Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls, eds.

TOWARD SIMULATION-BASED REAL-TIME DECISION-SUPPORT SYSTEMS FOR EMERGENCY DEPARTMENTS

Yariv N. Marmor

Industrial Engineering and Management
Technion - Israel Institute of Technology
Haifa, Israel

Segev Wasserkrug
Sergey Zeltyn
Yossi Mesika
Ohad Greenshpan
Boaz Carmeli

Haifa University Campus
IBM Haifa Research Labs
Haifa, Israel

Avraham Shtub
Avishai Mandelbaum

Industrial Engineering and Management
Technion - Israel Institute of Technology
Haifa, Israel

ABSTRACT

Emergency Departments (EDs) require advanced support systems for monitoring and controlling their processes: clinical, operational, and financial. A prerequisite for such a system is comprehensive operational information (e.g. queueing times, busy resources,…), reliably *portraying and predicting* ED status as it evolves in time. To this end, simulation comes to the rescue, through a two-step procedure that is hereby proposed for supporting real-time ED control. In the first step, an ED manager *infers* the ED's *current* state, based on historical data and simulation: data is fed into the simulator (e.g. via location-tracking systems, such as RFID tags), and the simulator then completes unobservable state-components. In the second step, and based on the inferred present state, simulation supports *control* by *predicting* future ED scenarios. To this end, we estimate time-varying resource requirements via a novel simulation-based technique that utilizes the notion of *offered-load*.

1 INTRODUCTION

1.1 Objectives

The rising cost of healthcare services has been a subject of mounting importance and much discussion worldwide. Ample explanations have been proposed, yet regardless of their cause, rising costs impose pressures on healthcare providers to improve the management of quality, efficiency, and economics of their organizations.

Hospitals play a central role in the provision of health services and, within hospitals, ED overcrowding has been perhaps the most urgent operational problem (Sinreich and Marmor 2005, Hall 2006, Green 2008). Overcrowding in the ED leads to excessive waiting times and repelling environments which, in turn, cause: (1) Poor service quality (clinical, operational, perceived); (2) Patients in unnecessary pain and anxiety; (3) Negative emotions (of patients and escorts), which sometimes led to violence against staff; (4) Increased risk of clinical deterioration; (5) Ambulance diversion; (6) Patients' LWBS (Leave Without Being Seen); (7) Inflated staff workload; and more (e.g., Derlet and Richards 2000). In order to alleviate this overcrowding and, more generally, to significantly improve the overall management of the ED, the authors of the present work, jointly with a partner hospital, have been involved in an initiative to design and deploy a comprehensive command-and-control solution for EDs (Greenshpan et. al. 2009). One major requirement from such a system is to provide the ongoing detailed operational state of the ED (e.g., queue lengths, waiting times, resource utilizations, etc.). Another important requirement is to provide predictions, several hours (a shift) into the future, based on which intraday operational decisions, such as staffing of nurses and physicians, can be made.

Predictive capabilities, as described above, require a model of the ED. As analytical models are currently unable to capture the complexity of ED operations, one of the major components of our solution is an ED simulation model (as reported in Sinreich and Marmor 2005). Providing short term *predictions* within the context of a command-and-control system has its own unique challenges, in particular an analysis must be performed within a short time frame (in the order of minutes), and must be based on data regarding the current ED state. Moreover, it turns out that in the ED environment, such a simulation model has an additional important role, which is providing estimates regarding the *current* operational state. This is due to

the fact that presently in EDs, available information systems provide only partial, and often inaccurate, information regarding the current operational state. Therefore, an additional capability required by the command-and-control solution is a model-based estimation of the current state.

The focus of this paper is thus the use of simulation modeling for both an on-line estimation of the current operational state, and a short-term prediction regarding future ED states, where the estimation and prediction must be based on incomplete, and sometimes inaccurate, data. Such simulation-based modeling has practical applications in many domains, and it offers a variety of future research opportunities in the area of simulation modeling.

1.2 Alleviating ED overcrowding

The prevalent mean for addressing ED overcrowding is *staff (re)scheduling* (e.g. Sinreich and Jabali 2007, Badri and Hollingsworth 1993, Beaulieu et al. 2000), namely adding or shifting in time staff resources so as to uniformly maintain acceptable ED performance (e.g. time to the First Encounter with a Doctor, or FED time). Such works often focus on off-line steady-state decision making, as opposed to on-line operational and tactical control. Other researchers analyze alternative operational ED *designs* (King, Ben-Tuvim, and Bassham 2006; Liyanage and Gale 1995) – for example, comparing acuteness-driven models (e.g. triage) against operations-driven models (e.g. fast-track, which assigns high priority to patients with low resource requirements). Finally, Green (2008), which we recommend for further references on these and related issues, argues against the prevalent healthcare emphasis on high utilization, rightly raising also the patients' view: quality of care, in particular reduced waiting time as advocated above. In our setting, the aim is to improve the capabilities of intraday staffing, thereby alleviating the pitfalls of overcrowding, in real-time and in the short-term.

1.3 Simulation in support of hospital operations

Due to space limitations, we can only briefly touch on the use of simulation in hospital settings. As previous generations of the present conference manifest, the application of simulation has been instrumental in addressing the multi-faceted challenges that the healthcare domain is presenting (Kuljis, Paul, and Stergioulas 2007). Various problems have been analyzed, such as the improvement of hospital performance (Gunal and Pidd 2008) and patient experience in EDs (Khurma, Bacioiu, and Pasek 2008), staffing optimization (Takakuwa and Wijewickrama 2008) and reduction of overcrowding (Kolb et al. 2008). It is quite common to use simulation, mostly by researchers, to compare operational models (Wong et al. 2003) or to assess a model that addresses a specific research question (Medeiros, Swenson, and DeFlitch 2008). In other cases, simulation has been used in planning processes, determining capacity of resources (Ballard and Kuhl 2006), optimal staffing (Spry and Lawley 2005) and work scheduling (Guo, Wagner, and West 2004). For some reviews of the subject, see Jun, Jacobson, and Swisher (1999), White (2005) and Jacobson, Hall, and Swisher (2006). We are, however, unaware of any uses of simulation in a hospital setting for real time command and control. Nor are we aware of any work in which simulation was used to complete partial data regarding the current operational state.

1.4 Relation to symbiotic simulation, and other topics

Our research gives rise to a multitude of practical and theoretical challenges, many of which touch on active simulation-driven research. But, again, space constraint limits our references to related existing literature: for example input modeling (Biller and Nelson 2002) or historical (trace-driven, resampling) simulation (Asmussen and Glynn 2007; McNeil, Frey, and Embrecht 2005), both related to the problem of properly incorporating *actual* ED data into our simulator.

Yet, deserving of attention 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". For our ED implementations, however, the interaction between the simulator and its underlying physical system must go beyond the common symbiotic simulation framework. Specifically, we obtain real data in real time regarding the current state, then complete the data when necessary via simulation, next predict short-term evolution and workload, and finally proceed with simulation and mathematical models as decision support tools, all this in real time or close to real time.

1.5 Structure of the paper

The rest of the paper is organized as follows. In Section 2, we describe the ED of our partner hospital. Then we outline how simulation can be used for data completion, and provide specific examples in an ED setting in Section 3. Techniques for short term forecasting regarding some aspects of the ED status is then proposed in Section 4. In addition, we describe how such forecasting can be used as the basis for quick intraday ED scheduling. In Section 5, we describe some experiments that tested our methodology. We continue with a brief description of the overall decision support system into which the simulation-based modeling is integrated in Section 6. Section 7 concludes with an outline of some worthy future work.

2 THE ED OF OUR PARTNER HOSPITAL

In Figure 1, we depict two perspectives of the care process that patients undergo in the ED: the resource (i.e. physicians, nurses, etc.) perspective, and the process (activities) perspective. In this care process, two types of *queues* portray the delays that patients experience: first are *resource* queues (rectangular, in red), which are due to limited resources (e.g. nurses, imaging equipment); the second are *synchronization* queues (triangular, in green), which arise when one process activity awaits another (e.g. a patient waiting for results of blood tests and x-ray, in order to proceed with the doctor's examination).

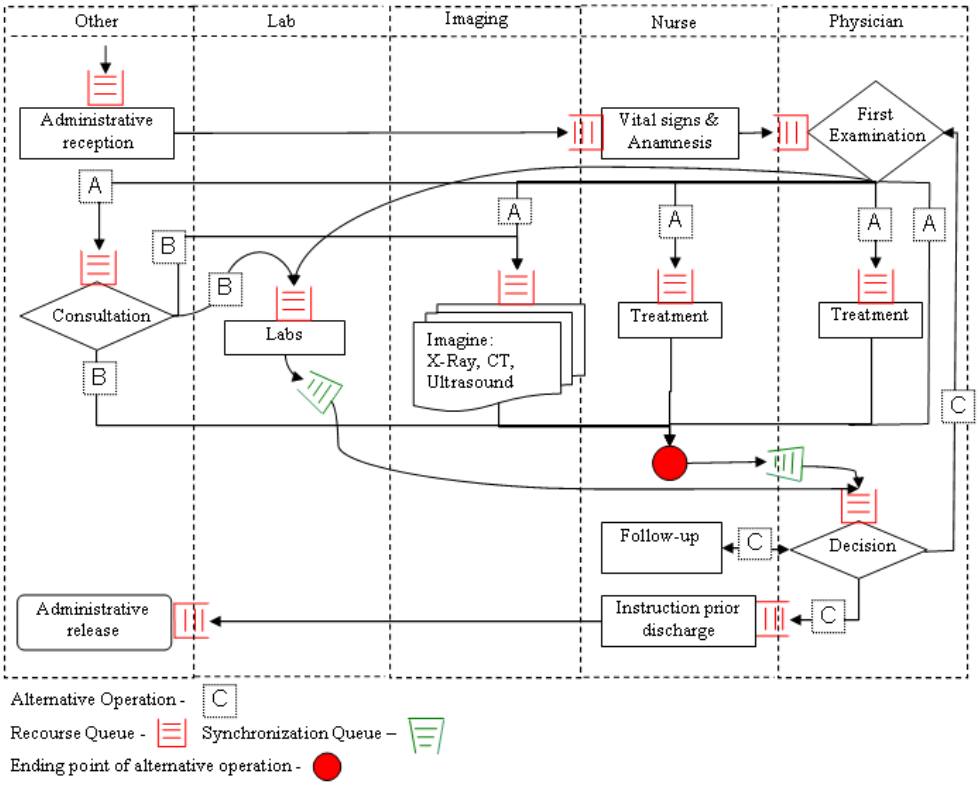

Figure 1: ED resource-process chart

The care process was captured in a simulation model, created with the generic simulation tool of Sinreich and Marmor (2005). In addition to the care process, the simulation model requires data for patient arrival processes, for each patient type, and staffing levels of the medical staff, with their respective skills. For our purpose, the model was configured to the ED specs of our partner hospital, as follows. There are six types of patients, which also require different skills from the caring Physicians. Patient types 1 and 2, which are Internal Acute and Internal Walking respectively, are treated by internal physicians. Patient types 3 and 4, which are Surgical Acute and Surgical Walking respectively, require treatment by surgical physicians. Finally, Patient types 5 and 6, Orthopedic Acute and Orthopedic Walking respectively, require an orthopedic physician. Acute patients need a bed while walking patients use chairs. In addition, data of patient types differ by the arrival process (e.g. number of arrivals per hour and by day-of-week), and by the decisions made in the patient care process (e.g. the percentage of patients sent to X-ray).

3 INFERENCE OF CURRENT STATE

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 queue (number) of patients waiting to be seen by a physician. One expects the amount and quality of usable data to constantly improve over time, due to the introduction of additional data entry systems or new technologies (e.g. sensor technologies, such as 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*. (Note that inaccurate data adds complexity to generating 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 actual arrival data into our model for a long enough period of time to ensure 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 part of the admittance process. The medical state of the majority of arriving patients is such that they actively participate in the registration process, as the first step upon arrival. Registration of the others, acute patients incapable of self-registration, is carried out by the accompanying paramedics shortly after arrival. Therefore, arrivals data accurately captures actual patients' arrival times – it can be thus fed as is into the simulator. (Receptionists also record patient type - Internal, Surgical, or Orthopedic - upon arrival.) To this end, we modified our generic simulator, which originally generates arrivals as a stochastic process (Poisson or relatives). It can now generate realizations consistent with the arrival data, when the latter is fed from an external database.

*Inaccurate data - taking discharges into account***:** Data about patients' discharge (departure) time, 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 bed/chair occupancy and ED staff utilization (due to time spent on catering to these delayed patients). Additional inaccuracies occur due to patients leaving without being see (Green 2008), with or without their medical files, and some other accounting-related reasons.

We found no efficient way for generating simulation realizations that are consistent with our discharge data, except for discarding inconsistent simulation paths. 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 condition on the *number* of patients of each type that were discharged from the ED according to the data. Namely, we considered a (short-term) simulation realization to be consistent if, at the end of the simulation run, the number of patients that were allowed discharge (of each type) equals, to *within* some *accuracy* constant, the number of patients of this type that were discharged according to the data. The results were satisfactory though, clearly, more thought is required here.

4 PREDICTION OF SHORT-TERM FUTURE STATE – REQUIRED STAFFING LEVELS

With the present ED state assumed given (following Section 3), simulation is now to be used for predicting ED evolution, say several hours (a shift, a day) 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 multi-objective problem. It must tradeoff conflicting objectives such as (1) Minimizing costs, (2) Maximizing resource utilization, (3) Minimizing waiting time of patients, (4) Maximizing quality of care. The complexity of such a multi-objective optimization, more so in a stochastic environment (e.g. randomness with respect to patient arrivals, routing, service durations, resource availability, and more) renders the optimization problem intractable analytically. This has thus led researchers to simulation-based heuristic solutions.

One approach is to decompose the problem by focusing only on one type of resource. An example is an effort to schedule nurses while ignoring the scheduling of other resources (Draeger 1992); or scheduling physicians and nurses, one after the other (Sinreich and Jabali 2007). These attempts, based on simulation models, 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, of each patient by each resource type, exactly as in the simulation that we are using.

A prerequisite for staffing is accurate forecasting of patient arrivals, as described in the subsequent Section 4.1. We then continue with predicting resource utilization; this leads to feasible staffing, based on pre-specified goals for resource utilizations (Section 4.2). But the resources' view cannot accommodate the experience of patients – for example, controlling the time till first encounter with a physician (Section 4.3). To control the latter, we calculate, for each resource type, its *offered load* as a function of time; then a classical staffing principle (square-root safety-staffing), in conjunction with the appropriate queueing model, yields our recommended time-varying staffing levels. We conclude, in Section 4.4, with a summary of our methodology.

4.1 Forecasting ED arrivals

For simulating an ED future evolution, one must simulate patient arrivals to the ED. Figure 2, from our partner 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 ED arrivals, leading also to related literature, see Channouf et al. (2007).

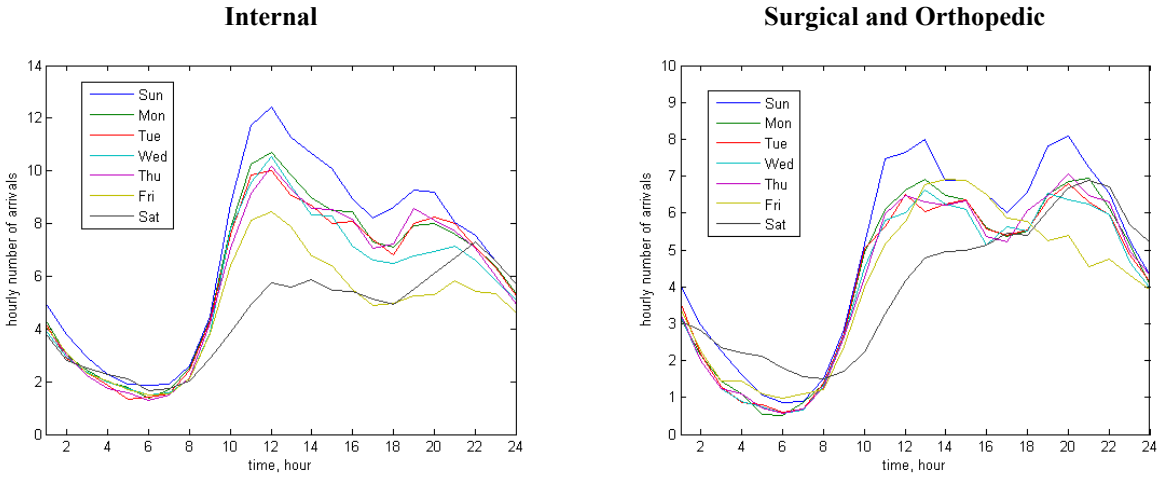

Figure 2: Hourly arrival rates per patient type (averaged over 4 years)

Arrivals in our simulation model are Poisson processes, with hourly rates that are forecasted for each future hour in question (say a shift, or a day) and each patient type. We use long term MA (Moving Averages) in order to predict hourly arrival rates. For example, in order to predict the arrival rate (assumed constant) on Tuesday during 11-12 a.m., we average the corresponding arrival rates during the last 50 "Tuesdays 11-12 a.m.", excluding those that are holidays or days after holidays.

The reason for choosing long-term MA is that we found it provides 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 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 (MLE) estimate for the Poisson parameter, namely the arrival rate.

4.2 First rough solution – RCCP

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 time on 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 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 time required from each resource r at time t:

$$RCCP_r(t) = \sum_i A_i(t) d_{ir}$$

- The recommended number of units of resource r at time t, $n_r(RCCP,t)$, would be the load $RCCP_r(t)$, amplified by safety slack/staffing, or f_s (we have used in our experiments f_s =90%): $n_r(RCCP,t) = RCCP_r(t) / f_s$.

We expect RCCP to achieve pre-planned resource utilization levels; its shortcoming, however, is that it ignores the patients point of view. This is remedied by our next approach.

4.3 Second refined solution: Offered-Load

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 (in order to let 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, 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. 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

$$n = R + \beta\sqrt{R}, \tag{1}$$

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 non-homogeneous and depend on the day-of-week and hour-of-day. Piecewise stationary approximations (such as SIPP - Stationary Independent Period by Period; Green, Kolesar, and Soares 2001), work fine if the arrival rate is slowly-varying with respect to the durations of services. This, however, does not happen in the ED case.

Assume that arrivals can be modeled by a non-homogeneous 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 (equivalently served customers), in a corresponding system with an *infinite* number of servers (Feldman et al. 2008). Specifically, any one of the following four representations gives it:

$$R(t) = E[A(t) - A(t-S)] = E[\lambda(t-S_e)]\cdot E[S] = E\left[\int_{t-S}^{t}\lambda(u)du\right] = \int_{-\infty}^{t}\lambda(u)P(S>t-u)du, \tag{2}$$

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 (2)). Then, for calculating time-varying performance, we recommend to substitute (2) into the corresponding steady-state model, which is the classical M/M/n queue, or Erlang-C, in our case. 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. Our QED approximation then gives rise to

$$1-\alpha = P\{W_q > T\} = P\{W_q > 0\}\cdot P\{W_q > T \mid W_q > 0\} \approx h(\beta_t)\cdot e^{-T\mu\beta_t\sqrt{R_t+\beta_t\sqrt{R_t}}}, \tag{3}$$

where $h(\beta_t)$ is the Halfin-Whitt function (Halfin and Whitt 1981). Equation (3) can now be solved numerically with respect to β_t, and the staffing rule (1) is replaced by the time-varying staffing function:

$$n(OF,t) = R(t) + \beta_t\sqrt{R(t)}. \tag{4}$$

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 are mathematically justified by asymptotic analysis, as workload (and hence the number of servers) increases indefinitely. (The practical motivation was large telephone call centers). However, ample experience (as well as recent research; e.g. Janssen, Van Leeuwaarden, and Zwart 2008) demonstrates amazing levels of high 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 formu-

lae. And indeed, we tested these exact calculations against the QED approximations (in our experiments below), and the sults were essentially unaltered.)

Our proposed offered-load methodology, for ED staffing, proceeds as follows:

- First, we are running the simulation model with infinitely many resources (e.g. physicians, or nurses, or both).
- Second, for each resource *r* (e.g. physician or nurse) and each hour *t*, we calculate the number of busy resources (equals the total work required), 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 $n_r(OL,t)$,via formula (4).

4.4 Methodology for forecasting short-term future ED state

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 run MA, to generate stochastic arrivals in the simulation; (3) Simulate and collect data every hour, for 8 future hours, using infinite resources (nurses, doctors); (4) From step 3, calculate staffing recommendations, both $n_r(RCCP,t)$ and $n_r(OL,t)$; (5) Run the simulation from the current ED state with the recommended staffing; (6) Calculate performance measures. The above can be repeated with existing staffing (in Step (5)), which enables it to be compared to RCCP and Offered-Load staffing.

5 EXPERIMENTS

We now demonstrate our methodologies through simulation experiments. First, we demonstrate the ability of our simulation-based tool to estimate the current ED state, using a database from our partner hospital (Section 5.1). For that, we randomly chose a month (August 2007) in the database, for comparing the known number of patients in the system with the simulation's outcome (following Section 3). In the second experiment (Section 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 a Sunday, which, in Israel, is a busy day of the week, being the first day following the weekend.) We then conclude, in Section 5.3, with a comparison of some ED performance measures, under different staffing methods (again following Section 4).

5.1 Current state

We ran 100 replications of each scenario, in order to compare our simulation results with the hospital's database. For each date and hour, we calculate the average number of patients over the simulation replication (*Avg*), and the corresponding standard deviation (σ), an Upper Bound ($UP = Avg + 1.96\ \sigma$), and a Lower Bound ($LB = Avg - 1.96\ \sigma$). In Figure 3, we depict 4 days, chosen to test our methodology against the (actual) number of patients from the database (*Wip*). These dates were selected for having no holidays during their preceding 7 days. We chose the last day of the weekend and the first working day of the week because they are typically the calmest and busiest, correspondingly. We are demonstrating the comparison on two subsequent weeks. Note that the night and early morning shifts (hours 1-10 in Figure 3) are not overloaded (see the utilization profiles during 09-10, in Table 1), and performance measures are then less accurate. 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 prediction is usefully accurate.

5.2 Forecasting – staffing level

Next, we looked at performance measures in the near future, to see if there is a way to improve ED operations via staffing. We looked at the offered load of all the relevant resources: Internal physician (Ip), Surgical physician (Sp), Orthopedic physician (Op) and Nurses (Nu). For our example, we use ED data until 16:00 and then apply simulation to forecast each succeeding hour, until the end of the day. In Table 1, we display the ED state until 16:00, then continued with the simulation-based forecast; the staffing level used in the simulation is the one exercised in our partner ED – we refer to it as "the existing staffing", and it appears in Table 2, under n(Current). Columns Ip, Sp, Op, and Nu list utilization levels of the respective staff (For nurses, this accounts for the time devoted to patient care, and excluding administrative duties; Physicians are exempted from the latter.). #Beds and #Chairs represent the number of occupied beds and chairs, respectively; %W is the fraction of patients that are exposed to unsatisfactory care, which here is taken to be "physician's first encounter occurs later than 30 minutes after arrival".

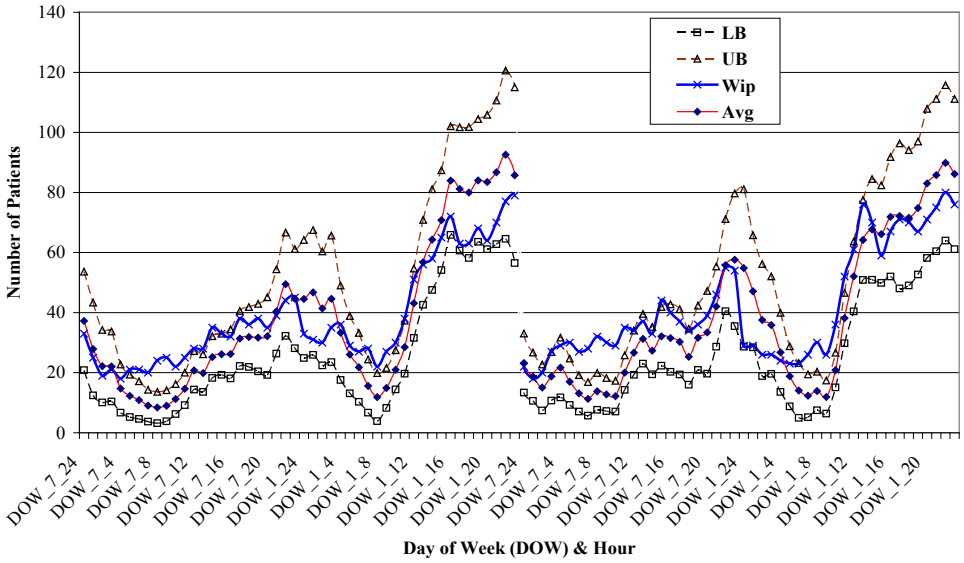

Figure 3: Comparing the Database with the simulated ED current-state (Weekdays and Weekends)

In Table 2, we display the staffing level in each category: first using ED existing staffing - n(Current), the offered load level (as explained in Section 4.3) - Offered Load; the recommended staffing level based on the offered load (aiming to achieve %W< 0.25) – n(OL); the RCCP level (as explained in Section 4.2) – RCCP Load; and its recommendation aiming at less than 90% staff utilization – n (RCCP).

Table 1: Simulation performance measures – current and forecasted (existing staffing)

Hour	Ip	Sp	Op	Nu	#Beds	#Chairs	%W
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%

Table 2: Staffing levels (present 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

5.3 Forecasting – performance measures

In Table 3, we record a simulated performance, under staffing levels calculated via the Offered Load 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 in excess of 1-2).

It is interesting to compare Table 3 (planned staffing) with Table 2 (existing staffing): the latter has obvious hours of under- and over-staffing while the formers' performance is rather stable. Preplanned staffing, either for resource utilization or, better yet, patients' service level (%W), clearly has its merit.

Table 3: Simulation performance measures (using OL and RCCP)

	Performance measures using OL recommendation							Performance measures using RCCP recommendation						
Hour	Ip	Sp	Op	N	Bed	Chair	%W	Ip	Sp	Op	N	Bed	Chair	%W
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%

6 INEDVANCE: A SUPPORT SYSTEM FOR RECORDING, PREDICTING AND DISPLAYING ED EVENTS

Input to our system originates from numerous data sources. For example, the ED's current state is based on information from a multitude of hospital IT systems, such as the Admit Discharge Transfer (ADT) system, the Picture Archiving and Communication System (PACS), the Lab Order Reservation system and the Electronic Medical Records system. Yet these systems provide only minimal *operational* information, such as the start and end of an activity. In particular, no information on queue lengths or waiting times is available (and here our simulation-based capabilities of ED state completion and prediction comes handy).

The hospital IT system collects its information and presents it to the user as a set of indicators and parameters. To interact with this hospital system, we have designed *InEDvance* (Greenshpan et al. 2009): a decision support system that can *record, process, simulate, and present* event data that hospital IT systems record and send, along with current (as in Section 5.1) and future performance measures (as in Sections 5.2 and 5.3). The InEDvance system comprises algorithms that assist the ED manager in planning resource allocation for the next several hours for handling forecasted resource scarcity. In particular, InEDvance has, at its core, a simulation-based module that is fed (in real-time) data from the hospital IT systems and then, through simulation (as described above), identifies and presents patient flow bottlenecks (e.g. excessive lines at the X-Ray) and consequently alert ED management.

The information arriving from the various IT systems generates a *dashboard* of past, present and predicted activities within the ED. We sample-demonstrate the use of such a dashboard by combining it with our ED simulator, and graphically presenting (potentially in real-time) information on the dashboard, using a graphical user interface. Figure 4 below shows a snapshot of the dashboard that presents, in various ways, past, current, and future occupancy of the different ED rooms. Figure 5 demonstrates a dashboard that could alert, based on calculated forecasting indicators, against predicted congestion and resource shortage.

7 CONCLUSIONS AND FUTURE RESEARCH

The wide scope of our project opens up ample opportunities for future research directions. We now list some that are related to simulation. First is *Validation* of our methods, continuing the limited pilot experiments in Section 5, and expanding to compare them against actual ED measurements. The simulation tool should also be refined, for example to account for patients who leave without being seen (LWBS), or ambulance diversions (see Green 2008) – both phenomena reduce effective ED workload. Simulation accuracy also calls for a better understanding (note the varying levels of accuracy in Figure 3). Related to that is the need for improved calibration with analytical models ((4) and (3), refinements or alternatives) that generate

staffing schedule. Here one could incorporate into the simulation also optimization and staffing constraint capabilities - indeed, ED staff availability is severely limited, as it is restricted by hospital needs beyond the ED as well as HR laws.

An attractive goal is to use the simulation to automatically trigger managerial interventions, present (for load balancing) and future (adaptive staffing), say based on present (predicted) workload that is matched against present (planned) resources: this takes running the simulation on a rolling-horizon basis (e.g. every hour, each run 8 hours into the future), which requires research support. With our partner hospital, we are analyzing the feasibility of a pilot RFID or Ultrasound implementation, integrated with the simulator and used for tracking the location of patients and hopefully also staff. Then, dashboard design, and integration with simulation, call for expertise from Human Factor Engineers, who are also collaborating with us towards an eventual InEDvance deployment. And finally there is the issue of real-time implementation: for the present research, we used Arena, which takes *minutes* per short-term replication; we also have an implementation with AnyLogic, which takes only *seconds* per replication; and each has its (dis)advantages (and costs), which must all be taken into account.

Figure 4: Dashboard snapshot showing rooms occupancy

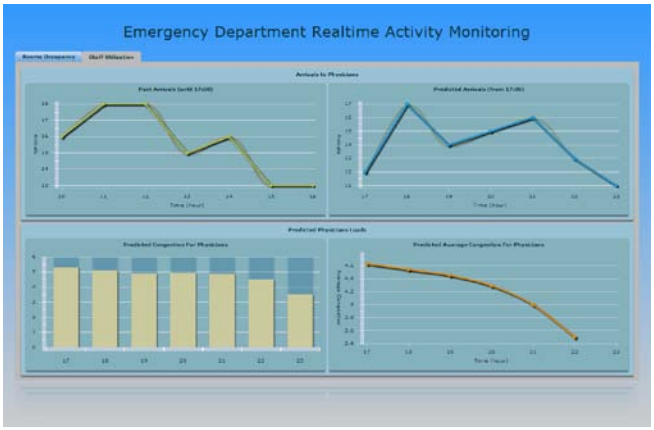

Figure 5: Predicted arrivals and physicians load

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic simulation*. New York: Springer.

Badri, M. A., and J. Hollingsworth. 1993. A simulation model for scheduling in the emergency room. *International Journal of Operations & Production Management* 13:13–24.

Ballard, S. M., and M. E. Kuhl. 2006. The use of simulation to determine maximum capacity in the surgical suite operating room. In *Proceedings of the 2006 Winter Simulation Conference*, eds. L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 433–438. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Beaulieu, H., J. A. Ferland, B. Gendron, and P. Michelon. 2000. A mathematical programming approach for scheduling physicians in the emergency room. *Health Care Management Science* 3: 193–200.

Biller, B., and B. L. Nelson. 2002. Answers to the top ten input modeling questions. In *Proceedings of the 2002 Winter Simulation Conference*, eds. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 35–40. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Borst, S., A. Mandelbaum, and M. Reiman. 2004. Dimensioning large call centers. *Operations Research* 52(1):17–34.

Channouf, N., P. L'Ecuyer, A. Ingolfsson, and A. N. Avramidis. 2007. The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science* 10:25–45.

Derlet, R. W., and J. R. Richards. 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*, eds. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson, 1057–1064. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Feldman, Z., A. Mandelbaum, W. Massey, and W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science* 54:324–338.

Fujimoto, R., D. Lunceford, E. Page, and A. M. Uhrmacher. 2002. Grand challenges for modeling and simulation. Technical Report No. 350, Schloss Dagstuhl.

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

Green, L. V., P. J. Kolesar, and J. Soares. 2001. Improving the SIPP approach for staffing service systems that have cyclic demand. *Operations Research* 49:549–564.

Green, L. V., P. J. Kolesar, and W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* 16:13–39.

Greenshpan, O., Y. N. Marmor, S. Wasserkrug, B. Carmeli, P. Vortman, F. Basis, D. Schwartz, and A. Mandelbaum. 2009. InEDvance: advanced IT in support of emergency department management. In *The 7th Conference on Next Generation Information Technologies and Systems*. Springer.

Gunal, M. M., and M. Pidd. 2008. DGHPSim: Supporting smart thinking to improve hospital performance. In *Proceedings of the 2008 Winter Simulation Conference*, eds. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler, 1484–1489. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Guo, M., M. Wagner, and C. West. 2004. Outpatient clinic scheduling - a simulation approach. In *Proceedings of the 2004 Winter Simulation Conference*, eds. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 1981–1987. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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., W. Cai, S. J. Turner, W. J. Hsu, S. Zhou, M. Y. H. Low, R. Fujimoto, and R. Ayani. 2006. A generic symbiotic simulation framework. In *Proceedings of the 20th Workshop on Principles of Advanced and Distributed Simulation*. ed. S. Ceballos, 131. Washington, DC: IEEE Computer Society.

Jacobson, S. H., S. Hall, and S. R. Swisher. 2006. Discrete-event simulation of health care systems. In *Patient Flow: Reducing Delay in Healthcare Deliver*, ed. R. W. Hall, 211–252. Springer US.

Janssen, A. J. E. M, J. S. H. Van Leeuwaarden, and B. Zwart. 2008. Refining square root safety by expanding Erlang C. Technical Report (http://www.win.tue.nl/~jleeuwaa/paper20.pdf).

Jun, J. B., S. H. Jacobson, and J. R. Swisher. 1999. Application of discrete-event simulation in health care clinics: a survey. *Journal of the Operational Research Society*, 50:109–123.

Khurma, N., G. M. Bacioiu, and Z. J. Pasek. 2008. Simulation-based verification of lean improvement for emergency room process. In *Proceedings of the 2008 Winter Simulation Conference*, eds. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler, 1490–1499. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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* 18:391–397.

Kolb, E. M. W., J. Peck, S. Schoening, and T. Lee. 2008. Reducing emergency department overcrowding - five patient buffer concepts in comparison. In *Proceedings of the 2008 Winter Simulation Conference*, eds. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler, 1516–1525. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kuljis, J., R. J. Paul, and L. K. Stergioulas. 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*, eds. S. G. Henderson, B. Biller, M. H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 1449–1453. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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

Maman, S., A. Mandelbaum, and S. Zeltyn. 2009. Uncertainty in the demand for service: the case of call centers and emergency departments. Research in progress.

McNeil, A., R. Frey, and P. Embrecht. 2005. *Quantitative Risk Management*. Princeton University Press.

Medeiros, D. J., E. Swenson, and C. DeFlitch. 2008. Improving patient flow in a hospital emergency department. In *Proceedings of the 2008 Winter Simulation Conference*, eds. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler, 1526–1531. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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

Sinreich, D., and O. Jabali. 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.

Spry, C.W., and M. A. Lawley. 2005. Evaluating hospital pharmacy staffing and work scheduling using simulation. In *Proceedings of the 2007 Winter Simulation Conference*, eds. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 2256–2263. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Takakuwa, S., and A. Wijewickrama. 2008. Optimizing staffing schedule in light of patient satisfaction for the whole outpatient hospital ward. In *Proceedings of the 2008 Winter Simulation Conference*, eds. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler, 1500–1508. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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*, eds. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 04–07. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Vollmann, T. E., W. L. Berry, and D. C. Whybark. 1993. *Integrated Production and Inventory Management*. Homewood, Ill: Business One Irwin.

Wong, C., G. Geiger, Y. D. Derman, C. R. Busby, and M. W. Carter. 2003. Redesigning the medication ordering, dispensing, and administration process in an acute care academic health sciences centre, In Simulation Conference, 2003. Proceedings of the 2003 Winter, vol.2, 1894–1902, 7–10 Dec. 2003.

AUTHOR BIOGRAPHIES

YARIV N. MARMOR is a doctoral student at the Technion – Israel Institute of Technology. He received his M.Sc. in Industrial Engineering from the Technion in 2003. His research activities involve implementing industrial engineering methods in the Health Care Industry. His email is <myariv.il@gmail.com>.

SEGEV WASSERKRUG is the manager of the Business Optimization group in IBM's Haifa Research lab. He has a B.A. and M.Sc. in Computer Science and a Ph.D. in Information Systems from the Technion. He has strong academic and practical background in a variety of areas, including simulation, optimization, Bayesian Networks and software engineering. He has practical experience, including researching and applying a variety of techniques. His email is <segevw@il.ibm.com>.

SERGEY ZELTYN is a researcher at IBM Haifa Research Labs, focusing on applications of operations research and statistical methods to business optimization in service systems. He received his M.Sc. and Ph.D. in statistics from the Technion. His research interests include queueing theory and applied statistics. His email is <sergeyz@il.ibm.com>.

YOSSI MESIKA is a research staff member at IBM Research Labs in Haifa. He received his B.Sc. in Computer Science from the Technion. He is currently an MBA student at the Open University, Israel. He is taking an active part in projects within the Healthcare & Life Sciences group, with the focus on Medical Imaging, Interoperability, and Epidemiological modeling. His email is <mesika@il.ibm.com>.

OHAD GREENSHPAN is a researcher at IBM Haifa Research Labs, focusing on advanced technologies for the Healthcare and Life Sciences domain. He received his M.Sc. in Computer Science from Ben-Gurion University. He is doing his PhD in Tel-Aviv University, focusing on advanced data management and integration on the web. His email is <ohadg@il.ibm.com>.

BOAZ CARMELI is a research staff member and the manager of the IT for Healthcare & Life Science group at the IBM Haifa Research Lab. He holds a B.Sc. in computer science. He is involved in the definition, design, and implementation of IT solutions for Healthcare and Life Sciences projects. He made contributions to fields ranging from wireless networks for hand held devices, business processes and integration systems for the healthcare market. His email is <boazc@il.ibm.com>.

AVRAHAM SHTUB is the Stephen and Sharon Seiden Chair in Project Management at the faculty of Industrial Engineering and Management, Technion, Israel. His current research focuses on the development of advanced simulation tools for training. He is the recipient of the 2008 Project Management Institute Professional Development Product of the Year Award for the training simulator "Project Team Builder – PTB". His email is <shtub@ie.technion.ac.il >.

AVISHAI MANDELBAUM is the Benjamin & Florence Free professor, at the faculty of Industrial Engineering and Management, Technion, Israel. He has been doing research in the area of stochastic processes, from the perspectives of operations research, statistics, engineering, and management. His research has been applied mainly to service operations, with a focus on telephone call centers and emergency departments. His email is <avim@tx.technion.ac.il>.