELATED WORK: SIMULATION IN SUPPORT OF ED OPERATIONS

The application of simulation has been instrumental in addressing the multi-faceted challenges that the healthcare domain presents [Kuljis, Paul, and Stergioulas 2007]. A wide spectrum of ED problems has also received significant attention in this line of research.

Researchers commonly use simulation to compare operational models or to assess a model that addresses a specific research question. For example, Medeiros, Swenson, and DeFlitch [2008] present a simulation-based validation of a novel approach to a change in ED processes, placing an emergency care physician at triage, while Kolb et al. [2008] study different policies of patient transfer from ED to internal wards, in order to decrease the resulting overcrowding and delays. Tseytlin [2009] addresses a similar problem for our hospital, but uses an analytical approach, based on queuing models.

Key reviews include Jun, Jacobson, and Swisher [1999], White [2005] and Jacobson, Hall, and Swisher [2006]. Meanwhile, the improvement of patient experiences in EDs using a combination of simulation and lean manufacturing tools was considered in Khurma, Bacioiu, and Pasek [2008].

The prevalent approach to ED overcrowding lies in the staff scheduling [Sinreich and Jabali 2007; Badri and Hollingsworth 1993] and focus on off-line steady-state decision-making, as opposed to on-line operational and tactical control. Others analyze alternative ED designs [Garcia et al. 1995; King, Ben-Tovim, and Bassham 2006; Liyanage and Gale 1995]. For example, acuteness-driven models such as triage are compared against more operational solutions such as fast-tracking, which assigns high priority to patients with low resource requirements.

A widespread approach is to "divide and conquer" a complex problem by focusing only on one type of resource at a time. An example is an effort to schedule nurses while ignoring the scheduling of other resources [Draeger 1992]; or scheduling physicians and nurses hierarchically [Sinreich and Jabali 2007]. These attempts predict performances of the ED as a function of staffing and scheduling decisions. The simulation models require input in the usual form of patient arrivals and service durations, for each patient by each resource type, exactly as in the simulation that we are using here.

Over a broader perspective, we note a progress in active simulation-driven research. For example, input modeling [Biller and Nelson 2002] and historical (trace-driven, resampling) simulation [Asmussen and Glynn 2007; McNeil, Frey, and Embrecht 2005] are both related to the problem of properly incorporating actual ED data into our simulator. Another related field of interest is symbiotic simulation [Fujimoto et al. 2002; Huang et al. 2006], defined as "one that interacts with the physical system in a mutually beneficial way", "driven by real time data collected from a physical system under control and needs to meet the real-time requirements of the physical system" [Huang et al. 2006]. Additionally [Fujimoto et al. 2002], symbiotic simulation is "highly adaptive, in that the simulation system not only performs 'what-if' experiments that are used to control the physical system, but also accepts and responds to data from the physical system".

Although a simulation-based approach is the focus of our research, we emphasize that an optimization approach to real-life ED problems should combine simulation and analytical insights. For example, Beaulieu et al. [2000] present a deterministic mathematical programming approach to staff scheduling. The RCCP approach, demonstrated in Section 5 [Vollmann, Berry, and Whybark 1993], is also based on deterministic considerations. However, in our opinion, stochastic models, based on queuing theory, are more appropriate for capturing the volatile and inherently nondeterministic ED reality. Although it is hard to design a tractable comprehensive queuing model for the ED, it is possible to develop simpler models and combine them with simulation. The research on the offered-load concept, presented in Section 5 provides an example of this approach. Using the offered-load technique, applied to time-varying queuing systems in Feldman et al. [2008], we develop a novel staffing algorithm. Readers are referred to Green [2008] for further references on staff scheduling and related issues.

3 RESEARCH FRAMEWORK

This research is a part of an Open Collaborative Research program, a combined research effort of three organizations partnered together: the Faculty of Industrial Engineering & Management at the Technion Institute, IBM's Haifa Research Laboratory and the government-affiliated Rambam hospital – which is Israel's largest northern medical center, serving over 2 million citizens (about one-third of Israel's population). It has 36 wards accommodating around 1,000 patients and admits 75,000 patients a year. In this research project, we focus on several hospital units including the ED, which is the gate into and the window on, the hospital, and which must operate in a mass-customized mode by following a structured care process while providing each individual with the specific care required.

The ED accepts 82,000 patients per year, with 58% classified as internal patients and 42% as surgical or orthopedic patients. The ED contains three major areas:

1) internal acute: waiting and treatment space for acute internal patients treated by dedicated internists physicians and nurses;
2) trauma acute: waiting and treatment space for surgical and orthopedic patients treated by dedicated nurses, but shared by orthopedic and surgical physicians;
3) walking: an area for walking patients (patients that do not need a bed and use chairs, usually with mild problems). It contains a waiting lobby and specialist unique treatment rooms for internal, surgical, and orthopedic physicians. In the walking area, there is also a psychiatric unit.

In addition, there are other emergency rooms detached from the main ED, which are dedicated to special issues such as pediatrics and ophthalmology.

The average length of stay (ALOS) in the Rambam hospital equals 4:38 hours, with a large variance over individual patients.

4 BASIC SIMULATION MODEL OF THE EMERGENCY DEPARTMENT

In Fig. 1, we depict two perspectives of the care process that patients undergo at the ED: the resource (i.e. physicians, nurses, etc.) perspective, and the process (activities) perspective. In this figure, two types of queues correspond to two types of delays encountered by patients. The first are resource queues (denoted by rectangles), due to limited resources (e.g. nurses, imaging equipment). The second are synchronization queues (denoted by triangles), which arise when one process activity awaits another. An example of this might be when a patient waits for results of a blood test in order to proceed with the doctor's examination. Note that Fig. 1 presents a somewhat simplified model of the care process. A more complete model is described in Sinreich and Marmor [2005].

The care process in an ED was captured in a simulation model, created with the generic simulation tool of Sinreich and Marmor [2005]. This model is based on field studies, performed in Emergency Departments of nine Israeli hospitals. The required data was gathered either from the IT systems of these hospitals or in person by research staff. In addition to the care process, the simulation model required patient arrival processes, for each patient type, and staffing levels of the medical staff, with their respective skills. Service times in our model were assumed to be exponentially distributed and a statistical analysis validated this fit for most data types.

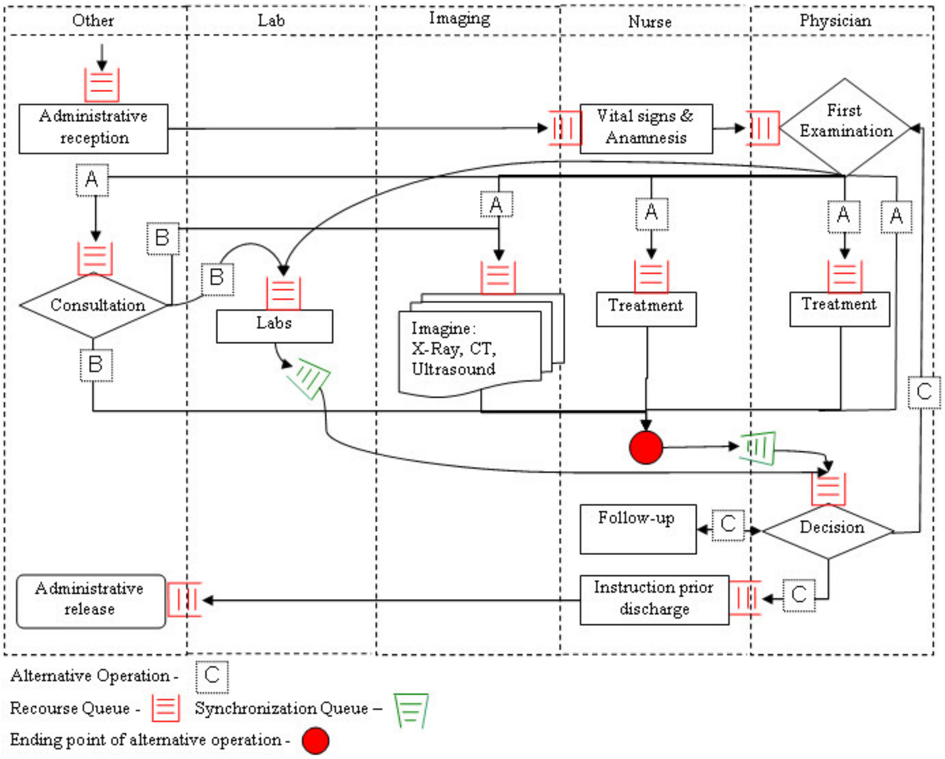


Fig. 1. ED resource-process chart.

The model was configured to the ED using six types of patients, which require different skills from the caring physicians:

- patient types 1 and 2 are internal acute and internal walking respectively, and are treated by internal physicians;
- patient types 3 and 4 are surgical acute and surgical walking respectively, and require treatment by surgical physicians;
- patient types 5 and 6 are orthopedic acute and orthopedic walking respectively, who require an orthopedic physician.

Acute patients need a bed while walking patients use chairs. In addition, patient types differ by the arrival process (e.g., number of arrivals per hour and by day-of-week; see Fig. 2), and by the decisions made in the patient care process (e.g., the percentage of patients sent to X-Ray).

The actual simulation tool is comprised of the following three modules:

1. The first module is a Graphical User Interface (GUI) that describes the general unified process, partially presented in Fig. 1. Through the GUI, the user can input data and customize the general process to fit the specific ED modeled and receive operational results from the ED after the simulation run. The detailed GUI description with screen shots can be found in Section 2.1 of Sinreich and Marmor [2004].
2. The second module includes two mathematical models used to estimate patient arrivals and staff walking time. The simulation tool uses the models for patient arrival estimation that were developed in Sinreich and Marmor [2005].
3. The third and final module is the simulation model itself. This model receives data from both the GUI and the mathematical models. The simulation is updated and customized automatically to fit a specific ED based on data and information the user passes on to the GUI. The simulation model is transparent to the user who is only required to interact with a user friendly GUI without the need to learn a simulation language syntax.

5 OPERATIONAL HORIZON: SIMULATION-BASED MODELING FOR ONLINE DECISION SUPPORT AND OPERATIONS PLANNING IN THE ED

In this section, we start to apply our simulation-based modeling approach to real-life ED problems. We show that this approach can help ED managers to infer the missing information on the current ED state, provide a reliable forecast of the ED state in the short-term and perform operational staff scheduling decisions.

5.1 Simulation-Based Validation of the Current ED State

As discussed in Section 1.1, reliable information on the current state of ED is crucial for online decision support and operational planning. Typically, only partial data of the current ED state is maintained and available from the hospital's electronic data systems. For example, in our case, no data exists regarding the patients waiting to be seen by a physician. One expects the amount and quality of usable data to improve constantly over time, due to the introduction of additional data-entry systems or new technologies, such as Radio Frequency Identification (RFID) and ultrasound, for accurate location tracking of patients, staff and equipment. However, within the chaotic ED environment, it is reasonable to expect that some data will always remain unavailable or too costly to acquire.

We now discuss how to infer missing data, using the simulation model described above. Such simulation-based inference must deal with several issues. The first is consistency: how to generate simulation paths that are consistent with available ED data. Another important issue is data inaccuracy that adds complexity to generation of simulation realizations that are consistent with the provided data. A third challenge, arising due to the availability of only incomplete data, is the identification of an appropriate initial state for the simulation. The way we overcome this last hurdle is to feed in actual arrival data for a long enough period of time (we used three weeks) that ensures that the simulation warm-up period is over, prior to estimating the missing data.

Coping with consistency and inaccuracy raises interesting research questions. Here we content ourselves with two ED-specific practical examples of accommodating actual ED data – accurate and inaccurate.

Accurate data - taking actual arrivals into account: In our partner ED, receptionists enter data into the IT systems, in particular regarding patient arrivals, as a part of the admission process. Therefore, arrival data accurately captures actual patients' arrival times – it can be fed as is into the simulator. Receptionists also record patient type upon arrival. To this end, we modified, in an obvious manner, our generic simulator, which originally generates arrivals as a Poisson process or its relatives, such as normal approximation to Poisson; see Sinreich and Marmor [2005].

It can now generate realizations consistent with the arrival data (e.g., time and patient type), when the latter is fed to the simulation package as a link from an external database.

Inaccurate data - taking discharges into account: Data about patients' discharge (departure) times, in our partner hospital, may be inaccurate. Specifically, each departure time is registered by the receptionist upon completion of the ED treatments – the patient is then ready to leave, for either home or to other hospital wards. In the common case when there is no ward immediately available to accept the patient, inaccurate data arises. Then, patients spend additional time waiting in the ED, which not only goes unrecorded but it also influences subsequent beds/chairs occupancy and ED staff utilization (due to time spent on catering to these delayed patients). Additional inaccuracies occur due to patients' leaving without being seen [Green 2008], with or without their medical files, and some other accounting-related reasons.

We found no efficient way for generating simulation paths that are consistent with our discharge data, except for discarding inconsistent realizations. Note, however, that the probability of generating a realization in which the simulated departure times correspond exactly to the provided departure times is negligible. To this end, and to overcome both inaccuracy issues, we validate the current state simulation by conditioning it on the number of patients of each type that were discharged from the ED according to the data. Namely, we considered a simulation realization to be consistent if, at the end of the simulation run, the number of patients of each type that were discharged equals, within some accuracy constant, the number of patients of this type that were discharged according to the data. In our case, we used 1.96-standard-deviation accuracy and accepted around 42% of the simulations results.

Section 5.5.1 applies the described techniques to the actual ED data.

5.2 Forecasting ED Arrivals

For simulating an ED future evolution, one must simulate patient arrivals to the ED. Fig. 2, based on IT data from the Rambam hospital, demonstrates that ED arrival rates strongly depend on day-of-week and hour-of-day. In addition, holidays and days after holidays have unusual patterns as well: holidays are lightly loaded and days after holidays are, as a rule, very heavily loaded. For a reference on forecasting and modeling of ED arrivals, leading also to related literature, see Channouf et al. [2007].

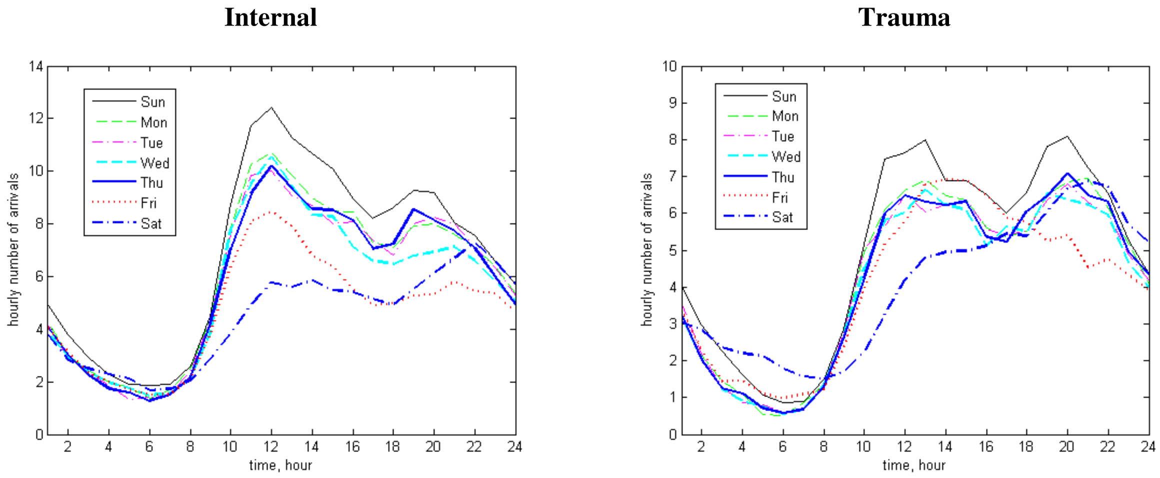

Internal **Trauma**

Fig. 2. Hourly arrival rates per patient type averaged over 4 years.

Arrivals in our simulation model are nonhomogeneous Poisson processes, with hourly rates that are forecasted for each future hour in question (say a shift, or a day) and each patient type. The nonhomogeneous Poisson assumption was validated in Maman, Mandelbaum and Zeltyn [2009], using the test developed in Brown et al. [2005]. Sinreich and Marmor [2005] demonstrated approximately normal distribution of square root of the arrival volumes, which is also consistent with the Poisson assumption (again, see Brown et al. [2005]). We assume that arrival rates are constant on an hourly scale.

Long-term moving average (MA) was used in order to predict hourly arrival rates. For example, in order to predict the arrival rate (assumed constant) on Tuesday during 11–12am, we average the corresponding arrival rates during the last 50 "Tuesdays 11–12am", excluding those that are holidays or days after holidays. We can also see that arrival patterns of internal and trauma patients are not similar—internal peak at about 8pm is much smaller than the one at noon. In contrast, the corresponding peaks of the trauma intraday arrival rate are of similar height. In particular, it means that we cannot predict the total number of arrivals and assign fixed probabilities to patient types.

The reason for choosing long-term MA is that we found it to provide essentially the same goodness-of-fit as more complicated time-series techniques. Indeed, long-term MA, applied to the overall arrival rate over a test period of 60 weeks, gave rise to a Mean Square Error (MSE) equal to 3.56, while two methods, based on Holt-Winters exponential smoothing, provide a MSE=3.55 and 3.54. Another argument in favor of the use of long-term MA stems from the level of stochastic variability in historical samples, calculated for each hour-of-week, which fits that of a Poisson process [Maman, Mandelbaum and Zeltyn 2009]; then, the historical mean (or MA) is a natural (Maximum Likelihood) estimate for the Poisson parameter, namely the arrival rate.

5.3 Staff Scheduling Approaches

With the present ED state assumed given, simulation is now to be used for predicting ED evolution several hours into the future; the goal is to determine appropriate staffing levels of resources – nurses, physicians and support staff, as a function of time.

Staffing the ED is a complex multiobjective problem. It must trade off conflicting objectives such as (i) Minimizing costs, (ii) Maximizing resource utilization, (iii) minimizing waiting time of patients, (iv) Maximizing quality of care. In this paper, we concentrate on the control of operational performance measures—utilization and waiting time. The complexity of theoretical analysis for a large complicated service network in a stochastic environment renders the optimization problem intractable analytically. This has thus led researchers to simulation-based solutions.

A prerequisite for staffing is accurate forecasting of patient arrivals, as described in Section 5.2. We then continue with predicting resource utilization; this leads to a staffing method, based on pre-specified goals for resource utilization (Section 5.3.1). However, the resources' view cannot accommodate the experience of patients—for example, controlling the time until the first encounter with a physician (Section 5.3.2). To control the latter, we calculate, for each resource type, its *offered-load* as a function of time; then a classical square-root safety-staffing principle, in conjunction with the appropriate queuing model, yields our recommended time-varying staffing levels. In Section 5.4, a summary of our methodology is presented.

5.3.1 Staff Scheduling via Rough Cut Capacity Planning

Rough Cut Capacity Planning (RCCP) is a technique for projecting resource requirements in a manufacturing or a service facility. As such, RCCP supports decisions regarding the acquisition and use of resources. Procedures for RCCP are listed in Vollmann, Berry, and Whybark [1993]. These procedures are based on the estimated processing time of each product or service unit, and the allocation of the total time among the different resource types. The goal is to match offered capacity with the forecasted demand for the capacity of each resource type. Thus, RCCP algorithms translate forecasts into an aggregate capacity plan, taking into account the time each resource type spends on each type of product or service.

We are proposing to apply RCCP in the ED environment, as follows:

- For each patient type i, calculate its average *total* time required from each resource type r (e.g. physician, nurse), d_{ir}.
- For each forecasted hour t, calculate the average number of *external* arrivals of patients of type i, $A_i(t)$.

Deduce the expected processing time required from each resource type r at time t:

$$\mathrm{RCCP}_r(t) = \sum_i A_i(t) d_{ir}. \tag{1}$$

- The recommended number of units of resource r at time t, $n_r(\mathrm{RCCP},t)$, is equal to the load $\mathrm{RCCP}_r(t)$, amplified by Safety Factor, or SF. SF is the maximum utilization we are targeting. In other words, the RCCP staffing recommendation is given by $n_r(\mathrm{RCCP},t) = \mathrm{RCCP}_r(t) / \mathrm{SF}$.

We expect RCCP to achieve preplanned resource utilization levels; its shortcoming, however, is that it ignores the time lag between arrival times of patients and actual times when these patients receive service or treatment from ED resources. Since patients spend, on average, several hours in ED this time lag can be significant: the patient arrival rate frequently reaches maximum before the workload for a specific resource reaches maximum. This problem is remedied by our next approach.

5.3.2 The Offered-Load Approach

The concept of *offered-load* is central for the analysis of operational performance. It is a refinement of RCCP in the sense that it spreads workload more accurately over time. For example, suppose that a nurse is required twice by a patient, once for injecting a medicine (10 minutes) and then, 3 hours later, after the medicine take its effect, for testing the results (also 10 minutes). RCCP would "load" 20 minutes of nurse-work upon a patient's arrival; the offered-load approach, in contrast, would acknowledge the 3-hour separation between the two 10-minute requirements. Such time-sensitivity enables one to accommodate time-based performance measures, notably those reflecting the quality of care from the patients' viewpoint.

In the simplest time-homogeneous steady-state case of a single service station, when the system is characterized by a constant arrival rate λ and a constant service rate μ, the offered-load is simply $R = \lambda/\mu = \lambda E(S)$ where $E(S)$ is the average service time. The quantity R represents the amount of work, measured in time-units of service, which arrives to the system per the same time-unit (say, hours of work that arrive per hour). Staffing rules can be naturally expressed in the terms of the offered-load: for example, the well known "square-root staffing rule" [Halfin and Whitt 1981; Borst, Mandelbaum, and Reiman 2004] postulates staffing according to

$$n = R + \beta\sqrt{R}, \tag{2}$$

where $\beta>0$ is a service-level parameter, which is set according to some service level agreement (SLA) or goal. This rule gives rise to quality and efficiency-driven (QED) operational performance, in the sense that it carefully balances high service quality with high utilization levels of resources. Arrival rates to an ED are, however, manifestly nonhomogeneous and depend on the day-of-week and hour-of-day. Piecewise stationary approximations work fine if the arrival rate is slowly varying with respect to the duration of services [Green, Kolesar, and Soares 2001]. This, however, does not happen in the ED case.

Assume that exogenous arrivals to a service system can be modeled by a nonhomogeneous Poisson with arrival rate $\lambda(t), t \geq 0$. In this case, our definition of the offered-load is based on the number of busy servers in a corresponding system with an *infinite* number of servers [Feldman et al. 2008]. Specifically, any one of the following four representations gives it:

$$R(t) = E[A(t) - A(t-S)] = E[\lambda(t-S_e)] \cdot E[S] = E\left[\int_{t-S}^{t} \lambda(u)du\right] = \int_{-\infty}^{t} \lambda(u)P(S > t-u)du, \tag{3}$$

where $A(t)$ is the cumulative number of arrivals up to time t, S is a (generic) service time, and S_e is its so-called excess service time. (See the review paper by Green, Kolesar, and Whitt [2007] for more details, as well as for useful approximations of (3).) Then, for calculating the time-varying performance in the case of a single service station, we recommend to substitute (3) into the corresponding steady-state model. In our case, the classical *M/M/n* queue, or Erlang-C, is used. To be concrete, assume that our service goal specifies a lower bound α, to the fraction of patients that start service within T time units. The QED approximation, based on Halfin and Whitt [1981] then gives rise to

$$1-\alpha = P\{W_q > T\} = P\{W_q > 0\} \cdot P\{W_q > T \mid W_q > 0\} \approx h(\beta_t) \cdot e^{-T\mu\beta_t\sqrt{R_t+\beta_t\sqrt{R_t}}}, \tag{4}$$

where $h(\beta_t)$ is the Halfin-Whitt function. Specifically, $h(\beta)$ approximates the delay probability $P\{W_q > 0\}$ in the Erlang-C queue given staffing level (2). Equation (4) can now be solved numerically with respect to β_t, and the staffing rule (2) is replaced by the time-varying staffing function:

$$n(\mathrm{OL}, t) = R(t) + \beta_t\sqrt{R(t)}. \tag{5}$$

The above procedure has been called the "modified offered-load approximations" — readers are referred to Feldman et al. [2008] for additional details and further references.

Square-root staffing is mathematically justified by asymptotic analysis, as workload and hence the number of servers increase indefinitely. However, ample experience, as well as recent research [Janssen, Van Leeuwaarden, and Zwart 2008], demonstrates useful levels of accuracy, already for *single*-digit staffing levels. This renders the above staffing rule relevant for EDs, as well as other healthcare systems, where the number of servers is indeed single-digit. For small systems, one could always apply exact Erlang-C formulae. Indeed, we tested these exact calculations against the QED approximations in our experiments below, and the results were essentially unaltered.

Now we extend the above framework from a single service station to a service network, in order to apply it in the ED. We proceed via the following steps:

- First, the simulation model is run with infinitely many resources (e.g. physicians and nurses).
- Second, for each resource r and each hour t, we calculate the number of busy resources, and use this value as our estimate for the offered-load $R(t)$ for resource r at time t. The final value of $R(t)$ is calculated by averaging over simulation runs.
- Finally, for each hour t we deduce a recommended staffing level $n_r(\mathrm{OL},t)$ via formulae (4) and (5).

5.4 Methodology for Short-term Forecasting and Staffing

In the following section, we set short-term staffing levels for eight hours into the future. Our simulation-based methodology for short-term forecasting of the ED state is as follows:

1. Initialize with the simulation-based estimate of the current ED state.
2. Use the average arrival rate, calculated from the long term MA, to generate stochastic arrivals in the simulation.
3. Simulate and collect data every hour, for eight future hours, using infinite resources (nurses, doctors).
4. From step 3, calculate staffing recommendations $n_r(\mathrm{RCCP},t)$ and $n_r(\mathrm{OL},t)$ using RCCP and Offered-Load (OL) methods, described in Sections 5.3.1 and 5.3.2, respectively.
5. Run the simulation from the current ED state with the recommended staffing.
6. Calculate performance measures. The above can be repeated with the actual staffing (in Step 5), which makes it possible to compare it against RCCP and OL staffing.

5.5 Simulation Experiments

We now apply methodology from the previous section in simulation experiments. First, we demonstrate the ability of our simulation-based tool to estimate the current ED state, using a database from Rambam hospital (Section 5.5.1). For that, we randomly choose a month (August 2007) in the database, and compare the known number of patients in the system with the simulation's outcome. In the second experiment (Section 5.5.2), we use the ED state at a specific time (September 2nd, 2007, 16:00) to predict 1–7 hours ahead. The chosen day is Sunday, which, in Israel, is a busy day of the week, being the first day following the weekend. We continue, in Section 5.5.3, with a comparison of some ED performance measures, using two alternative staffing methods (RCCP and OL), and the methodology developed in Section 5.3. Finally, in Section 5.5.4 we compare the two staffing techniques given the same number of resources is used.

5.5.1 Current State

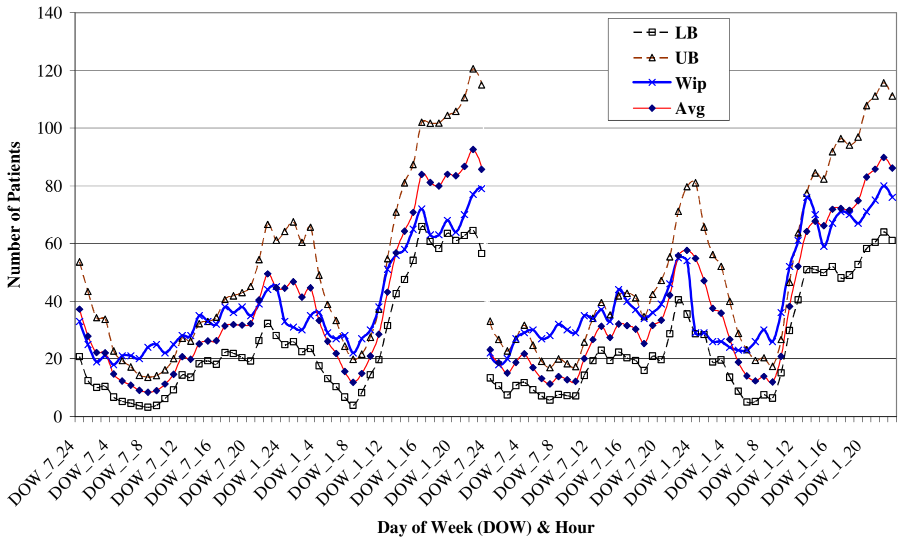


Fig. 3. Comparing the Database with the simulated ED current-state (weekdays and weekends).

We ran 100 one-month long replications of each scenario, in order to compare our simulation results with the data from the hospital's database. For each date and hour, we calculated the average number of patients over the simulation replication (Avg series in Figure 3), and the corresponding standard deviation (SD), an Upper Bound (UB = Avg + 1.96 SD), and a Lower Bound (LB = Avg - 1.96 SD). In Figure 3, we depict 4 days, chosen to test our methodology against the (actual) number of patients from the database (Wip is work in progress). We chose two periods that are two days long, the last day of the weekend (Saturday in Israel) and the first working day of the next week (Sunday). For example, DOW_7_4 at time axis stands for 4am on Saturday and DOW_1_16 denotes 4pm on Sunday.

These days are typically the calmest and busiest in the week, respectively. Note that the night and early morning shifts (hours 1–10 in Fig. 3) are not overloaded and performance measures are then less accurate. See, for example, the utilization profiles during 9–10am, in Table I. However, once the ED becomes congested, the simulation does yield an accurate prediction of the number of patients in the ED. At all times, though, the accuracy of prediction varies from reasonable to good.

A probable explanation for a somewhat worse fit of the simulation during lightly loaded hours is the following. When the load is low, the staff has more time for activities that are not incorporated into our simulation, such as department meetings. In contrast, during heavily loaded periods, there is virtually no time for such activities and reality becomes consistent with the simulation.

5.5.2 Calculation of Short-Term Staffing Recommendations

Next, we simulated the system in the near future using methodology from Section 5.4, to see if there is a way to improve ED operations via an appropriate staffing technique. We calculated the offered-load of all the relevant resources: internal physician (Ip), surgical physician (Sp), orthopedic physician (Op) and nurses (Nu). For this experiment, we used ED data until 16:00 and then applied simulation to forecast each succeeding hour, until the end of the day. Here and in the experiments described below, 100 simulations were performed. In Table I, we display the ED state until 16:00, and then continue with the simulation-based forecast; the staffing levels used in the simulation are the one exercised in our partner ED — we refer to it as "the actual staffing". Columns Ip, Sp, Op, and Nu list utilization levels of the respective staff. The column headings #Beds and #Chairs represent the average number of occupied beds and chairs, respectively; *%(W>T)* is the fraction of patients that are exposed to unsatisfactory care, which here is taken to be "physician's first encounter occurs later than *T* minutes after arrival to the ED". In our research, the value of *T* is equal to 30 minutes.

Table I. Simulation Performance Measures – Current and Forecasted (Actual Staffing)

Hour	Resource utilization				#Beds	#Chairs	*%(W>T)*
	Ip	Sp	Op	Nu			
09-10	73%	1%	23%	55%	15.7	8.6	7%
10-11	93%	25%	59%	68%	23.5	17.0	33%
11-12	94%	59%	67%	72%	29.3	22.8	51%
12-13	90%	45%	81%	58%	33.2	30.3	53%
13-14	95%	68%	94%	71%	36.2	34.7	77%
14-15	90%	62%	76%	63%	34.2	33.3	70%
15-16	91%	51%	46%	51%	34.4	30.5	77%
16-17	100%	43%	41%	53%	34.6	27.6	69%
17-18	95%	58%	46%	57%	33.4	23.6	52%
18-19	90%	46%	52%	50%	32.4	23.9	31%
19-20	89%	64%	70%	58%	29.3	25.3	40%
20-21	79%	64%	75%	56%	26.5	20.6	39%
21-22	84%	46%	60%	45%	23.4	17.0	23%
22-23	66%	38%	51%	46%	20.2	13.9	20%

In Table II we display the following characteristics:

- ED actual staffing is denoted by *n*(Current),
- the offered-load level (as explained in Section 5.3.2) in Offered-Load column,

- recommended staffing level based on the offered-load (aiming to achieve *%(W>T)*< 0.25) – *n*(OL),
- the RCCP level (as explained in Section 5.3.1) – RCCP Load columns,

RCCP staffing recommendations aiming at less than 90% staff utilization – *n* (RCCP).

Table II. Staffing Levels (Actual and Recommended)

	n (Current)				Offered-Load				N (OL)				RCCP Load				n (RCCP)			
Hour	Ip	Sp	Op	Nu	Ip	Sp	Op	Nu	Ip	Sp	Op	Nu	Ip	Sp	Op	Nu	Ip	Sp	Op	Nu
16-17	4	1	2	5	7.8	0.8	0.8	4.1	9	2	2	5	3.0	0.5	0.6	2.4	4	1	1	3
17-18	4	1	2	5	3.7	0.4	0.9	2.5	5	1	2	3	3.3	0.4	0.7	1.3	4	1	1	2
18-19	4	1	2	5	3.2	0.4	1.1	2.7	4	1	2	4	2.3	0.4	0.4	1.3	3	1	1	2
19-20	4	1	2	5	2.3	0.5	1.2	2.5	3	1	2	3	2.4	0.5	0.6	1.0	3	1	1	2
20-21	4	1	2	5	2.7	0.6	1.5	2.7	4	1	2	4	2.3	0.5	0.4	1.0	3	1	1	2
21-22	4	1	2	5	2.4	0.4	1.3	2.4	3	1	2	3	2.8	0.5	0.4	1.1	4	1	1	2
22-23	4	1	2	5	2.3	0.2	0.9	2.0	3	1	2	3	2.4	0.3	0.2	1.0	3	1	1	2

5.5.3 Short-term Staffing Recommendations – Performance Forecasting

In Table III, we record simulated performance, under staffing levels calculated via OL and RCCP methods. As anticipated, the offered-load method achieved good service quality: indeed, the fraction of patients getting to see a physician within their first half hour at the ED is typically less than half of those under RCCP, the latter being also more influenced by the changes in the arrival rate. RCCP of course yields good performance at the resource utilization column, all being near the 90% target for the resources with staffing levels larger than 1–2.

It is interesting to compare Table III that presents recommended staffing with Table II that displays levels of actual staffing and the corresponding performance: the latter has obvious hours of under- and over-staffing while the formers' performance is relatively stable. For example, *n*(Current) in Table II implies understaffing during 16-17 and overstaffing for 22-23 period. Preplanned staffing, either for resource utilization (RCCP) or, better yet, patients' service level (OL), clearly has its merit.

Table III. Simulation Performance Measures (Using OL and RCCP)

	Performance measures using OL recommendation							Performance measures using RCCP recommendation						
Hour	Resource utilization				#Beds	#Chairs	*%(W>T)*	Resource utilization				#Beds	#Chairs	*%(W>T)*
	Ip	Sp	Op	Nu				Ip	Sp	Op	Nu			
16-17	62%	38%	40%	58%	36.0	29.0	56%	90%	54%	60%	59%	38.3	35.3	78%
17-18	59%	33%	35%	67%	34.8	31.6	36%	82%	47%	65%	81%	39.3	40.2	82%
18-19	75%	49%	53%	76%	32.2	29.9	46%	80%	45%	69%	92%	40.6	46.2	86%
19-20	84%	48%	57%	80%	31.5	31.1	38%	72%	43%	79%	97%	42.3	52.2	90%
20-21	76%	52%	65%	71%	28.7	28.4	38%	68%	46%	85%	99%	43.4	57.7	91%
21-22	83%	49%	59%	75%	27.8	27.9	42%	55%	45%	89%	99%	44.7	62.4	91%
22-23	85%	45%	50%	73%	25.7	25.4	50%	63%	39%	87%	99%	45.9	64.9	91%

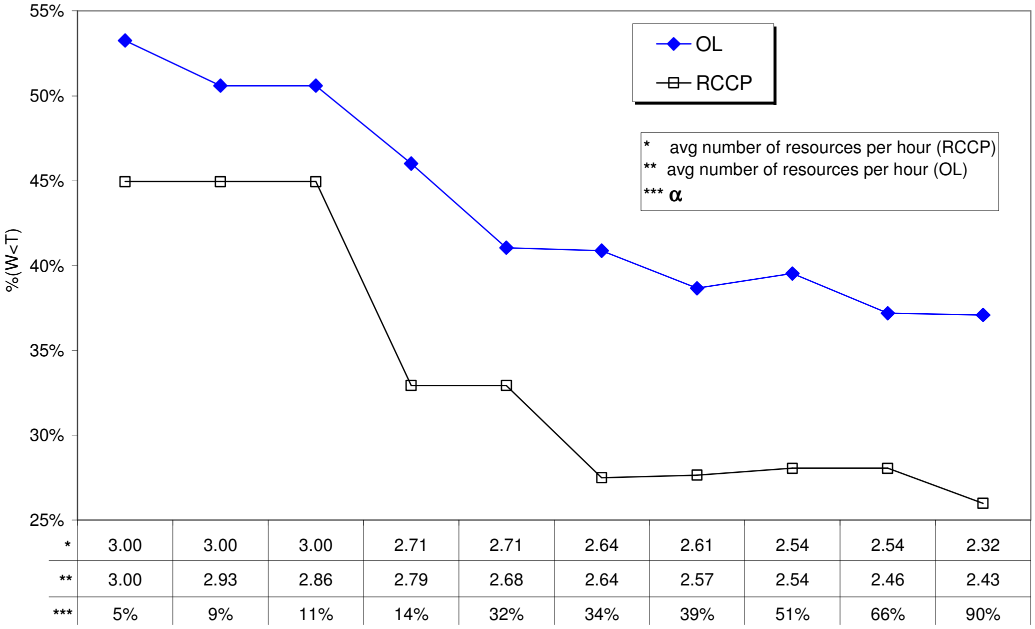
Table IV. Standard Deviation of Performance Measures (Using OL and RCCP)

	Performance measures using OL recommendation							Performance measures using RCCP recommendation						
Hour	Resource utilization				#Beds	Chairs	*%(W>T)*	Resource utilization				#Beds	#Chairs	*%(W>T)*
	Ip	Sp	Op	Nu				Ip	Sp	Op	Nu			
16-17	1.7%	3.0%	2.9%	2.2%	0.8	1.0	2.8%	1.5%	3.4%	3.8%	2.6%	0.7	0.9	3.1%
17-18	2.1%	3.0%	4.1%	2.8%	0.8	1.2	3.5%	2.0%	3.3%	23%	3.0%	0.7	1.1	3.6%
18-19	1.9%	2.7%	2.1%	2.4%	0.9	1.3	3.8%	2.3%	3.0%	2.6%	2.5%	0.8	1.2	3.7%
19-20	2.0%	2.8%	2.0%	2.3%	1.0	1.4	3.9%	2.2%	2.8%	4.4%	2.4%	0.9	1.3	4.0%
20-21	2.0%	3.0%	2.2%	2.7%	1.0	1.4	3.7%	1.6%	2.6%	2.9%	1.3%	0.9	1.4	3.5%
21-22	1.9%	2.9%	2.2%	2.1%	1.1	1.5	3.5%	1.4%	2.6%	2.5%	1.1%	1.0	1.6	3.4%
22-23	1.8%	3.5%	5.3%	3.7%	1.1	1.6	3.4%	1.8%	2.5%	2.3%	1.4%	1.1	1.8	3.2%

Table IV presents the standard deviations of performance measures calculated in Table III. We observe that these values are relatively small. The standard deviations in the other numerical experiments are of the same order.

5.5.4 Comparing RCCP and OL Given the Same Average Number of Resources

In this section, we provide a “fair comparison” between RCCP and OL staffing techniques. The same simulation model for the same time period, as in Sections 5.5.2 and 5.5.3, was used. However, in the previous sections, we allowed a different amount of resources for the two methods, obtaining better results for OL with more resources. Here we targeted the two staffing methods to use the same average number of resources (Ip, Sp, Op, and Nu) per hour. We used the following algorithm to reach this goal. First, different values of the targeted service level *α=%(W>T)* were used to get the OL recommendations on the number of resources per hour, via Equations (4) and (5). The overall average utilization was computed for each case. Then we modified the overall number of resources in the RCCP Equation (1), in order to target the same values of the overall average utilization. Finally, simulations were run in order to compare the quality of service *%(W < T)* for the two methods; the results are presented in Fig. 4. The simulation results are conclusive — the OL is the superior method, which implies the higher quality of service with the same number of resources for all values of α.



*	3.00	3.00	3.00	2.71	2.71	2.64	2.61	2.54	2.54	2.32
**	3.00	2.93	2.86	2.79	2.68	2.64	2.57	2.54	2.46	2.43
***	5%	9%	11%	14%	32%	34%	39%	51%	66%	90%

Fig. 4. Quality of service of RCCP and OL with the same number of resources per hour.

6 TACTICAL HORIZON: SIMULATION-BASED MODELING FOR THE CONTROL OF SEASONAL LOAD EFFECTS IN THE ED

Although the patient intraweek arrival pattern does not change over time, there are midterm load effects (e.g. flu epidemic months) that must be addressed when one plans and schedules the ED resources. Assume that we have an arrival load forecast for a certain time period. Our goal is to calculate hourly staffing recommendations. For this goal, we do not need an on-line simulation, and we can look on the average effects of a model, which uses the OL number of resources per hour, and a model, which uses RCCP recommendations. For a fair comparison, we forced the total number of resource-hours for both methods to be the same. The same technique as in Section 5.5.4 was used for this purpose. One hundred simulations for each special case with a three-day warm-up period were performed. The only difference with respect to the on-line-simulation was that here we used a simulation model with shared physicians instead of specific ones for simplicity reasons. We compared the two staffing methods with respect to the following performance measures: *%(W>T)*; average length of stay (ALOS); and number of average occupied chairs and beds. We fixed ten values of the targeted service level α (from 0.1 to 1.0 with a step 0.1), got OL recommendations for the number of resources and, then, calculated RCCP recommendations with the same overall utilization. We ran the simulation again to receive the quality of service for comparison. The results are presented in Table V.

Table V. Simulation Performance Measures Using OL and RCCP (Off-line)

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Hourly average *%(W>T)* (OL)	6.3%	10.4%	13.9%	16.9%	20.5%	21.2%	24.6%	26.0%	27.8%	31.0%
Hourly stdev *%(W>T)* (OL)	17.4%	23.0%	26.7%	29.6%	32.3%	33.2%	34.8%	35.6%	36.5%	38.4%
Hourly average *%(W>T)* (RCCP)	15.7%	21.6%	23.9%	26.5%	29.4%	30.9%	35.6%	36.3%	39.1%	42.3%
Hourly stdev *%(W>T)* (RCCP)	30.9%	35.5%	36.9%	38.4%	39.7%	40.2%	41.9%	42.1%	42.9%	43.6%
Average *%(W>T)* (OL)	6.4%	10.5%	14.3%	17.3%	21.1%	21.8%	25.9%	27.0%	28.7%	31.5%
Average *%(W>T)* (RCCP)	11.2%	16.5%	18.4%	21.1%	23.0%	24.7%	29.1%	30.5%	32.9%	36.1%
ALOS(OL)	200.9	211.2	221.5	227.6	232.5	237.7	241.1	245.8	253.0	254.7
ALOS(RCCP)	211.2	226.2	238.9	244.6	251.8	256.6	267.7	270.6	279.4	291.4
Average Beds(OL)	13.4	14.0	14.4	14.9	15.2	15.1	15.7	15.9	16.0	16.4
average Chairs(OL)	9.7	10.7	11.5	11.9	12.5	12.3	13.0	13.3	13.4	14.1
average Beds(RCCP)	14.2	14.9	15.4	15.9	16.3	16.3	17.5	17.4	18.1	18.3
average Chairs(RCCP)	10.6	11.6	12.2	12.7	13.1	13.2	14.4	14.4	15.0	15.4

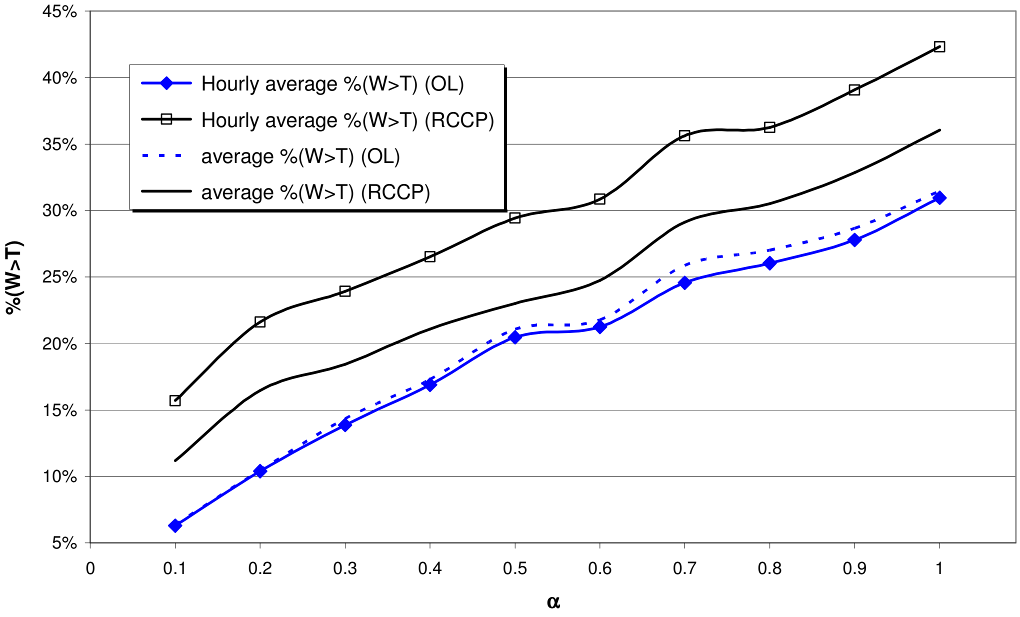


Fig. 5. Quality of service of RCCP and OL by using a similar number of resources per hour (off-line).

In Fig. 5 we observe that if the comparison is done over *%(W>T),* OL is dominating RCCP by 5% approximately if averages over all patients are compared, and by 10% if hourly averages are compared. (In the latter case, we first calculate performance for each hour and then average the results.) The superiority of the OL approach is also clear for ALOS, and for the average occupied beds and chairs indices. If the performance is analyzed on an hourly basis, we observe that the OL approach is not always dominant. It can be shown that the number of resources per hour is not too different for the two methods. For example, see Fig. 6 for α = 0.3 on an average day, where *R*(OL,Dr) and *R*(OL,Nu) mean the offered-load (3) for physicians and nurses, respectively; RCCP(Dr) and RCCP(Nu) denote the expected processing time per resource (1); and, finally, *n*(OL,Dr), *n*(OL,Nu), *n*(RCCP,Dr) and *n*(RCCP,Nu) denote staffing levels for a corresponding method and resource type.

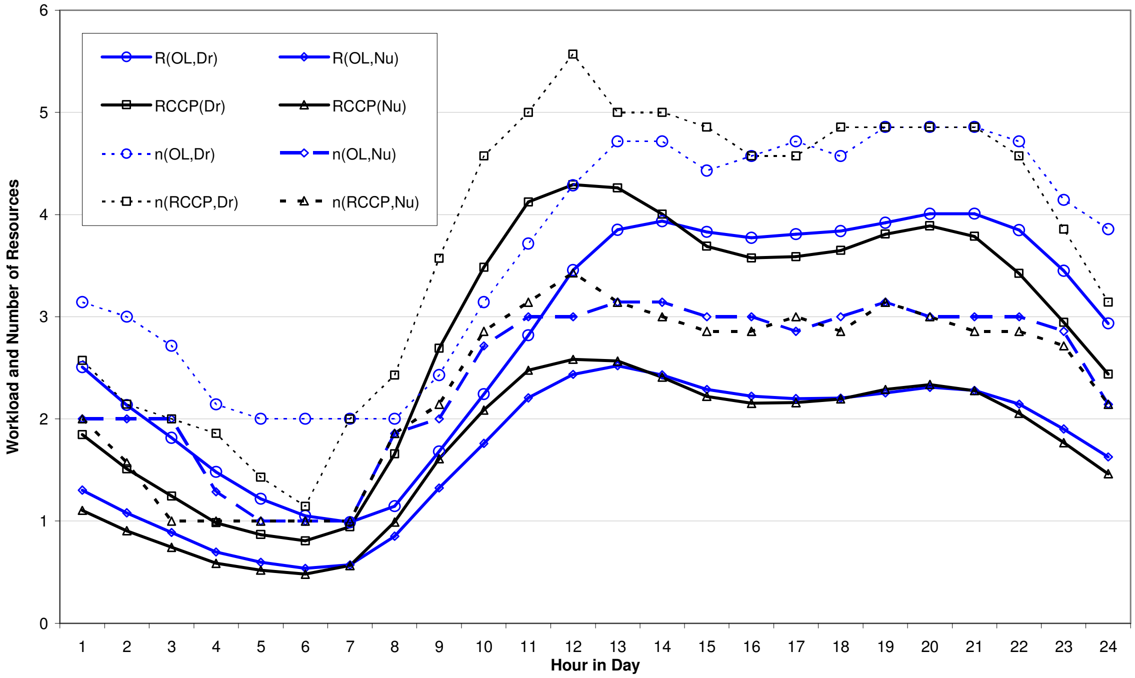


Fig. 6. *n* and *R* per average Hour of a day (α = 0.3) (off-line).

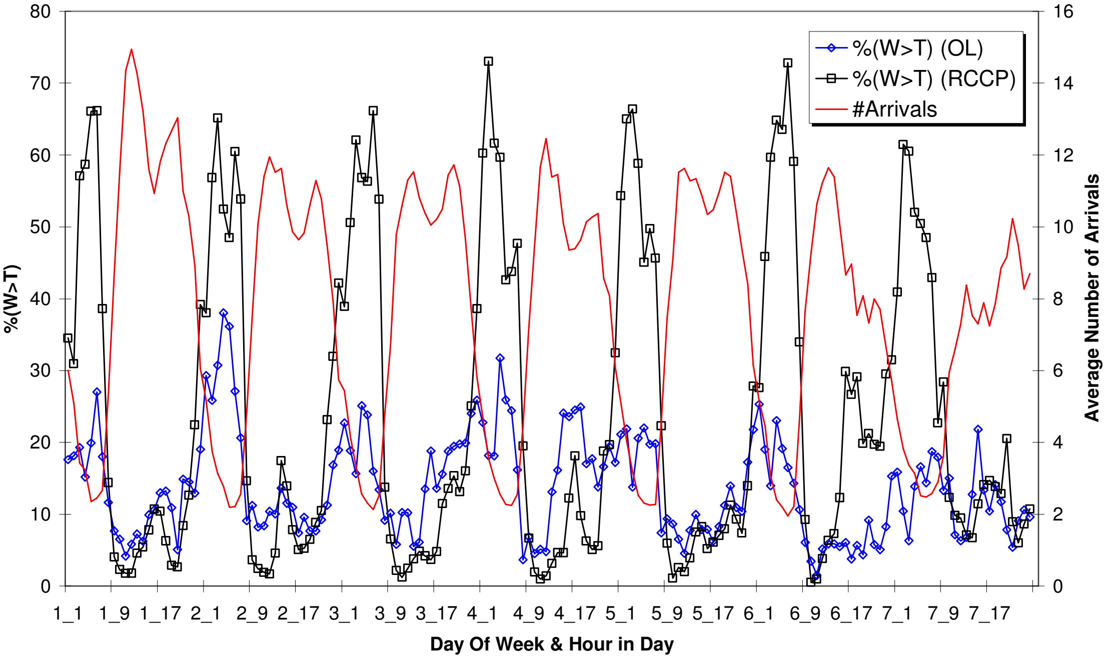


Fig. 7. %(W>T) (and #Arrivals) per Hour by Method in an Average Week (α = 0.3) (off-line).

In Fig. 7 (α = 0.3), we observe that OL maintains a steady quality of service during the week, while RCCP is gaining slightly better results during increasing arrival rate periods and fails when the arrival rate declines. The reason is the economies-of-scale phenomenon, which is well-known in queuing theory. RCCP targets the utilization level, but a system with a larger number of servers provides a better performance given the same utilization.

Summarizing, the OL method provides better and more stable performance. Since tactical planning is performed weeks or months in advance, it is much easier to schedule the needed workforce for the tactical horizon than in the case of operational planning. A possible limitation of tactical planning is related to forecast reliability. Say, if load forecasting quality for flu epidemic periods is low, the staffing recommendations will be far from optimal.

7 STRATEGIC HORIZON: SIMULATION-BASED MODELING FOR ED REDESIGN UNDER PHYSICAL RELOCATION

7.1 Background

The managers of Rambam hospital came to the conclusion that the previous ED could not provide a sufficient service quality given the increasing load and growing demands to clinical and operational service levels. Therefore, the decision to design and construct a new ED was taken, in order to create a more comfortable environment for patients and hospital staff. The ED transfer is implemented in two stages. First, in 2008, the ED was transferred to a temporary location in the basement of the hospital. In 2010, a new permanent ED will be opened at the same location as the previous ED. Both transfers increase uncertainty concerning many issues related to ED functioning. Two undergraduate student projects were performed in order to help hospital management deal with this uncertainty and provide recommendations on the issues that are still open. Here we concentrate on the first project that was dedicated to the transfer from the previous location to the temporary location.

In both locations, patients were classified either as acute or as walking. The process structure is similar for both types of patients. First, they are sent to triage, classified as internal, orthopedic or surgical (orthopedic and surgical patients are referred to as trauma patients) and then transferred to the corresponding room of the ED. Here a nurse performs initial checks and a physician provides an initial assessment. The patient is then sent to additional tests if needed. When the results of the tests arrive, a physician assesses the patient once again and, unless additional tests are required, decides either to transfer him/her to the internal ward or to release the patient home.

The area of the temporary ED is significantly larger than the area of the previous ED: 2,000 square meters versus 1,000, respectively. Hence the design in both cases had to be different. For example, the problem of large walking distances can potentially arise for the temporary ED. The number of physicians and nurses that work in the two EDs was initially assumed the same and, due to large distances between walking and acute patients of trauma type, there was a special need to evaluate the walking distances of physicians that had to treat both types of patients.

In addition, the nurses' schedule had to be changed after the transfer. In the previous state, different teams of nurses treated internal and trauma walking patients. Moreover, all trauma patients were treated by the same team. In the temporary ED, all walking patients are concentrated at the same location, so-called ambulatory ED, and a single nursing team treats them. Therefore, the need to compare different configurations of nursing teams arose.

Another important problem was related to the X-Ray unit. This is an important issue since approximately two thirds of ED patients are sent for an X-Ray check. In the previous location, ED had its own X-Ray room that functioned between 8am and 2pm. During these hours, 42% of ED patients that had to perform the check were sent to this room and the others were sent to the external X-Ray room that gave service to all hospital wards. During the period when the ED X-Ray room was closed, all patients were sent to the external X-Ray. In the temporary state, it was planned initially to prolong the working hours of the ED X-Ray room to 10 hours or even to 24 hours per day. Validation of this preliminary decision has been an important issue.

In order to explore these challenges, our main simulation model [Sinreich and Marmor 2005, 2004] has been used. Since it was not feasible to perform arbitrary staffing changes in the temporary ED, we used the straightforward simulation approach, instead of the offered-load approach, applied in the previous sections.

7.2 Nurse Staffing in the Temporary ED

The ambulatory ED room, where all walking patients are located, brought an important change in the temporary ED design with respect to the previous state. Initially, it was assumed that the team dedicated to internal walking patients in the previous state would be able to treat all walking patients in the ambulatory room. This team consisted of a single full-time nurse and a second nurse, added during several high-loaded hours only. Our simulation analysis demonstrated that, in order to sustain a reasonable service level under new conditions, it is necessary to add a second nurse during most hours of the day.

Table VI. Average Length of Stay under Different Scenarios in an Ambulatory Room

Patient type	Second nurse added 9 hours/day		Second nurse added 18 hours/day		Difference is statistically significant
	ALOS, min	CI	ALOS, min	CI	
Acute internal	441.20	128.78	438.31	98.49	No
Acute surgical	161.67	33.18	165.54	33.76	No
Acute orthopedic	170.46	17.96	173.03	32.86	No
Walking internal	455.70	174.78	272.75	107.72	Yes
Walking surgical	328.38	62.88	176.97	14.96	Yes
Walking orthopedic	392.05	90.46	194.72	47.48	Yes
Overall	381.59		276.95		

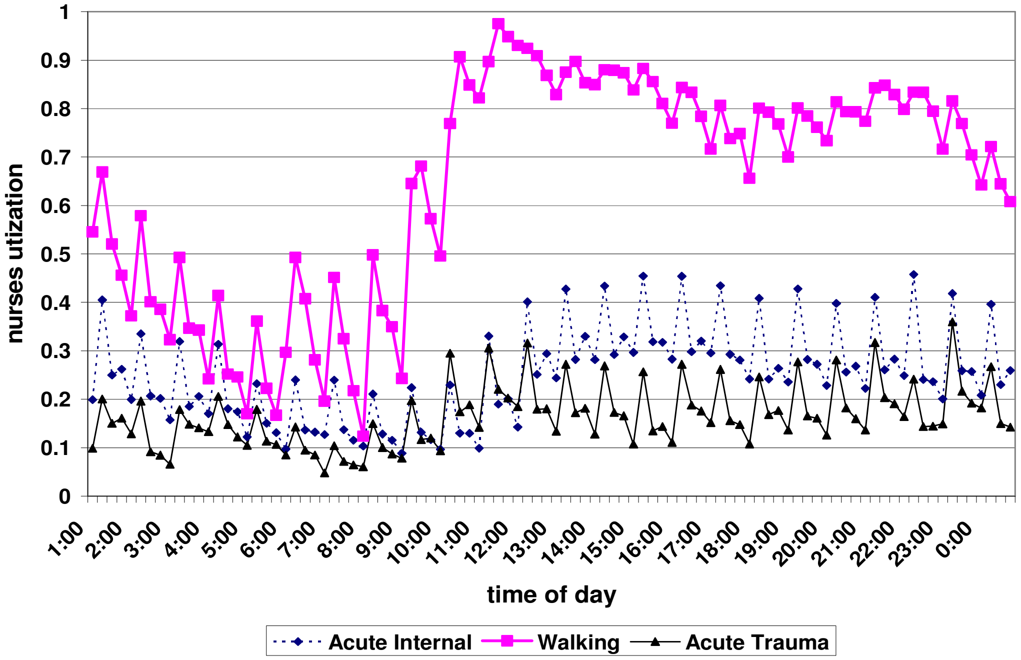


Fig. 8. Workload of nursing teams.

Table VI compares two cases: a second nurse is added in the Ambulatory room for 9 and 18 hours, respectively. In the second case, two nurses are working in the unit during all hours of the day, except for the 5am–11am period. The difference with respect to ALOS of Walking patients and, hence, overall ALOS, is striking. (CI is the width of 95% confidence intervals.)

However, even after a second ambulatory nurse is added for 18 hours per day, the nursing team that treats walking patients remains overloaded. Fig. 8 illustrates this phenomenon, comparing workload per nurse of the three nursing teams. (There are four nurses in the acute internal team, three nurses most of the day in the acute trauma team and two nurses most of the day in the ambulatory walking team.) Note that around 12:00–13:00 the workload of the ambulatory nurses is too close to one: nurses would work in a heavily overloaded regime probably implying undesirable consequences from the operational service-level and clinical points of view.

Since an additional workforce was not available, it was decided to transfer a nurse from one of the other two teams to the ambulatory team. Transfers from acute internal and acute trauma were modeled via our simulator; it turned out that the transfer from the trauma team is slightly more preferable. This conclusion could be expected from Fig. 8, where we observe that the acute trauma team is the least loaded one. Fig. 9 shows the workload of the nursing teams after the transfer: now the load on the walking team is reasonable.

Following this research, our specific recommendations on a nurse transfer from trauma unit to ambulatory unit and on addition of a second nurse to ambulatory unit for 18 hours per day were actually implemented.

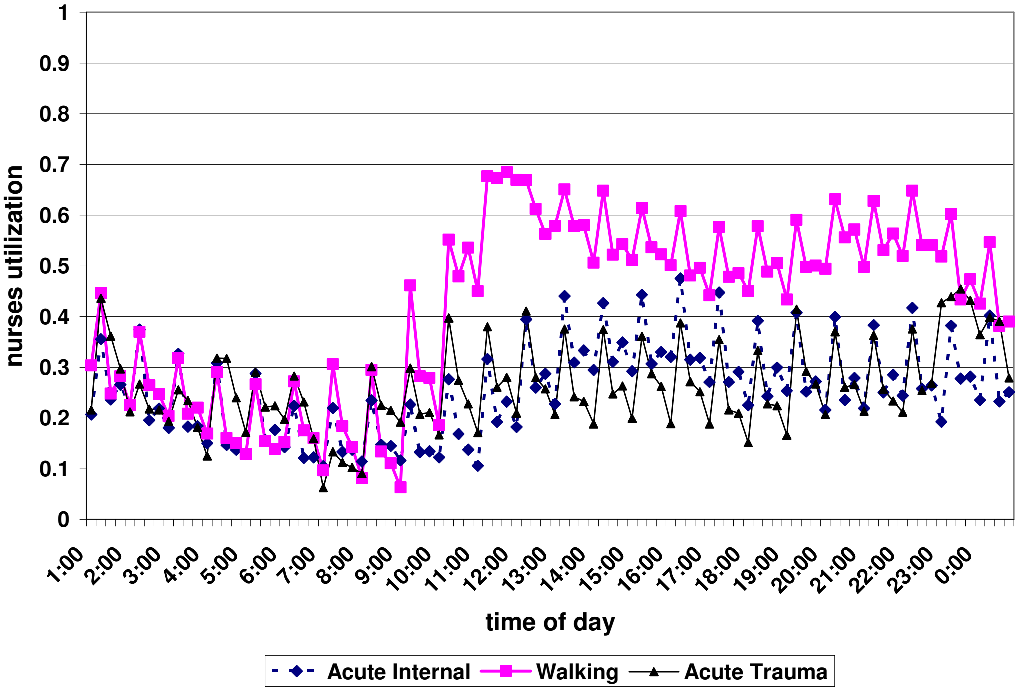


Fig. 9. Workload of nursing team after transfer of one nurse from Trauma to Walking.

7.3 Summary of Proposed Changes in the Temporary ED and Assessing the Expected Gain

In addition to the changes in nurses staffing, other recommendations were provided in order to improve functioning of the temporary ED. Below we summarize the three most significant recommendations:

- Simulation experiments were performed to compare several options for the ED X-Ray room's opening hours. We compared between 8am–6pm, 8am–2pm (current practice), and 12am–6pm options. It turned out that 12am–6pm opening hours imply small mean waiting times and queue lengths during the day. For example, the mean queue length to X-Ray rooms during the day never exceeds two patients. In addition, the utilization of the ED X-Ray room under this scenario is stable, slightly changing around 90%. In contrast, under the alternative scenarios, the X-Ray unit would be underloaded in the early hours of the morning. This conclusion can be explained by the high load on X-Ray units during late afternoon hours: if the ED X-Ray is closed during these hours, the load on the external X-Ray is very high. Simulation was also used to determine the optimal threshold value for the queue length in the ED X-Ray. If the queue length in the ED X-Ray reached this value (eight turned out to be optimal value that minimized average waiting of patients), the patients were sent to the external X-Ray.
- Pooling treatment policy was studied for the physicians that treat internal acute patients. The pooling policy suggests that each internal physician can treat each patient in the internal acute room. The alternatives to this approach suggest that responsibility on different groups of internal acute patients, located in several sectors of the internal acute room, is divided between different physicians. We had to analyze the tradeoff between well-known advantages of pooling and potentially larger walking distances of physicians. Since the increase in the walking distances for the pooling option was not significant, a decrease of ALOS was observed and the pooling method was chosen.
- Surgery and orthopedic physicians treat both acute and walking (ambulatory) patients. It turned out that the walking distance between the ambulatory room and the room where acute trauma patients were located is relatively long. Our simulations have shown that walking time for these types of physicians can reach 6.5% of overall time that they spend in the hospital. In order to eliminate this undesirable factor, we suggested a new design of the temporary ED, which significantly decreased the walking distances of the physicians. In addition, we decreased the walking distances of the nurses via optimal locations of the shelves with the medicine.

Table VII displays ALOS that has been calculated using our simulation models in the three cases. The first column displays ALOS in the case when internal physicians are not pooled. Several non-pooled working protocols were compared and here we display the output of the one that gives the best results. The second column shows the state after pooling was performed; we observe that ALOS of acute internal patients decreased significantly. Finally, the third column displays ALOS for the optimal state, where all our suggested design changes were performed. We observe a very significant improvement of service level for some patient types with respect to the pooling scenario.

The number of beds needed in an ED is an additional important operational metric. It turns out that our redesign improves this metric significantly. For example, consider the following metric: number of beds in the internal unit that is enough to absorb the load 95% of time. Our simulation experiments demonstrate that this metric decreases from 43 to 36 beds on weekdays and from 28 to 26 beds on weekends.

Table VII. Average Length of Stay under Different ED Design

Patient type	Internal physicians are not pooled	Internal physicians are pooled	Optimal design
	ALOS, min	ALOS, min	ALOS, min
Acute internal	489.66	438.31	389.64
Acute surgical	167.60	165.54	157.01
Acute orthopedic	172.00	173.03	166.98
Walking internal	264.70	272.75	212.07
Walking surgical	166.20	176.97	123.49
Walking orthopedic	185.74	194.72	144.11
Overall	286.30	276.95	230.62

8 DISCUSSION

Below we discuss the theoretical and practical impact of our research, for each of the three staffing horizons within which we worked.

Online Decision Support, Short-term Forecasting and Operational Planning. In Section 5, we have shown how the setup problems for the staffing algorithm are solved in this case. We emphasized simulation-based inference of the current state and, especially, the specific problem of inferring patient discharge times, which seems to be a widespread example of incomplete data. Starting with a trustable estimate of the current ED state, we were able to simulate future ED evolution and thus generate staffing recommendations, based on the simulated offered-load.

The Offered-Load (OL) Framework. We believe that the offered-load framework for staffing constitutes the main theoretical contribution of the paper. The method is applicable over all planning horizons. It is based on simulation with infinite resources and thus generalizes, to a complex ED service network, the single-station approach of Feldman et al. [2008]. We compared the offered-load method with the prevalent RCCP technique in several different setups and, overall, staffing based on the offered-load implies better performance given comparable resources. The main reason for this outcome is that the offered-load concept refines RCCP in the sense that it allocates workload accurately over time, while RCCP accounts for *all* the workload brought in by a patient right at the arrival time of that patient.

Simulation-based Staffing over a Tactical Horizon. In Section 6, we considered the problem of midterm staffing weeks or months ahead. The two simulation-based staffing methods, OL and RCCP, were compared here as well. As indicated, for all considered performance measures, which included ALOS, probability of a long wait for the first physician encounter and average number of occupied beds, the OL approach turned out to be preferable.

Scenario Analysis and Strategic Planning. In Section 7, we considered an actually implemented decision of our partner hospital, namely to establish a new ED and to move the ED operations to a temporary location while the future permanent ED is redesigned and rebuilt. Our simulation-based approach was used to evaluate the consequences of the first ED transfer, focusing mainly, but not solely, on resource staffing questions. This approach turned out instrumental for validation of specific changes that had been planned by hospital management and for scenario analysis of alternatives.

9 CONCLUSIONS AND PROPOSALS FOR FUTURE RESEARCH

In this paper, we applied a simulation model of an emergency department to staff scheduling problems over several different time horizons. The results turn out to be useful and promising. Our approach helped our partner hospital to solve strategic planning problems that arose during ED relocation. In addition, we introduced a simulation-based offered-load staffing technique that performs better than a prevalent alternative. This combination of a flexible simulation model and of an advanced staffing technique can certainly be used in other hospitals. In order to enhance our approach, it would be helpful to design IT systems that integrate these tools with real-time decision support systems and RFID technology.

Since this research covers several heterogeneous topics, many future research directions can arise out of it. For example:

- *Research on RFID Implementation in EDs.* RFID technologies could be very helpful for real-time control and operational planning in Emergency Departments. However, an RFID implementation would affect many aspects of ED functioning and, hence, it should be justified from operational, financial and ethical perspectives. We are presently involved in a research project on this issue.
- *ED Design and Redesign.* Numerous practical and research challenges exist regarding ED design (or redesign, as a consequence of moving the ED to a new location). For example, the new Rambam ED will employ multi-skilled EM (Emergency Medicine) physicians, in contrast to its previous scheme where internal and trauma patients were treated by expert physicians. Simulation-based analysis of the tradeoffs between these two options, as well as additional alternatives for operational design, is ongoing.
- *Enhancing Forecasting Algorithms.* In this paper, a simple MA technique was used to forecast arrival volume since we failed to improve its goodness-of-fit vs. more sophisticated approaches. However, this issue deserves additional research effort. For example, an alternative approach to arrival load forecasting is presented in Kuhl, Sumant and Wilson [2006], Kuhl and Wilson [2000], Kuhl, Wilson and Johnson [1997], where the authors estimate the parametric rate function of a non-homogeneous Poisson process. Verifying if such methods provide a better goodness-of-fit to our data is an interesting research topic.
- *Integration between ED Simulators and Hospital Data Repositories.* The Service Engineering Enterprise (SEE) Center at the Faculty of Industrial Engineering and Management in the Technion has created and maintained data repositories from service systems. These are all based on the DataMOCCA model (Data Model for Call Centers Analysis, see Trofimov et al. [2006]). The model provides a uniform presentation of operational data from various sources for statistical analysis, operations research and simulation. Initially designed for call center data storage and processing, DataMOCCA was generalized to accommodate other sources and types of data, including healthcare data in general, and ED in particular. Indeed, the SEE data repository, partially available at http://ie.technion.ac.il/Labs/Serveng/, now includes data from EDs and internal wards of several hospitals. Such data provides a testing ground for integrating an ED simulator with a hospital's data-base, so that the simulator can be operated in real- or near real-time.

REFERENCES

ASMUSSEN, S., AND GLYNN, P.W. 2007. *Stochastic Simulation*. Springer, New York.

BADRI, M.A., AND HOLLINGSWORTH, J. 1993. A simulation model for scheduling in the emergency room. *International Journal of Operations & Production Management,* 13, 13–24.

BEAULIEU, H., FERLAND, J.A., GENDRON, B., AND MICHELON, P. 2000. A mathematical programming approach for scheduling physicians in the emergency room. *Health Care Manage Science,* 3, 193–200.

BILLER, B., AND NELSON, B.L. 2002. Answers to the top ten input modeling questions. In *Proceedings of the 2002 Winter Simulation Conference*. Yücesan, E., Chen, C.-H., Snowdon, J.L., and Charnes, J.M. (Eds), 35–40. Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

BORST, S., MANDELBAUM, A., AND REIMAN, M. 2004. Dimensioning large call centers. *Operations Research,* 52(1), 17–34.

BROWN, L., GANS, N., MANDELBAUM, A., SAKOV, A., ZELTYN, S., ZHAO, L., AND HAIPENG, S. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association,* 100, 36–50.

CHANNOUF, N., L'ECUYER, P., INGOLFSSON, A., AND AVRAMIDIS, A.N. 2007. The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Manage Science,* 10, 25–45.

DERLET, R.W., AND RICHARDS, J.R. 2000. Overcrowding in the nation's emergency departments: complex causes and disturbing effects. *Annals of Emergency Medicine,* 35, 63–68.

DRAEGER, M.A. 1992. An emergency department simulation model used to evaluate alternative nurse staffing and patient population scenarios. In *Proceedings of the 1992 Winter Simulation Conference*. Swain, J.J., Goldsman, D., Crain, R.C., and Wilson, J.R. (Eds) 1057–1064. Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

FELDMAN, Z., MANDELBAUM, A., MASSEY, W., AND WHITT, W. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science,* 54, 324–338.

FUJIMOTO, R., LUNCEFORD, D., PAGE, E., AND UHRMACHER, A.M. 2002. Grand Challenges for Modeling and Simulation. Technical Report No. 350, Schloss Dagstuhl.

GARCÍA, M.L., CENTENO, M.A., RIVERA, C., AND DECARIO, N. 1995. Reducing time in an emergency room via a fast-track. In *Proceedings of the 27th Conference on Winter Simulation*, 1048–1053, Arlington, VA, USA, December 3–6.

GARNETT, O., MANDELBAUM, A., AND REIMAN, M. 2002. Designing a call center with impatient customers. *Manufacturing and Service Operations Management*, 4(3), 208–227.

GREEN, L.V. 2008. Using Operations Research to reduce delays for healthcare. In *Tutorials in Operations Research*. Chen, Zhi-Long and Raghavan, S. (Eds), 1–16. INFORMS, Hanover, MD.

GREEN, L.V., KOLESAR, P.J., AND SOARES, J. 2001. Improving the SIPP approach for staffing service systems that have cyclic demand. *Operations Research,* 49, 549–564.

GREEN, L.V., KOLESAR, P.J., AND WHITT, W. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management,* 16, 13–39.

HALFIN, S., AND WHITT, W. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research,* 29, 567–588.

HALL, R.W. 2006. *Patient Flow: Reducing Delay in Healthcare Delivery*. Springer.

HUANG, S.Y., CAI, W., TURNER, S.J., HSU, W.J., ZHOU, S., LOW, M.Y.H., FUJIMOTO, R., AND AYANI, R. 2006. A generic symbiotic simulation framework. In *Proceedings of the 20th Workshop on Principles of Advanced and Distributed Simulation*. Ceballos, S. (Ed), 131. IEEE Computer Society, Washington, DC.

JACOBSON, S.H., HALL, S., AND SWISHER, S.R. 2006. Discrete-event simulation of health care systems. In *Patient Flow: Reducing Delay in Healthcare Delivery*, Hall, R.W. (Ed), 211–252. Springer USA.

JANSSEN, A.J.E.M., VAN LEEUWAARDEN, J.S.H., AND ZWART, B. 2008. Refining Square Root Safety by Expanding Erlang C. Technical Report (http://www.win.tue.nl/~jleeuwaa/paper20.pdf).

JUN, J.B., JACOBSON, S.H., AND SWISHER, J.R. 1999. Application of discrete-event simulation in health care clinics: a survey. *Journal of the Operational Research Society*, 50, 109–123.

KHURMA, N., BACIOIU, G.M., AND PASEK, Z.J. 2008. Simulation-based verification of lean improvement for emergency room process. In *Proceedings of the 2008 Winter Simulation Conference*. Mason, S.J., Hill, R.R., Mönch, L., Rose, O., Jefferson, T., and Fowler, J.W. (Eds), 1490–1499. Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

KING, D.L., BEN-TOVIM, D.I., AND BASSHAM, J. 2006. Redesigning emergency department patient flows: application of Lean thinking to health care. *Emergency Medicine Australasia,* 18, 391–397.

KOLB, E.M.W., PECK, J., SCHOENING, S., AND LEE, T. 2008. Reducing emergency department overcrowding - five patient buffer concepts in comparison. In *Proceedings of the 2008 Winter Simulation Conference*. Mason, S.J., Hill, R.R., Mönch, L., Rose, O., Jefferson, T., and Fowler, J.W. (Eds), 1516–1525. Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

KUHL, M.E., SUMANT, S., AND WILSON, J.R. 2006. An automated multiresolution procedure for modeling complex arrival processes. *INFORMS Journal on Computing,* 18(1), 3–18.

KUHL, M.E., AND WILSON, J.R. 2000. Least squares estimation of nonhomogeneous Poisson processes. *Journal of Statistical Computation and Simulation,* 67, 75–108.

KUHL, M.E., WILSON, J.R., AND JOHNSON, M.A. 1997. Estimating and simulating Poisson processes having trends or multiple periodicities. *IIE Transactions,* 29(3), 201–211.

KULJIS, J., PAUL, R.J., AND STERGIOULAS, L.K. 2007. Can health care benefit from modeling and simulation methods in the same way as business and manufacturing has? In *Proceedings of the 2007 Winter Simulation Conference*. Henderson, S.G., Biller, B., Hsieh, M.H., Shortle, J., Tew, J.D., and Barton, R.R. (Eds), 1449–1453. Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

LIYANAGE, L., AND GALE, M. 1995. Quality improvement for the Campbelltown hospital emergency service. In *IEEE International Conference on Systems, Man, and Cybernetics*. Gruver, W.A., Fraser, S., and de Silva, C.W. (Eds), 1997–2002. Institute of Electrical and Electronic Engineers, Vancouver, British Columbia, Canada.

MAMAN, S., MANDELBAUM, A., AND ZELTYN, S. 2009. Uncertainty in the demand for service: the case of call centers and emergency departments. Research in progress.

MARMOR, Y., SHTUB, A., MANDELBAUM, A., WASSERKRUG, S., ZELTYN, S., MESIKA, Y., GREENSHPAN, O., AND CARMELI, B. 2009. Toward simulation-based real-time decision-support-systems for emergency departments. *2009 Winter Simulation Conference,* December 13–16, Austin, TX.

MCNEIL, A., FREY, R., AND EMBRECHT, P. 2005. *Quantitative Risk Management*. Princeton University Press.

MEDEIROS, D.J., SWENSON, E., AND DEFLITCH, C. 2008. Improving patient flow in a hospital emergency department. In *Proceedings of the 2008 Winter Simulation Conference*. Mason, S.J., Hill, R.R., Mönch, L., Rose, O., Jefferson, T., and Fowler, J.W. (Eds), 1526–1531. Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

SINREICH, D., AND MARMOR, Y.N. 2005. Emergency department operations: the basis for developing a simulation tool. *IIE Transactions,* 37, 233–245.

SINREICH, D., AND MARMOR. Y. 2004. Emergency Department Operations: A Simple and Intuitive Simulation Tool based on the Generic Process Approach. Technion, Israel Institute of Technology, Technical Report. Available at http://ie.technion.ac.il/Home/Deceased/sinr/emergency2.pdf.

SINREICH, D., AND JABALI, O. 2007. Staggered work shifts: a way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Management Sciences,* 10, 293–308.

TSEYTLIN, Y. 2009. Queueing Systems with Heterogeneous Servers: On Fair Routing of Patients in Emergency Departments. M.Sc. Thesis, Technion. Available at http://ie.technion.ac.il/serveng/References/thesis-yulia.pdf.

TROFIMOV, V., FEIGIN, P.D., MANDELBAUM, A., ISHAY, E., AND NADJHAROV, E. 2006. DATA MOdel for Call Center Analysis: Model Description and Introduction to User Interface. Technion, Israel Institute of Technology, Technical Report.
http://ie.technion.ac.il/Labs/Serveng/files/Model_Description_and_Introduction_to_User_Interface.pdf.

VOLLMANN, T.E., BERRY, W.L., AND WHYBARK, D.C. 1993. *Integrated Production and Inventory Management*. Business One Irwin, Homewood, IL.

WHITE, P.K. JR., 2005. A survey of data resources for simulating patient flows in healthcare delivery systems. In *Proceedings of the 2007 Winter Simulation Conference*, Kuhl, M.E., Steiger, N.M., Armstrong, F.B., and Joines, J.A. (Eds), 04–07. Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

Acknowledgements. This work would not have been possible without the hard work and dedication of the following people: Prof. Rafi Beyar, head of the Rambam hospital; Oded Cohen, head of IBM Haifa Research Lab; Prof. Boaz Golany, Dean of Technion's IE&M Faculty; The Open Collaborative Research project steering committee, first and foremost the Rambam Members - Prof. Shimon Pollack, Dr. Yaron Barel, Dr. Hana Adami and Amir Weiman; The dedicated professionals in the Rambam ED and trauma room, led by Dr. Moshe Michaelson and head nurse Hagar Baruch. We are deeply thankful to Technion undergraduate students Kamel Badran, Rasha Khawaly and Inbal Haas. The outputs of their project, performed under the supervision of Yariv Marmor, were used in Section 7.

Our paper has benefited greatly from the detailed and thoughtful comments by the Editor-in-Chief, a Guest Editor, an Associate Editor and four referees.