On Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective. An Extended Version

Mor Armony
NYU, marmony@stern.nyu.edu

Shlomo Israelit
Rambam Health Care Campus (RHCC), s_israelit@rambam.health.gov.il

Avishai Mandelbaum
Technion—Israel Institute of Technology, avim@ie.technion.ac.il

Yariv N. Marmor
ORT Braude, marmor.yariv@mayo.edu

Yulia Tseytlin
IBM, yuliatse@gmail.com

Galit B. Yom-Tov
Technion—Israel Institute of Technology, gality@tx.technion.ac.il

Patient flow in hospitals can be naturally modeled as a queueing network, where patients are the customers, and medical staff, beds and equipment are the servers. But are there special features of such a network that sets it apart from prevalent models of queueing networks? To address this question, we use Exploratory Data Analysis (EDA) to study detailed patient flow data from a large Israeli hospital.

EDA reveals interesting and significant phenomena, which are not readily explained by available queueing models, and which raise questions such as: What queueing model best describes the distribution of the number of patients in the Emergency Department (ED); and how do such models accommodate existing throughput degradation during peak congestion? What time resolutions and operational regimes are relevant for modeling patient length of stay in the Internal Wards (IWs)? While routing patients from the ED to the IWs, how to control delays in concert with fair workload allocation among the wards? Which leads one to ask how to measure this workload: Is it proportional to bed occupancy levels? How is it related to patient turnover rates?

Our research addresses such questions and explores their operational and scientific significance. Moreover, the above questions mostly address medical units unilaterally, but EDA underscores the need for and benefit from a comparative-integrative view: for example, comparing IWs to the Maternity and Oncology wards, or relating ED bottlenecks to IW physician protocols. All this gives rise to additional questions that offer opportunities for further research, in Queueing Theory, its applications and beyond. A shorter, more focused version of the paper appears in Armony et al. (2015).

Key words: Queueing Models, Queueing Networks, Healthcare, Patient flow, EDA

Contents

1. Introduction

Health care systems in general, and hospitals in particular, constitute a very important part of the service sector. Over the years, hospitals have become increasingly successful in deploying medical and technical innovations to deliver more effective clinical treatments. However, they are still (too) often rife with inefficiencies and delays, thus presenting a propitious ground for research in numerous scientific fields, specifically Queueing Theory.

Of particular interest to queueing scientists is the topic of *patient flow* in hospitals: improving it can have a significant impact on quality of care as well as on patient satisfaction. Indeed, the medical community has acknowledged the importance of patient flow management, as is illustrated by e.g. Standard LD.3.10.10 which the Joint Commission on Accreditation of Hospital Organizations (JCAHO (2004)) set for patient flow leadership. In parallel, patient flow has caught the attention of researchers in Operations Research, Applied Probability, Service Engineering and Operations Management; with Queueing Theory being a common central thread connecting these four disciplines. The reason is that hospitals experience frequent congestion which results in significant delays. Hence they fit naturally the framework of Queueing Theory, which addresses the tradeoffs between (operational) service quality vs. resources efficiency.

The analysis of patient flow raises ample theoretical challenges, and an almost prerequisite for the relevance of such analysis is real data. The present study builds on ample such data, and it seeks to promote a data-based queueing-science perspective of patient flow. Our hope is to stimulate relevant research, with the ultimate goal being delay reduction with its accompanying important benefits: clinical, financial, psychological and societal.

Our starting point is that a queueing network encapsulates the operational dimensions of patient flow in hospitals, with the medical units being the nodes of the network, patients are the customers, while beds, medical staff and medical equipment are the servers. But what are the special features of this queueing network in terms of its system primitives, key performance measures and available controls? To address this question, as indicated, we study an extensive data set of patient flow, through the lenses of a queueing scientist. Our study highlights some interesting phenomena that arise in the data, which leads to discussing: 1) their operational implications, 2) how these phenomena impact the queueing model, 3) how queueing theory might be used to explain these phenomena, and 4) what are some resulting significant and promising research opportunities?

1.1. EDA, the scientific paradigm and queueing science

Our approach of learning from data is in the spirit of Tukey's Exploratory Data Analysis (EDA) approach (Tukey (1977)). To quote from Brillinger (2002), John Wilder Tukey "...recognized two types of data analysis: exploratory data analysis (EDA) and confirmatory data analysis (CDA). In

the former the data are sacred while in the latter the model is sacred. In EDA the principal aim is to see what the data is "saying". It is used to look for unexpected patterns in data. In CDA one is trying to disconfirm a previously identified indication, hopefully doing this on fresh data. It is used to decide whether data confirm hypotheses the study was designed to test".

Using the EDA-CDA dichotomy, here we focus on EDA. This prepares the ground for future CDA which, in the present context, would be the application of data-based (queueing) models to confirm or refute prevalent hypotheses. Confirmation would contribute insight that supports the management of the originating system(s) (patient flow in hospitals, in this paper); refutation gives rise to new hypotheses, further EDA then CDA, with ultimately new insightful models. This EDA-CDA cycle is the day-to-day routine in natural sciences, where it is commonly referred to as The Scientific Paradigm (The OCR Project (IBM(2011)). Human complexity forced the paradigm into Transportation Science (Herman (1992)) and Behavioral Economics (Camerer et al. (2003)), and the present study aims at the same for the analysis of patient flow in hospitals. A similar approach has already proved successful in other settings, including semi-conductor manufacturing (Chen et al. (1988)), telecommunication (Leland et al. (1994)), new product development (Adler et al. (1995)) and, most recently, call centers (see Mandelbaum et al. (2000) and Brown et al. (2005) for the empirical findings, and Gans et al. (2003) and Aksin et al. (2007b) for surveys on follow-up work).

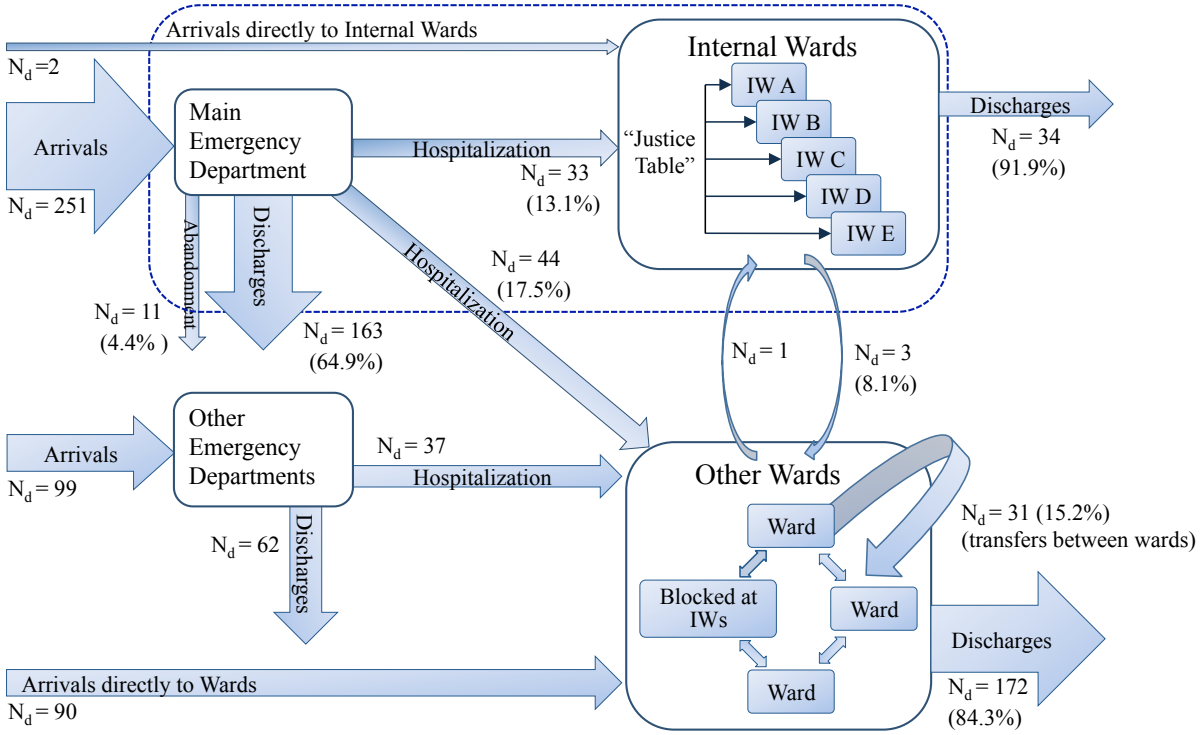
We conclude our discussion of EDA with two "apologies" to the Statistician. First, the goals of the present study, its target audience and space considerations render secondary the role of "rigorous" statistical analysis (e.g. hypothesis testing, confidence intervals). Indeed, we believe that its intensional omission is both necessary and justified. Accordingly, we either mention a statistical rationalization only in passing (eg. for the fact that the arrival process to the ED is over-dispersed Poisson, and that IW length-of-stay, in days, is Log-Normal), or we content ourselves with a convincing visual evidence (the privilege of having a large data set). Secondly, our data originates from a single Israeli hospital. This raises doubts as far as the scientific (universal) relevance of our findings is concerned, and rightly so. The "apology" here draws first from a previous study of *eight* Israeli EDs (Marmor (2003)), which already gave rise to findings similar to ours. This still leaves the concern that Israeli hospitals are perhaps too "unique", but they are not: for example, a parallel study by Shi et al. (2014), in a major Singapore hospital, reveals many phenomena that are shared by both hospitals. Most importantly, our hope is that plainly reading the manuscript will dispel all doubts concerning its relevance and significance (practical or statistical).

1.2. Rambam hospital

The data we rely on was collected at a large Israeli hospital.This hospital consists of about 1000 beds and 45 medical units, with about 75,000 patients hospitalized annually. The data includes

detailed information on patient flow throughout the hospital, over a period of several years (2004-2008). In particular, the data allows one to follow the paths of individual patients throughout their stay at the hospital, including admission, discharge, and transfers between hospital units.

Traditionally, hospital studies have focused on individual units, in isolation from the rest of the hospital; but this approach ignores interactions among units. On the flip side, looking at the hospital as a whole is complex and may lead to a lack of focus. Instead, and although our data encompasses the entire hospital, we chose to focus on a sub-network that consists of the Emergency Department (ED) and five Internal Wards (IW), denoted by A through E; see Figure 1. This sub-network,



N_d - daily average number of patients per weekdays (excluding holidays) total over 105 days, for period January 1, 2007 - May, 31, 2007

main ED - Internal, Surgery, Traumatology, and Orthopedic EDs

Figure 1 Patient Flow in Rambam—Zooming in on the ED+IW network

referred to as ED+IW, is more amenable to analysis than studying the entire hospital. At the same time, it is truly a system of networked units, which requires an *integrative* approach for its study. Moreover, the ED+IW network is also not too small: approximately 47% of the patients entering the hospital stay within this subnetwork, and 16% of those are hospitalized in the IWs; and the network is fairly isolated in the sense that its interactions with the rest of the hospital are minimal. To wit, virtually all arrivals into the ED are from outside the hospital, and 91.6% of the patient

transfers into the IWs are from within the ED+IW network. (These transfers include arrivals from outside the hospital and exclude transfers to the IWs from other hospital units.) Those numbers were obtained from the data by examining a detailed 90×90 transition matrix between hospital wards, as is illustrated by Figure 2. While focusing on the ED+IW network, we nevertheless rip the benefits of having access to overall hospital data. One such benefit is the use of other hospital units (e.g. Oncology, Maternity) as reference points. This improves one's understanding of specific phenomena that arise from the ED+IW data.

The ED+IW network. The ED has 40 beds and it treats on average 245 patients daily. An internal patient, whom an ED physician decides to hospitalize, is directed to one of the five Internal wards. The IWs have about 170 beds that accommodate around 1000 patients per month. Internal Wards are responsible for the treatment of a wide range of internal conditions, thus providing inpatient medical care to thousands of patients each year. Wards A-D share more or less the same medical capabilities - each can treat similar (multiple) types of patients. Ward E, on the other hand, attends to only the less severe cases; in particular, this ward cannot admit ventilated patients.

1.3. Some hints to the literature

Patient flow in hospitals has been studied extensively. Readers are referred to the many papers in Hall (2006), which are also sources for further references. In the present section, we merely touch on three dimensions, which are the most relevant for our study: a network view, queueing models and data-based analysis. Plenty additional references, on particular issues that arise throughout the paper, will be further cited as we go along.

Most research on patient flow has concentrated on the ED and how to improve its flows in within. There are a few exceptions that offer a broader view. For example, Cooper et al. (2001) identifies a main source of ED congestion to be *controlled* variability, downstream from the ED (eg. operating-room schedules that are customized to physician needs rather then being operationally optimized). In the same spirit, de Bruin et al. (2007) observes that "refused admissions at the First Cardiac Aid are primarily caused by unavailability of beds downstream the care chain." These blocked admissions can be controlled via proper bed allocation along the care chain of Cardiac in-patients; and to support such allocations, a queueing network model was proposed, with parameters that were estimated from hospital data. Broadening the view further, Hall et al. (2006) develops data-based descriptions of hospital flows, starting at the highest unit-level (yearly view) down to specific sub-wards (eg. imaging). The resulting flow charts are supplemented with descriptions of various factors that cause delays in hospitals, and then some means that hospitals employ to cope with these delays.

There has been a growing body of research that tackles operational problems in hospitals with Operations Research (OR) techniques. Brandeau et al. (2004) is a handbook of OR methods and

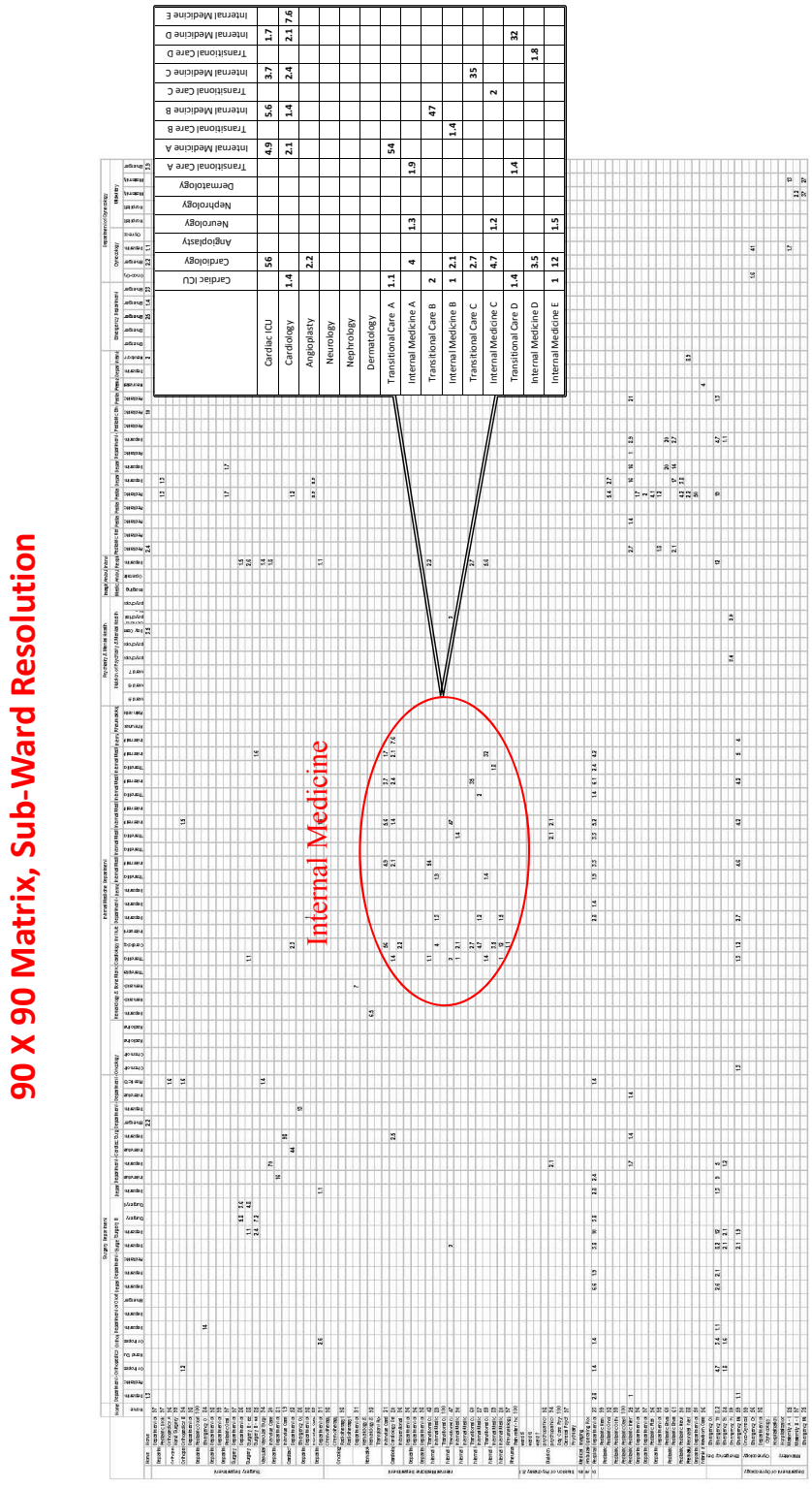


Figure 2 **Transition probabilities between hospital wards, at the resolution of sub-wards. For example, during the period over which the matrix was calculated (January 4th, 2005 to June 31st, 2005), 47% of the patients in the Transitional Care Unit of IW A were transferred to IW A itself. plausibly after their condition improved enough for the transfer.**

applications in health care; the part that is most relevant to this paper is its section on Health Care Operations Management, which is recommended also for additional references. More recently, Green (2008) surveys the potential of OR in helping reduce hospital delays, with an emphasis on queueing models. Jennings and de Véricourt (2011, 2008) and Green and Yankovic (2011) apply queueing models to determine the number of nurses needed in a medical ward. Green (2004) and de Bruin et al. (2009) rely on queueing models such as Erlang-C and loss systems, to recommend bed allocation strategies for hospital wards. Lastly, Green et al. (2007a) and Yom-Tov and Mandelbaum (2014) develop (time-varying) queueing networks to help determine the number of physicians and nurses required in an ED.

There is also a growing acknowledgement of the significant role that data can, and more often than not must, play in patient flow research. For example, Kc and Terwiesch (2009) use econometric methods to investigate the influence of workload on service time and readmission probability, in Intensive Care Units (ICUs). This inspired Chan et al. (2014) to model an ICU as a state-dependent queueing network, in order to gain insight on how speedup and readmission effects influence the ICU. Finally, Mandelbaum et al. (2012) identified heterogeneity of LOS across IWs. This gave rise to fair policies of ED-to-IW routing (see Section 5); fair in the sense that they balance workload across the wards while controlling for patient delays.

1.4. Data description

The data-set we obtained from Rambam hospital covers patient-level flow data, over the period of January 1, 2004, through December 1, 2008. There are four compatible data "tables" which capture the hospital operations. The first table (Visits) contains records of ED patients, including their ID, arrival and departure times, arrival mode (independently or by ambulance), cause of arrival, demographic data, and more. The second table (Justice Table) contains details of the patients that were transferred from the ED to the IWs. This includes information on the time of assignment from the ED to an IW, the identity of this IW, as well as assignment cancelations and reassignment times when relevant. The third table (Hospital Transfers) consists of patient-level records of arrivals to and departures from hospital wards. It also contains data on the responsible ward for each patient as, sometimes, due to lack of space, patients are not treated in their best-fit ward; hence, there could be a distinction between the physical location of a patient and the ward that is clinically in charge of that patient. The last table (Treatment) contains individual records of first treatment time in the IWs. Altogether, our data consists of over one million records, which enable the presently reported EDA and more.

*S*ome data challenges: Not surprisingly, the data obtained from the hospital was far from being amenable to analysis. Here, expertise gained at the Technion SEELab with call centers data, jointly

with our own significant efforts, turned out essential for an extensive data cleaning effort. To elaborate some, there were plenty of records that were flawed due to archiving or plainly system errors. These were identified via their inconsistency with trustable data and hence corrected or removed. But more challenging was the identification of records that had been included in the data due to some regulations, rather than physical transactions. For example, some unreasonable workload profiles lead to the discovery of a high fraction of transfers from the ED to a virtual ward, all occurring precisely at 11:59pm; subsequent analysis associated each of these transfers with a physical transfer, from the ED to some actual ward on the following day. The reason for the inclusion of such virtual transfers was financial, having to do with regulations of insurance reimbursement. And this is just one out of many examples.

All in all, our cleaning process demanded many months of work and, in parallel, it triggered two additional processes. The first was the creation of a corresponding data-repository, at the Technion SEELab. This repository, and others, are conveniently accessible via SEEStat, which is SEELab's environment for online EDA. Having SEEStat at our disposal has greatly facilitated our EDA. As an example, Figure 3 displays three snapshots from a SEEStat session. The far background is the working environment. The others are two outputs of a single data query, produced (simultaneously and almost instantaneously) as two Excel spreadsheets: a numerical one and its graphical animation. Here, we are reproducing Figure 19 (§4.1). The second process was the writing of a comprehensive manuscript, which served as the root and shaped a framework for the present study. This manuscript (Mandelbaum et al. (2011)), in the context of hospitals, plays the same role that Mandelbaum et al. (2000) played for call centers: simply yet importantly archiving routine operational performance, which paved the way for the more "exotic" phenomena that are reported in the present paper.

2. Summary of results

In Sections 3, 4 and 5 we analyze, respectively, the ED, IWs and the transfer of patients from the ED to the IWs. In each of these sections we highlight interesting phenomena that arise from the data and, subsequently, discuss their implications on system operations, queueing modeling, and research opportunities. Section 6 provides a broader view of some common themes that arise throughout the paper: patient-flow as a surrogate for overall hospital performance (operational, clinical, financial, societal and psychological), workload, capacity, fairness, time-scales and data-based research.

The present research has already provided the empirical foundations for several graduate theses, each culminating in a research paper: Marmor (2010) studied ED architectures and staffing (see Zeltyn et al. (2011)); Yom-Tov (2010) focused on time-varying models with customers' returns for

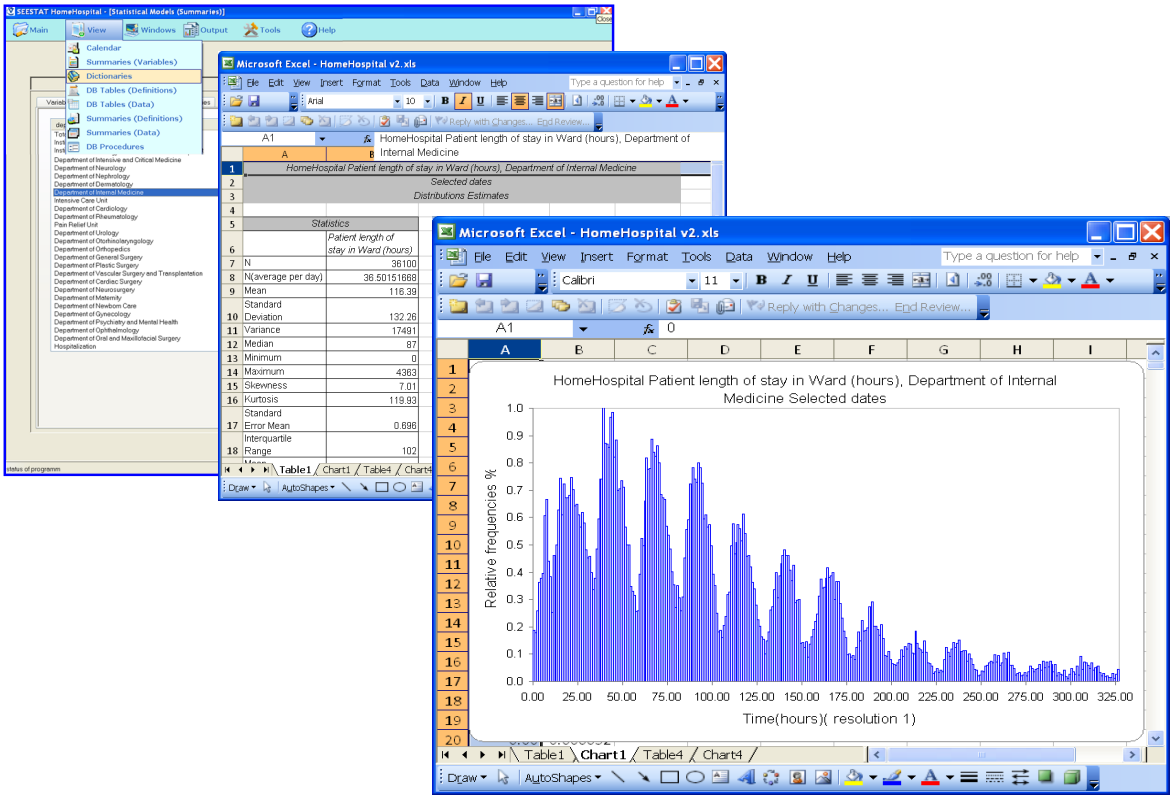


Figure 3 **SEEStat: EDA online environment, producing numerical output & its graphical animation (corresponding to Figure 19, §4.1)**

the ED (Yom-Tov and Mandelbaum (2014)) and the IWs; Tseytlin (2009) investigated the transfer process from the ED to the IWs (Mandelbaum et al. (2012)); and Maman (2009) explored over-dispersion characteristics of the arrival process into the ED (Maman et al. (2011)); these are all examples of the EDA-CDA process, alluded to in Subsection 1.1. We now proceed with a summary of our main findings.

2.1. ED analysis (Section 3)

In the analysis of the ED, we focus on the interrelations between workload distribution, time of day, and patient length of stay (LOS). To this end, we seek a simple queueing model that fits the complex ED reality, being encouraged by the success of Erlang models in the context of call centers (Brown et al. (2005), Aksin et al. (2007b)). Indeed, the dynamics of a call center can be usefully captured by dividing the day into successive time-intervals, say 24 hours, and then fitting independent stationary Birth & Death models to each hour. This approach, however, turns out to not work for EDs. The reason, broadly speaking, is that the effects of an arrival to the ED ripples over this arrival's LOS (several hours forward), and the ED state typically changes significantly during that LOS. (In call centers, on the other hand, the scale of LOS is minutes and the effects

of an arrival are hence local.) Nevertheless, we observe that a simple $M/M/\infty$ model is useful in describing the steady-state of ED patient count during times when the ED is congested.

This Birth & Death view of the ED yields a simple "black-box" ED model which is useful but nevertheless restricted: it accurately captures operational implications related to steady-state ED occupancy and LOS, but it is blind to the intricacies of patient flow. The model is thus descriptive as opposed to explanatory. It acknowledges the ED as a delay-node within a broader queueing network (analytical or simulation model). But getting to the source of time- and state-dependency requires the finer resolutions of explanatory models. We hint at several such models that could support physicians and nurses staffing decisions (Yom-Tov and Mandelbaum (2014)), the planning of physician scheduling priorities (Huang et al. (2015)) and the design of ED operational architectures (Marmor et al. (2012)).

2.2. IW analysis (Section 4)

In studying the IWs we find that, with respect to the LOS distribution, two time resolutions are relevant: days and hours. These two resolutions result in strikingly different histograms: the daily picture is unexplainably log-normal and the hourly one is periodically peaked (Figure 19). LOS as measured in days is mostly impacted by medical considerations, and is operationally associated with decisions of capacity and staffing levels. LOS in resolution of hours is related to operational decisions such as physician schedule, discharge policy, and time of transfer from the ED. These latter factors could potentially be manipulated relatively easily.

Further analysis of the four IWs reveals that their LOS distributions share the same daily and hourly shape, as described above. However, their average LOS (ALOS) differ, and the smallest Ward B is significantly the "fastest" (shortest ALOS). This asymmetry challenges the fair distribution of work among the wards, which we return to momentarily. We also put things in further perspective by analyzing LOS in two parallel *Maternity* wards. Here already the *shapes* of LOS histograms differ, which reflects their different patient mix: one ward is in charge of pre-birth complications and the second of post-birth; normal (vaginal) births are to be allocated between the wards so that workload is ultimately balanced. Again, challenges of fair work-allocation arise, but this goes beyond the present study as it involves both operational and emotional workloads (see §6.3 and §6.5).

Rate of return to hospitalization can also be inferred from our data. In the case of IWs, a patient's return within a relatively short time could possibly signal poor quality of care, and in fact is quite rare. In contrast, returns to other types of wards could be predictably routine: for example, returns to *Oncology* are scheduled to be part of the normal treatment regiment. We argue in Section 4.2.2 that, as far as resource dimensioning is concerned, these two return types (IWs vs. Oncology) call for differing considerations in formulating their supporting queueing models.

Finally, in the context of the IWs, we investigate forms of economies and dis-economies of scale with respect to ward size (number of beds). Of interest is the effect of scale on LOS, the probability of having to wait for a bed, and the waiting time until being first seen by a physician.

2.3. ED-to-IW flow (Section 5)

With respect to the ED-to-IWs transfer process, we focus on the routing mechanism and its relationship to transfer delays and fairness. We consider fairness towards the medical staff in the IWs as well as fairness towards patients. We argue that understanding how routing influences all of these factors is key to choosing the "right" routing mechanism.

Closely related to routing are incentive issues; in order to reduce delays in the ED-to-IWs transfer, one must prioritize work such that patient discharge and preparation for a new admission will have a minimal impact on these delays. But for this to happen, the wards must have the incentive to comply. In particular, if a "fast" ward is consistently assigned more patients, it may have an incentive to become less efficient. Our observations also reveal that the ownership of the time-of-transfer decision (i.e. when to actually transfer the patient) has a large impact on the delay prior to the transfer. Specifically, when the ED "pushes" patients to the IWs (vs. IWs "pull" patients from the ED), delays become significantly shorter.

2.4. System view

The above treats the three network components (ED, IWs, transfers) unilaterally. However, our study underscores the importance of looking at this network as a whole, as these three components are clearly interdependent. Indeed, delays in the ED-to-IW transfer are tightly coupled with IW workloads and are affected by operational decisions such as the IW discharge policy. At the same time, delays in transfer to the IWs increase the workload within the ED which, in turn, results in delayed emergency treatment as well as prolonged ED LOS. We now let the data tell part of this story. (Further details will be provided throughout the paper.)

Denote by L the process that describes the overall number of patients within the ED. Viewing L as a Birth-and-Death process, it is thus characterized by birth-rates (arrivals) and death- or service-rates (departures). Figure 4 shows that the service rate per ED-patient is *decreasing* in L, for $L \geq 35$. In Section 3 we provide a few explanations for this decrease, which pertain to limited ED resources and human psychology factors. We now present a more subtle plausible explanation, which is revealed only by examining the integrated ED+IW network.

According to Figure 5, the times of day during which L is relatively high are 12pm to 9pm. Hence, to understand the behavior of the departure rate per patient, for large L, one should examine the discharge process from the ED during these times.

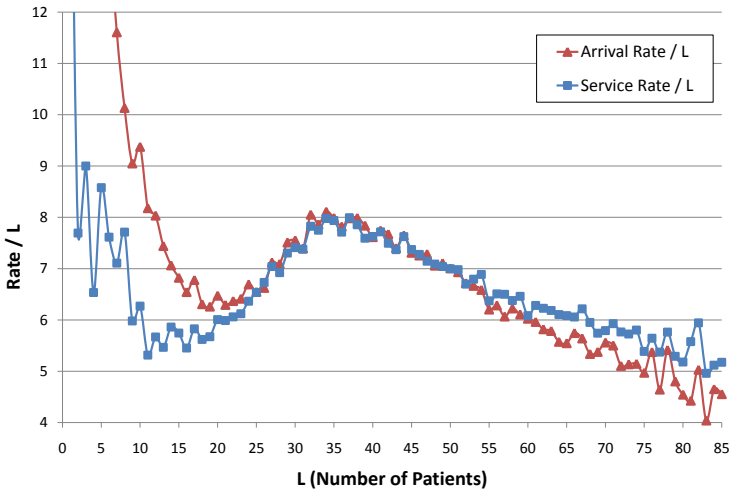


Figure 4 **Arrival rate and service rate as a function of L**

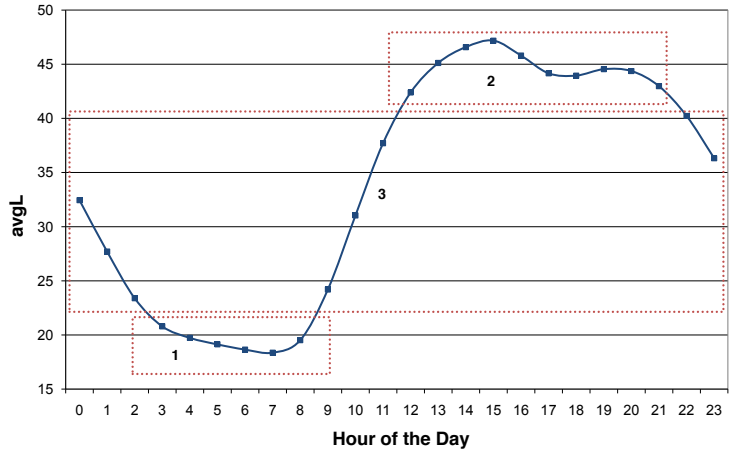


Figure 5 **Average number of patients in ED (avgL) per hour of the day**

Approximately 20% of the ED patients are internal patients, who end up being transferred and admitted to one of the IWs. These transfer patients remain in the ED until the respective IW is prepared to admit them. Switching attention to the IWs, Figure 6 shows that departure rates from the IWs peak between 3pm and 6pm, and are almost null otherwise. (This narrow time-window for discharge is an outcome of the IW routine where physicians perform their rounds during late mornings, and only thereafter can patients be made ready for release; Shi et al. (2014), in his analysis of a Singapore hospital, discovers a similar peak at 1pm.) One would thus expect that there would be a decline in the number of transfer patients following the IWs discharge peak. Indeed, we observe, in Figure 7, that there is a decline in the number of transfer patients starting

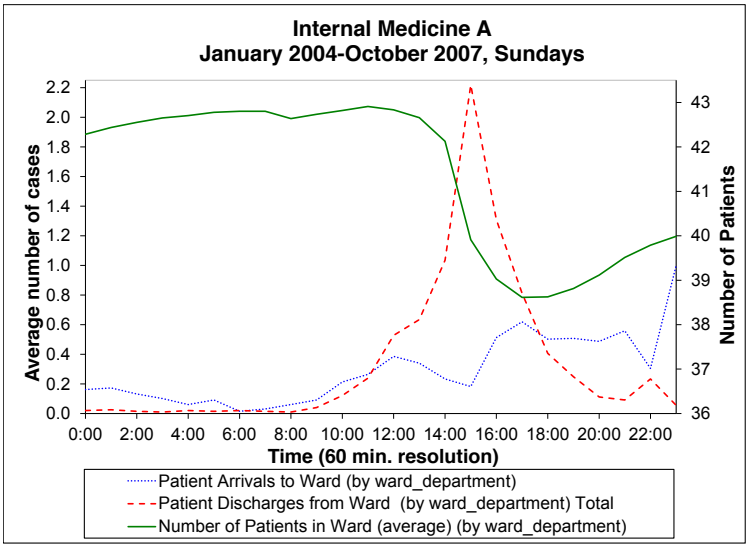


Figure 6 **Arrivals, departures, and average number of patients in Internal wards by hour of day**

at around 5pm (the decline is not as sharp as one might expect due to the time lag that is involved in preparing an IW for the arrival of any particular transfer patient).[1]

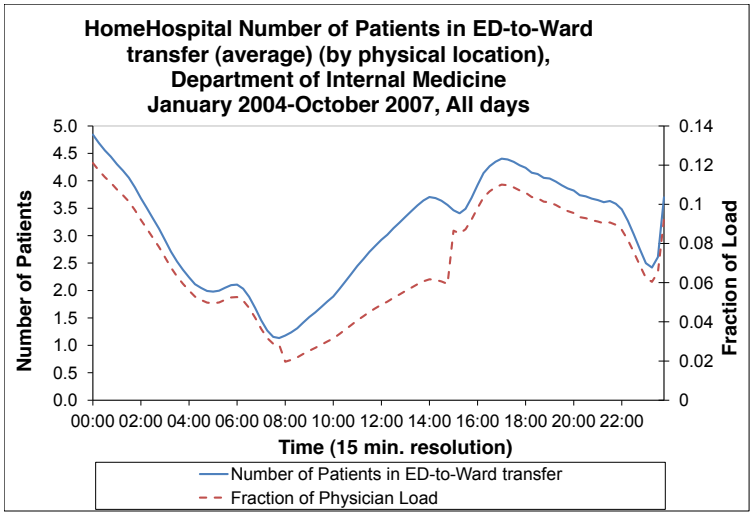


Figure 7 **Number of patients in ED-to-IW transfer and the fraction of ED physicians' time devoted to these patients**

Why do then service rates per patient decrease in $L \geq 35$? Our analysis shows that precisely at those times when the ED has high occupancy, the average number of transfer patients is also relatively high. The transfer patients, who are present in the ED during these times, require care

[1] The local plateaus before 3pm and 11pm and the sharp increases thereafter are due to shift changes in the ED at those times.

from the ED medical staff while, instead, they should be obtaining care in the IWs. Figure 7 also shows our estimates of the fraction of ED physicians time that is spent caring for the transfer patients, assuming every such patient requires 1.5 minutes of physician's time every 15 minutes. This extra workload for the ED staff, that occurs at times when their workload is already high, results in "wasted" capacity and *throughput degradation*: a phenomenon that is well acknowledged in transportation (Chen et al. (2001)) and telecommunication (Gerla and Kleinrock (1980)).

3. Emergency Department

Many research papers have been devoted to the topic of patient flow in the ED (Hall (2006)). In this paper, we focus mostly (though not exclusively) on modeling the ED as a node in a queueing network (the hospital), rather than on patient flow within the ED per-se. The latter involves a multitude of interrelated steps, as is illustrated in Figure 8. Our main objective is to identify a *simple* model that would usefully capture the distributions of ED occupancy and Length-of-Stay (LOS). Such a process could be used, for example, as a black-box model of the ED-node within a larger queueing network, which represents patient flow in the entire hospital.

A similar effort was devoted in the past to model a call center as a queueing system. A successful approach for call centers is to divide the day into 15-30 minutes intervals, and assume that the system has a fixed arrival rate, service rate and number of servers during each interval. Moreover, convergence to steady-state is relatively fast, and therefore the assumption that the system works in steady-state within each interval has been widely accepted. This approach works well for call centers because changes in system parameters occur at a longer time scale than an individual customer LOS. In the ED, things are different. As we shall see, it is typical for patients to spend several hours in the ED. Hence, the effect of each individual arrival is not local, but lingers for that length of time. Moreover, during a patient's stay in the ED, many operational factors may change such as the arrival rate, staffing level, equipment availability, etc. Therefore, it is necessary for a queueing model to capture the system dynamics throughout the day, rather than decomposing the day into shorter time intervals and assuming steady-state during these short intervals.

The so-called black box approach is simple and useful in supporting certain operational decisions that depend only on total patient count but not on internal dynamics, or they can model ED sojourn times within a larger hospital model. However, this approach is also lacking, as it is too crude in supporting other more refined operational decisions. To address this we discuss in Section 3.3 a hierarchy of models that we have proposed to support decisions such as staffing, task scheduling and overall design of ED architecture.

The Emergency Department (ED) of Rambam hospital attends to about 250 patients daily, with about 60% classified as Internal patients and 40% classified as Surgical or Orthopedic. The ED

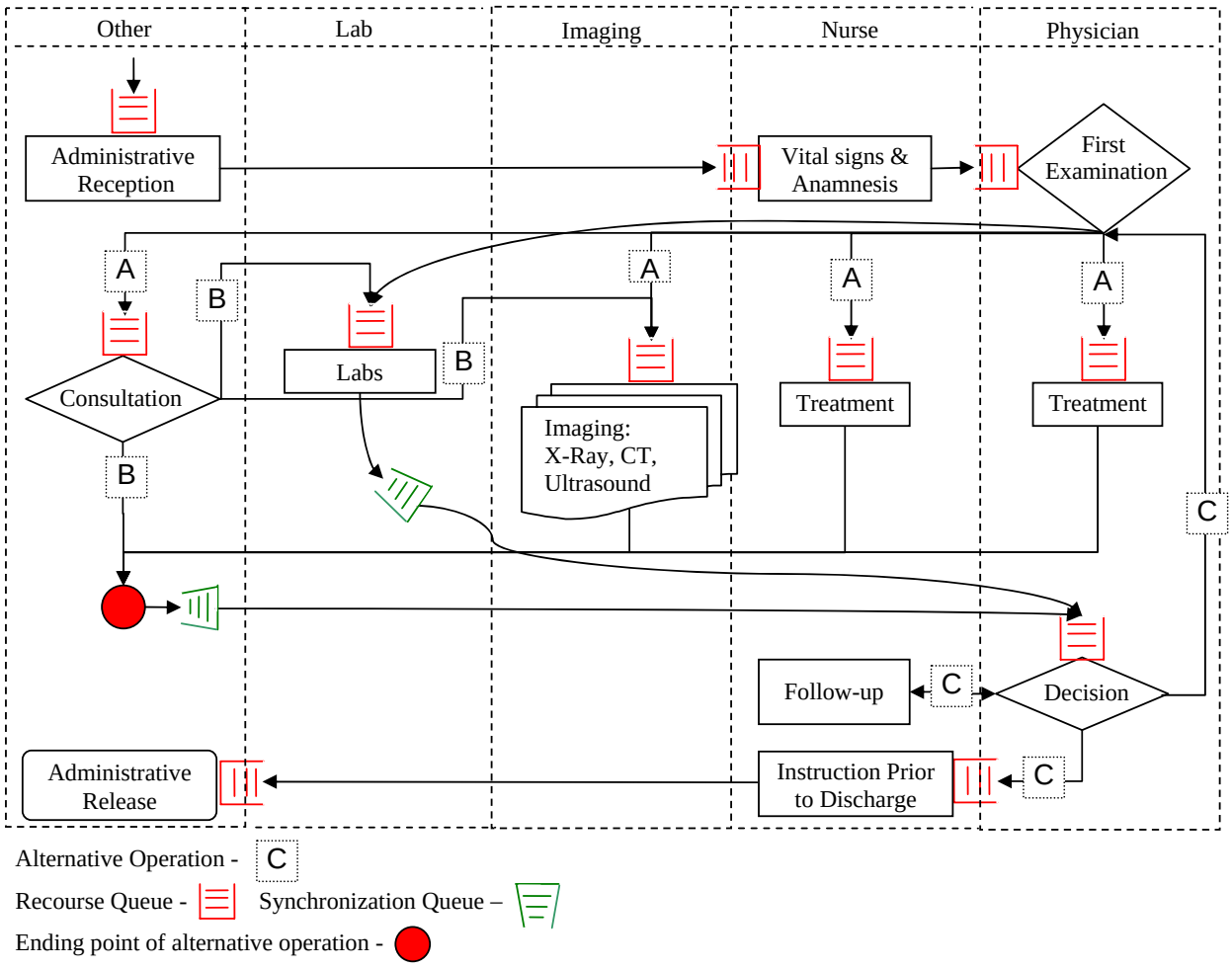


Figure 8 **Activities-Resources flow chart in the ED**

has three major areas: Internal acute, Trauma acute, and Walking patients area. There are other ED locations, physically detached from the main one we are focusing on, which are specialized to other areas such as Pediatrics or Ophthalmology. Mean sojourn time of Internal patients in the ED (ALOS) equals 4.25 hours, with a large variability over individual patients (17% of the Internal patients spend over 6 hours in the ED).

3.1. Empirical findings

In this section, we explore relationships between ED LOS, occupancy level, and arrival- and departure-rates. We regard patients arrival time as the time of their admission into the ED. The departure time is the time the patient has been discharged from the ED, either to go home or to be transferred to another hospital unit.

We start by studying the distribution of L - the number of patients in the ED[2]: Figure 9 shows its empirical distribution. We observe that the distribution has an unusual shape; it is skewed to the left, has a light right tail and has two local peaks. Further investigation reveals that the distribution of L, for each hour of the day, follows a Normal distribution with mean and variance that vary over time (this was confirmed by the appropriate statistical tests). Specifically, Figure 10 shows the distribution of $L(t)$, the number of patients in the ED at the time of the day t, for $t = 0, 1, ..., 23$.[3] The mean of L by hour of the day, $avgL$, was shown in Figure 5 (§ 2). Using these

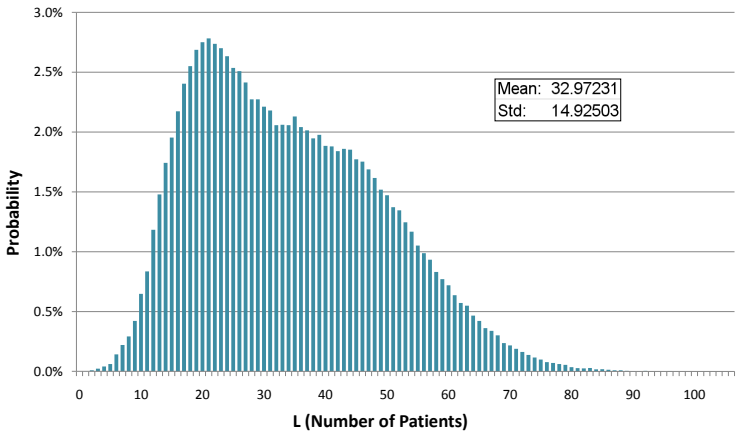


Figure 9 Distribution of the number of patients in the ED

figures, we identify three main patterns that compose the $L(t)$ distribution: (1) From 02:00 until 09:00, where the average number of patients ($avgL$) is about 20 and varies from 0 to 40; (2) From 12:00 until 22:00, where $avgL$ is about 45 and it varies from 0 to 90; and (3) The rest of the day (09:00-12:00, 22:00-02:00), when the distribution shifts from one group to the other.

Aligning $avgL$ and λ (the arrival rate of patients to the ED) together in Figure 11, reveals a time-lag between arrivals and load, which is common in many service systems[4]. As an aside, notice that the arrival rate may vary significantly over a period of several hours, which is the typical LOS of a patient in the ED (recall that average LOS is 4.5 hours). This is precisely the reason why the ED occupancy process may not be described as a piecewise stationary process.

[2] Note that L is not the number of occupied beds: L includes *walking* patients whose medical condition allows them to move independently (e.g. between physician's room, nurses station, lab, etc.). Note also that L includes patients prior to their first visit to a physician. Finally, there is no rigid constraint on the number of beds in the ED, namely, beds are added in accordance to congestion levels.

[3] The evolution of the distribution by hour resembles the findings of Edie (1954) with respect to the distribution of arrivals in transportation systems.

[4] The time-lag between arrival rate and occupancy has significant influence on staffing, as was discussed in Green et al. (2007a) and Feldman et al. (2008).

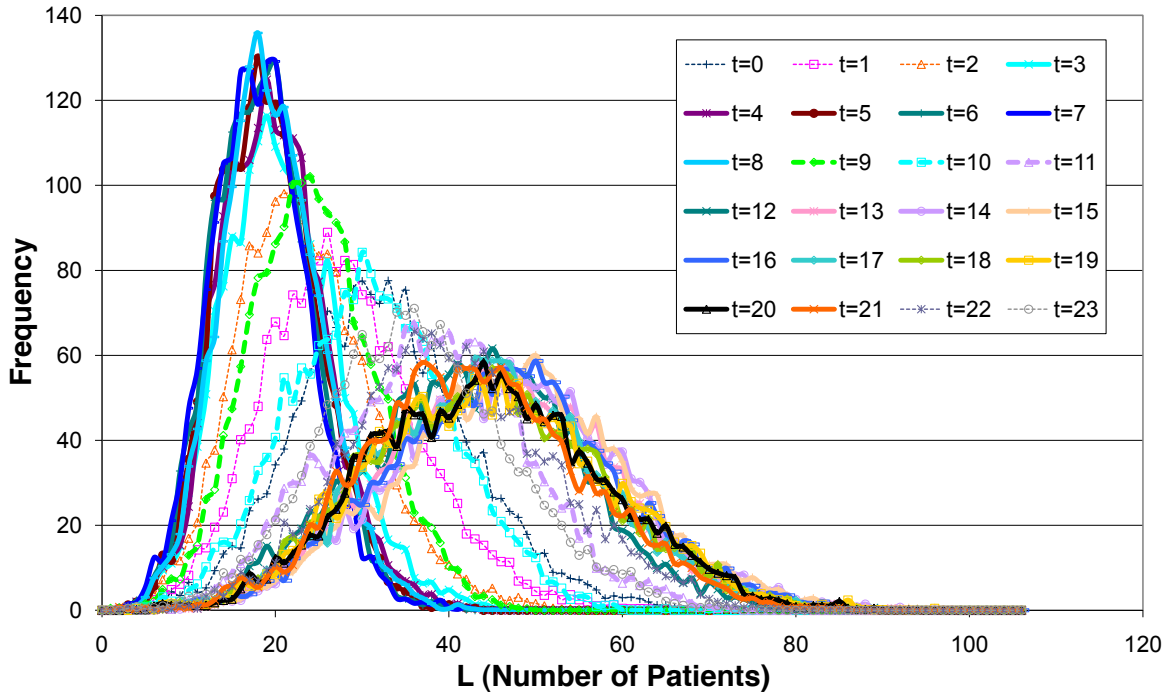


Figure 10 **Distribution of the number of patients in the ED per hour of the day** $(L(t))$

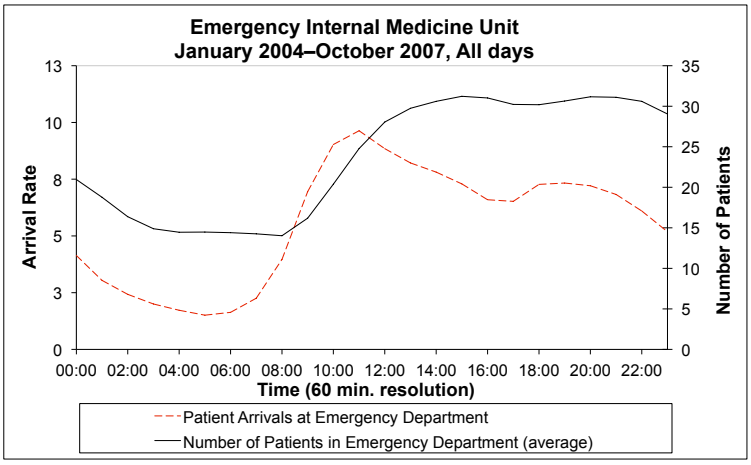


Figure 11 **Average number of patients and arrival rate by hour of the day**

The time-varying evolution of $L(t)$ has prompted researchers to explore the question of how to stabilize ED performance measures such as waiting times, either via adaptive staffing (Green et al. (2006), Yom-Tov and Mandelbaum (2014), Zeltyn et al. (2011)), or via admission control, that is, controlling the arrival rate, for example, by applying ambulance diversion policies (Hagtvedt et al. (2009)). Most of this work applied simple queueing models, such as M/M/N or M/M/∞, to capture system dynamics. In all of them, there is an underlying assumption that the arrival and service rates do not depend on the time of the day or the system state (occupancy) or time of day.

But given our observations that such complex features do appear in the ED dynamics, it is natural to ask whether simple models are still useful.

When viewing the system occupancy level as a Birth-and-Death process, and estimating the corresponding birth and death rates, we notice that the resulting rates heavily depend on the occupancy level L. Figure 4 (§ 2) shows our estimates of the total arrival and departure rates of each state L, divided by the ED occupancy L; these rates do not seem to fit any simple prevalent queueing model. We can clearly see that the arrival rate changes with the system state, and similarly the service rate. We observe that both $\lambda(L)/L$ and $\mu(L)/L$ decrease for $7 \leq L < 15$, increase for $15 \leq L < 35$, and decrease again for $35 \leq L \leq 60$. As L itself changes over time, we in fact observe a process that is both time- and state-dependent. The following questions naturally arise:

1. Which has a more significant effect on arrival and service rates: time dependency or state dependency?

2. What is, or is there, a simple way to accurately model the underlying process that determines ED occupancy?

3. How does one explain state-and-time dependency in the ED of Rambam hospital?

The first question is open for further rigorous research; We conjecture that the state-dependent arrival rates can be largely attributed to the time varying arrival patterns. Confirming this requires the development of statistical methods to distinguish between the two effects. We will address the second question at length in Section 3.2, and now provide several explanations for the third.

There are a few plausible explanations as to why the service rates are state- (and time-) dependent:

a. The first explanation involves the fact that the ED is actually a *multi-type server system*. That is, servers could be beds, but also medical staff or equipment. Therefore, it is natural that at times when the medical staff and equipment are stretched to satisfy service requirements, the total service rate per patient will be lower. This can explain the decrease in service rate per patient observed for $35 \leq L \leq 60$.

b. The second explanation involves *psychological effects* of load on service rate. It is known that pressure affects efficiency of service providers (Sullivan and Baghat (1992)). Accordingly, we conjecture that the medical staff speed up service when congestion starts to build ($15 \leq L < 35$), but feel overwhelmed by the pressure of the system, and therefore slow down, when the number of patients is high ($35 \leq L \leq 60$).

c. Another explanation is related to the *combination* of time-varying arrivals and multi-types of patients, who differ with respect to their LOS distribution. Recently, Marmor et al. (2011) observed a similar phenomenon in a Cardio Vascular ICU system. In that ICU, arrival rates vary across the days of a week. Their findings suggest that when two or more types of patients are analyzed

together, the proportion of patients with the shorter LOS ("fast") out of the total ICU patients, is amplified with the *increase* of arrival rate. This results in both the growth in occupancy level, and the increase in service rate per bed. Of course, when the arrival rate decreases, we expect to see a decrease both in occupancy level and in the proportion of the fast patients; Thus, the decline in the service rate per bed. When the arrival rate is stabilized or just slightly reduced, but not enough to alleviate congestion, there will be a drop in the proportion of fast patients. Hence, under this conditions, the system will experience both an increase of occupancy level, and decrease in service rate per bed. The ED of Rambam hospital also has multi-types of patients. For example, "Walking" patients have significantly shorter LOS than to-be-hospitalized patients. Hence, this explanation is relevant in the current environment too.

d. Our last explanation involves the notion of *wasted capacity* and *throughput degradation*, that we expanded on in Section 2. In a nutshell, the ED is not an isolated ward, but is part of a network of wards within the hospital. When patients wait to be transferred to other wards, they still need to be checked upon by a physician or a nurse every 15 minutes or so. Moreover, during such extended transfer waits, clinical conditions can deteriorate and emergencies arise. Hence, if there are many patients waiting to be transferred, the physicians and nurses might be overly occupied in caring for these patients, that less attention will be given to new and in-process patients. That, in turn, can cause a decrease in the service rate per patient, as we observe when $35 \leq L$. We call this phenomenon "choking", to underline the fact that the system is "choked" by to-be-transferred patients, and the capacity of its servers is therefore "wasted". This phenomenon is well known in other environments such as transportation Chen et al. (2001) and telecommunications (Gerla and Kleinrock (1980)) where it is also referred to as throughput degradation.

3.2. What simple queueing model best fits the ED environment?

In addressing various operational problems in the ED, one could benefit from having a simple way to model this very complex environment (recall the process depicted in Figure 8, to be elaborated on in Section 3.3). In addition, a simple ED model can be useful if one studies the ED in conjunction with other units at the hospital. Fitting an accurate simple queueing model for this system can be quite challenging; Internal ED processes involve many steps that require different resources and priorities, depending on the patient condition. Here we present a successful attempt to find such a model with a good fit to our data.

Current literature postulates that a time-varying Erlang-C (Green et al. (2006)) or Erlang-B (de Bruin et al. (2009)) can capture the dynamics of the ED occupancy process. Our empirical findings above regarding the state dependency of the LOS suggest that these models may be too simplistic. Nevertheless, our observations suggest that if one focuses on the time of day where the

ED is heavily loaded, then a steady-state approximation based on an even simpler model, namely, $M/M/\infty$, fits the data quite well. We elaborate on the relevant exploration process below.

Recall the distribution of the ED occupancy as depicted in Figure 9. An attempt to find a statistical fit with a mix of known distributions results in a non-insightful mixture of up to seven distributions, including Normal, Gamma and Weibull. To obtain a potentially more insightful distributional fit, we focus our attention on the occupancy distribution of the internal ED only. The relevant empirical histograms for the daily occupancy, as well as occupancy by hour are depicted in Figures 12 and 13.

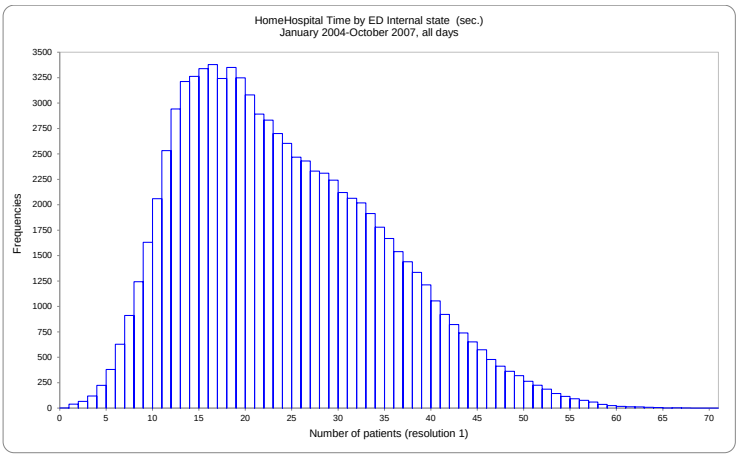


Figure 12 **Empirical histogram of Internal ED Occupancy**

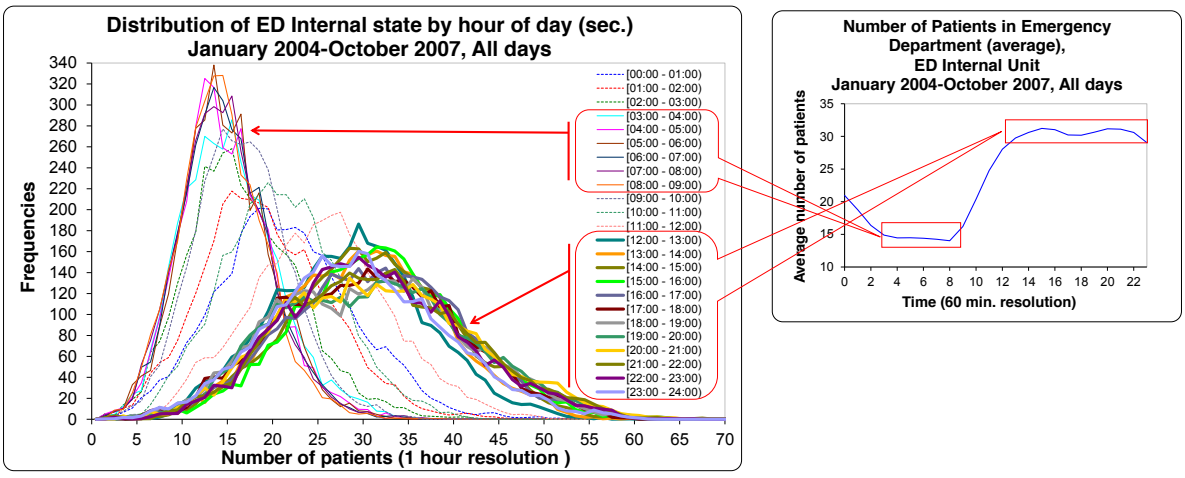


Figure 13 **Internal ED Occupancy histogram by hour of the day**

An attempt to find the best fit among all mixtures of three normal distributions has resulted in a distribution has a statistically significant fit to the corresponding empirical occupancy distribution

as illustrated in Figure 14. The figure also shows the empirical distribution of ED occupancy during

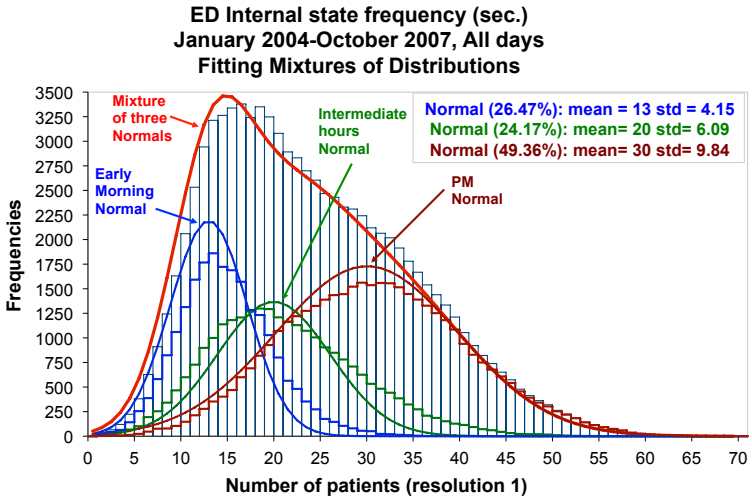


Figure 14 **A mixture of three normal distributions to the ED occupancy distribution**

the three periods identified in Figure 13; namely: early morning (3:00-8:59), PM hours (12:00-23:59) and intermediate hours (9:00-11:59 and 0:00-2:00). Remarkably, the three normal distributions capture the empirical distributions of the occupancy during these time intervals exceptionally well.

Focusing further on the PM hours (the peak hours in the ED), and narrowing down the data to the non-holiday Mondays during the year 2005 (so that most of the a-priori explained variability is controlled for), we find that a Normal distribution with mean 33.22 patients and a variance of 33.18 exhibits a statistically significant fit to the data (see Figure 15). By the central limit theorem,

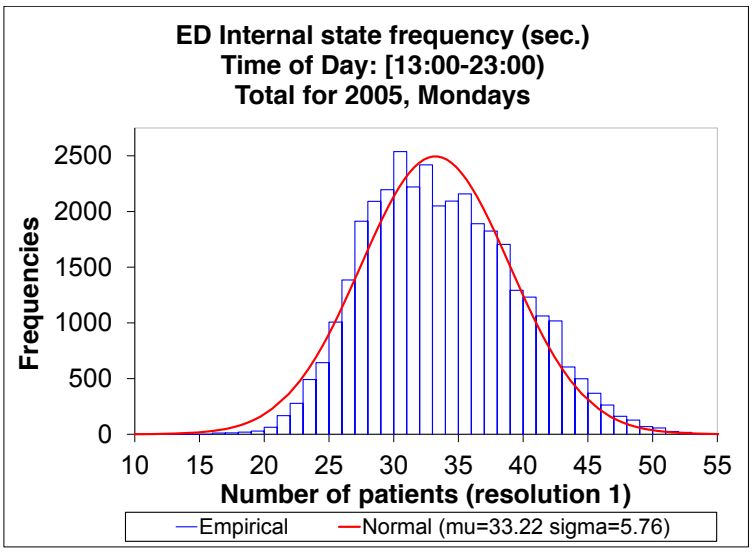


Figure 15 **A Normal fit to the occupancy distribution with equal mean and variance**

and because the mean is equal to the variance, this distribution is a good approximation for a Poisson distribution with rate ≈ 33.2. The latter is the same distribution as the steady-state of an $M/M/\infty$ queue with offered load $\lambda/\mu = 33.2$. In summary, we have observed that when controlling for variability associated with factors such as time-of-day, day-of-week, calendar year, and patient type, the resulting empirical distribution is well approximated by the steady-state distribution of an $M/M/\infty$ queue. This is a useful observation for decisions that directly rely on the occupancy distribution. In addition, an $M/M/\infty$ queue has the same distribution of an $M/M/N+M$ queue, where the abandonment rate is equal to the service rate. This is useful in situations where one is explicitly interested in abandonment induced performance measure such as the fraction of patients who have left without being seen (LWBS) and / or those who have left against medical advice (LAMA).

In sum, we have observed that if one focuses on occupancy data during peak times, and data is taken from an a-priori homogeneous set of observation, it is plausible that a simple model would emerge, as we have seen in this example. Moreover, while the ED environment is clearly a time varying one, our observations suggest that when focusing on peak hours, a steady-state approximation may be appropriate.

3.3. Opening the Black Box: A Hierarchy of Models

A model at the level of a birth and death process could suffice for decisions that treat the ED as a node in a larger queue network (hospital). However, one may need more detailed models to support more involved decisions that may alter the ED internal dynamics. In this section, we outline a few such queueing models and the operational decisions that are associated with them. In effect, we present a hierarchy of models, where one gradually adds more detail (possibly at the cost of simplicity and tractability) to support more involved operational decisions. All models are part of research projects that were motivated by our data.

We start with the simplest. In Yom-Tov and Mandelbaum (2014), the authors propose a so-called time-varying Erlang-R model (where "R" stands for Reentrant customers, Repetitive service, or Return to service). According to this model, which is depicted in Figure 16, customers oscillate between Needy and Content states. Whenever customers become needy, they join a FCFS queue to wait for service until a server becomes available. This model captures the essence of the ED reality in the sense that patients are attended by medical staff only when in need, and the rest of the time they occupy a bed or equipment. Yom-Tov and Mandelbaum (2014) shows that this model captures enough of the complex ED reality, to render it useful for generating staffing recommendations of physicians and nurses. It is important to note that, for Erlang-R to be a useful ED model, it *must* acknowledge time-variability. This is due to arrival rates that change significantly over an LOS, as describes previously. This is further elaborated on in Yom-Tov and Mandelbaum (2014).

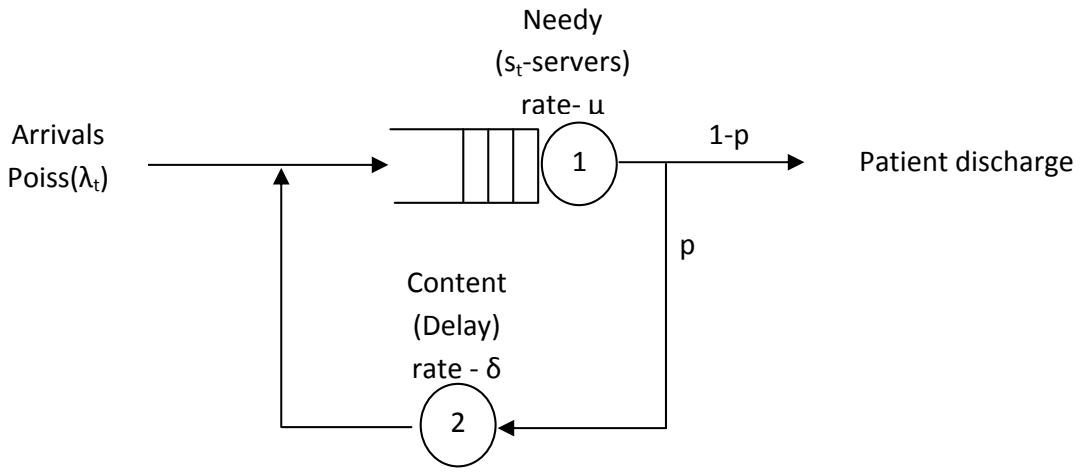


Figure 16 **The time varying Erlang-R queueing model**

At the next level of detail, a *multiclass* queueing model with deadlines and feedback is proposed by Huang et al. (2015) (see Figure 17), to help prioritize the work of ED physicians. The goal is to balance between Triage and In-Process (IP) patients which is formalized by minimizing congestion costs subject to triage deadline constraints. The authors propose a simple threshold-based priority

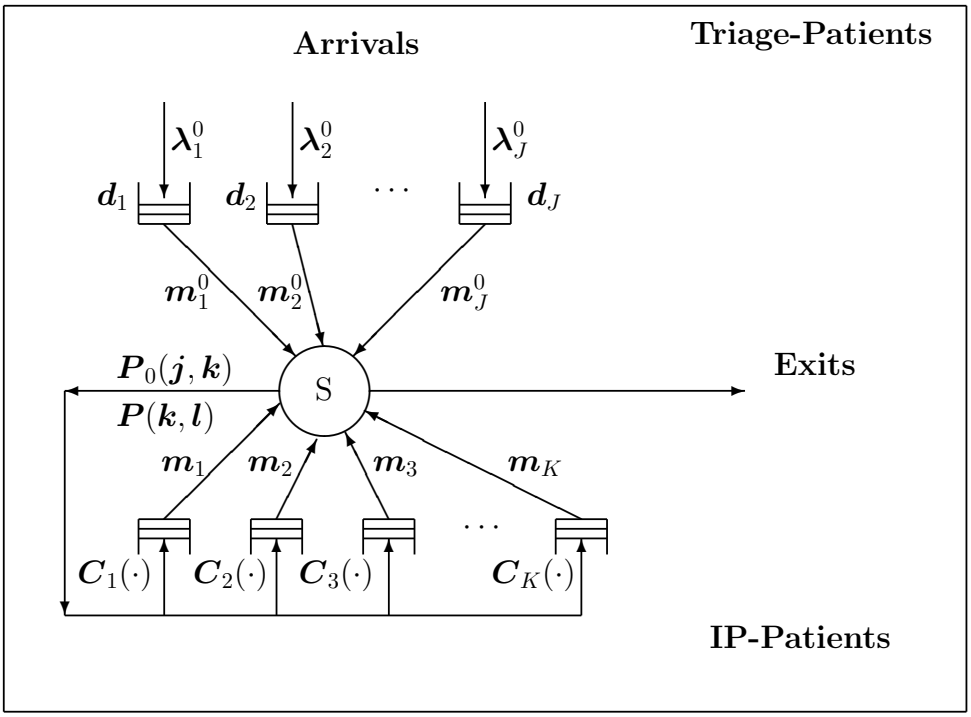


Figure 17 **An ED modeled as a multiclass queueing network with feedback and priorities**

rule combined with a modified $Gc\mu$ rule, and prove its asymptotic optimality in conventional heavy traffic.

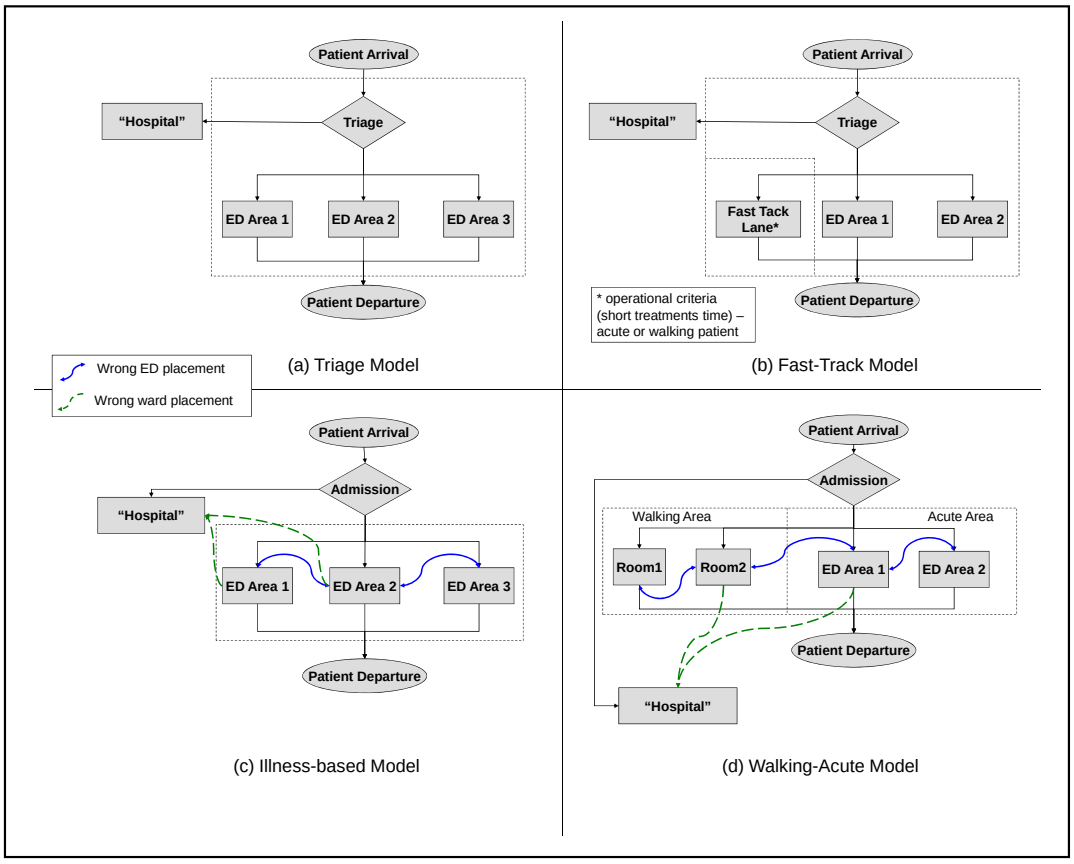


Figure 18 Emergency Department design of prevalent operational models

An additional level of modeling complexity has been proposed by Marmor et al. (2012) to support decisions of how to *design* the patient flow in Emergency Departments. The authors consider several ED operational architectures (see Figure 18), which were identified based on ED patient flow data from eight Israeli hospitals of various sizes and operational modes. Finally, Data Envelopment Analysis (DEA) is used to ascertain which operational model has dominant performance under various scenarios.

Marmor et al. (2012) shows that different operational models have weaknesses and strengths over various uncontrollable parameters. For example, hospitals that admit a high volume of elderly patients would preferably provide a Fast Track to cater to patients that are expected to have short ED LOS. This is a model where priorities are given based on *operational* criteria, while other EDs can use Triage rules, where priorities are given based on *clinical* criteria. When Triage and Fast Track are not feasible options (e.g., lack of space or staff), using a different track for Acute and

Walking patients (WA) is the most effective operational model (especially when the number of elderly arrivals to the ED is relatively high).

Further opportunities in queueing research: While our discussion in the current section has focused on examining whether a simple steady-state model can capture the occupancy dynamics of the ED during peak hours, we have also observed that time- dependency is extremely prevalent, especially in the context of arrival patterns (e.g. recall Figure 11). Thus, while in some cases a steady-state model may suffice, we argue that time variability must be included in queueing models to render them useful in supporting a wide array of decisions in the ED. Time-varying queues, i.e. where there are high-levels of predictable time variability, are notoriously hard to analyze but, in the ED environment one cannot ignore them, specifically because time variability is significant during an individual LOS. This also raises the importance of fluid-models, as being proxy-models for this predictable variability. Hence, ED-relevant standard queueing models require re-interpretations in time-varying environments.

Our discussion of ED process configurations (Figure 18) raises another interesting question, which is at the heart of emergency medicine, namely: which is the more appropriate model: an ED (Emergency Department) or an ER (Emergency Room)? In particular, should the "gate to the hospital" serve as just a router to its wards (ER model) or should it be a medical ward in itself (ED) where patients receive medical treatment beyond mere routing? There are many issues that arise or are intimately related to the "ED. vs. ER" question. To name just a few: what are the benefits vs. disadvantages of an "ED specialization"? (which, in some sense, is an oxymoron - after all, being an ED physician entails being a generalist that is trained in multi-disciplines); under what circumstances should one operate a Triage? and/or Fast-Track?; how are all the above questions related to the physical architecture, operational processes, human-factors and information design within the ED/ER? Data-based research is the key to addressing these important questions.

4. Internal Wards

As discussed in the Introduction, Internal Wards (IWs) are the "clinical heart" of a hospital. Yet, compared to EDs and Operating Rooms, IWs have received less attention in the Operations (Research, Management) literature. Some exceptions include papers on nurse staffing and bed allocation in medical wards (Jennings and de Véricourt (2011), Green (2004), Green and Yankovic (2011), and Yom-Tov (2010)). We find medical wards in general, and IWs in particular, to offer a rich environment for OR and OM research, and thus propose a deeper view into the world of IWs, and the research challenges that they give rise to. As we investigate the IW data from Rambam hospital, we identify many interesting phenomena. In Section 4.1, we observe non-standard LOS distributions; in Section 4.2 we extend our discussion to other medical wards and show how Maternity

wards differ in LOS distribution and that Oncology wards have unique return-to-hospitalization patterns. Lastly, in Section 4.3, we discuss the operational regimes of the Internal wards. Specifically, we ask whether multiple operational regimes can co-exist and explore the question of how does ward size affect its patients' LOS.

Various aspects of these phenomena can be attributed to the following characteristics of the IW system:

- Network design (topology): Offline Matching of patient types with admitting wards.
- Ward design: Bed capacity, staffing, patient mix.
- Operational policies: Discharge procedures, physician rounds, and more.

In each of the following subsections, we discuss how each phenomenon observed is related to the characteristics described above, and discuss related research challenges.

4.1. LOS distribution in Internal wards: Separating medical and operational influences

Rambam hospital has five Internal wards. Wards A–D are identical from a clinical perspective; the patients treated in these wards are of the same array of clinical conditions. Ward E is different in that it admits only less severe patients, for example, those who walked into the ED, as opposed to arriving via an ambulance. The size of the wards ranges from 20 to 45 beds. In this section, we examine the LOS distribution in each of those units. It is obvious that clinical conditions influence patient LOS. The influence of operational and managerial characteristics is, however, less obvious.

Figure 19 shows the LOS distribution in one of the IWs, in two time scales: days and hours. When considering daily resolutions, the Log-Normal distribution turns out to fit the data well. This is typical of other service systems as well, though the reason for its prevalence is unclear (Brown et al. (2005)). The second graph has a 1-hour resolution. We observe a completely different LOS distribution, with peaks that are periodically 24 hours apart. In this graph, the LOS distribution looks like a mixture of daily normal distributions.

These two graphs reveal the impact of two operational decisions: The daily time scale represents physician decisions, made every morning, on whether to discharge the patient that day or to extend hospitalization by at least one more day. The second decision is at what time will the patient be actually discharged during that day. This latter decision follows this discharge process: it starts with the physician who writes the discharge letters (after finishing the morning rounds); then nurses take care of paper work, instructing patients (and their families) on how to continue medical treatment after discharge, and arranging transportation (if needed). The discharge procedure is performed over "batches" of patients and, hence, takes a few hours. This results in a very low variance of the discharge time, as most patients are released between 3pm and 4pm (see Figure 6, §2). This explains the shape of the LOS distribution in the hourly resolution, and why we observe

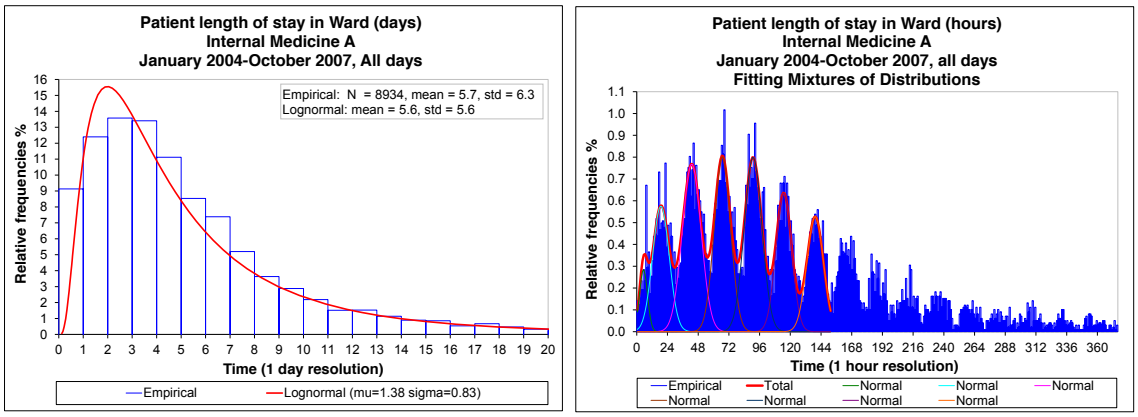


Figure 19 **LOS distribution of IW A in two time-scales: daily and hourly**

peaks at intervals of 24 hours. The variation around these peaks is determined by the arrival process: patients arrive almost exclusively over a 12-hour period (10am–10pm), with a peak in arrival rate between 3pm–7pm (Figure 6). We saw earlier, in Section 2, that the discharge policy has a significant influence on the workload in the ED.

Figure 20 displays the LOS distribution for the other four IWs in resolution of days. They all adhere to the Log-Normal distribution, though the parameters of that distribution are different. The implication of this difference, mainly the ALOS, will be further discussed in Section 5.5.

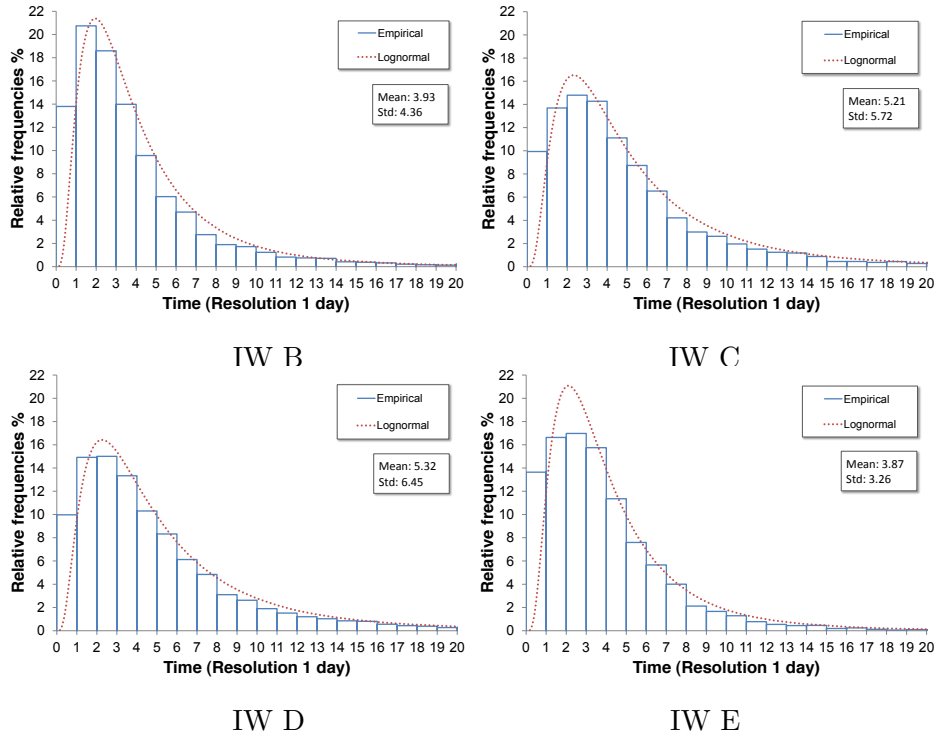


Figure 20 **LOS distribution of IW B-E**

Implications to queueing models: Attention must be given to both time scales: daily and hourly. Each time scale corresponds to a model with difference server entity, in the corresponding queueing system. Daily resolution is natural when considering beds to be the servers, as beds are occupied for days. For this case the challenge is to use a traditional time-varying queueing model with a Log-Normal service distribution to determine bed allocation. Hourly resolution is appropriate when considering nurses and physicians as servers, since service times are in the order of minutes. The offered load of nurses varies during the day for two reasons: (a) changes in the number of patients during the day (see Figure 6, § 2), and (b) patient admission and discharge require different attention than routine care. Combining both of these time variations, it is clear that the number of personnel must (and actually does) vary during the day; hence, the importance of observing the system in hourly resolution and understanding its behavior. As mentioned above, some efforts to develop queueing models for nurse staffing in IWs have been carried out by Jennings and de Véricourt (2011), Green and Yankovic (2011), and Yom-Tov (2010), but none of these works can explain the LOS distribution observed in our data.

There are additional interesting operational questions that are related to the above, for example: How will changes in the discharge process influence the system? Will balancing discharges more uniformly over the day be beneficial to the entire hospital? How would such a change influence delays of patients waiting to be transferred into the IW from the ED? (The connection between the ED and IWs will be discussed further in Section 5.4.) Queueing models are natural for addressing all these questions, and more.

4.2. Comparison between IWs and other medical wards

Naturally hospitals include medical wards other than IWs, which may have significantly different characteristics. To hint on some of these differences, we now examine two phenomena: the first is patient mix among IWs, which we compare against Maternity wards. The second is return to hospitalization, which we contrast with Oncology wards.

4.2.1. LOS in Maternity wards Different wards give rise to different LOS distribution. We use Maternity ward data from Rambam hospital as an illustration, and observe how mixing different patient types influences the LOS distribution. Figure 21 shows the LOS distribution of Maternity wards A and B in hourly resolution. We observe that, unlike the IWs, here the two wards have different LOS distributions, which are also different from the one observed earlier for the IWs. The difference stems from the mix of patients that are hospitalized in these wards.

The Maternity wards serve three types of patients: high risk pregnancy, women after normal (vaginal) deliveries, and women after Cesarean sections. Each type has different hospitalization requirements, and a different LOS distribution. A woman after normal delivery usually stays at the

hospital for two days, while a woman after a Cesarean section stays longer, usually 5 days. LOS of high-risk pregnancy patients has much higher variability; such a patient may be hospitalized anywhere between a few hours to a couple of months. There are also differences in medical treatments, and the equipment required. Due to these varying considerations, the assignment protocol to Maternity wards differ from the one used in the IWs (see Figure 22). Indeed, each ward caters to only one of the two more complicated conditions, while women after normal deliveries are routed so as to balance the workload between wards. Specifically, Maternity ward A treats high-risk pregnancies, while Maternity ward B takes care of women after Cesarean sections. Naturally, this affects the LOS distributions of these wards. We can clearly observe a 48-hour peak for both wards, and an additional peak around 120 hours (5 days) for ward B only. In IWs, we also have three types of patients: Ventilated, Special care, and Regular. But all of them are routed across all the four similar IWs, as seen in Figure 22 (for illustration purposes, Figure 22 only shows two such IWs), and discussed further in Section 5.

Implications to queueing modeling: The patient mix of Maternity wards raises the issue of fair allocation of work between wards: This kind of question has been addressed for cases in which service rates are server dependent (Armony and Ward (2010), Ward and Armony (2013), Mandelbaum et al. (2012)), but not for cases in which the service time is determined by the patient class (or by both patient class and the server). For more discussion on fairness see Section 5.5. Note that differences between patient classes typically go beyond the operational perspective, and could accommodate also emotional and cognitive considerations (see Section 6.3).

4.2.2. Return to hospitalization An important characteristic of a medical ward is its *rate of return* of patients to hospitalization. It is captured by the probability of return to hospitalization within, say, 3 months after release. We distinguish between two types of returns: unplanned and planned. Unplanned returns arise in wards where a patient is admitted for a one-time treatment, such as in an IW. In this case, if a patient returns shortly after a previous visit, the return is considered negative, as it is possibly due to a lack of appropriate treatment or a misdiagnosis. (This is certainly not always the case, therefore a better measurement would distinguish between patients who return with a similar condition and those who return with an unrelated problem. This information is not available in our database.) The return probability is then an operational surrogate for the medical quality-of-care. The hospital seeks to minimize such return occurrences. Planned returns, on the other hand, are recognized to be part of the treatment regiment and therefore do not reflect poor clinical care. For example, in Oncology ward where the treatment is executed in cycles, and patients are expected to visit the hospital repeatedly.

Table 1 compares the average number of returns per patient in IWs and Oncology, and the probability of return within three months, calculated over the studied period. We note that the two

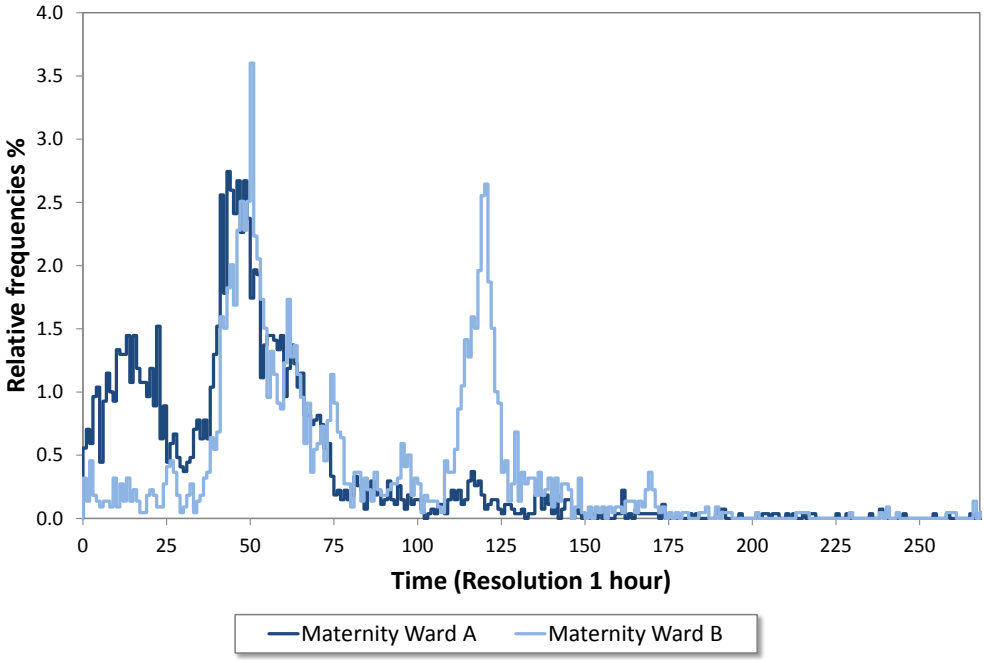


Figure 21 **Patient LOS in Maternity wards**

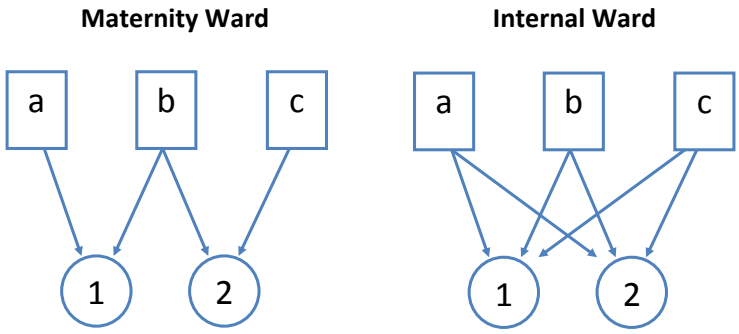


Figure 22 **Routing Schema of Maternity and Internal wards**

types of wards are indeed very different. We discuss momentarily how the above re-hospitalization phenomenon is significant in determining bed allocation.

Table 1 **Returns to hospital**

Ward	Average returns per patient	Average time between successive returns of a patient (days)	Probability of return within 3 months	ALOS of individual visit (days)
Internal	1.76	208	22%	4.8
Oncology	5.46	22	75%	3.4

Implications to queueing models: The distinction between unplanned and planned returns, and the data of Table 1, contribute to our understanding of the key factors in making bed allocation

decisions. The main question is: does one need to account for returns in making such allocation decisions? We acknowledge three important factors that determine the method needed: duration between returns, planned vs. unplanned returns, and the offered load ratio (which will be defined shortly).

Table 1 presents two different time scales for returns. In an IW, the time between returns is long enough to assume that each visit of a patient is independent of previous visits, and therefore traditional queueing models, such as Erlang-B, suffice (de Bruin et al. (2009), Green (2004)). This is not the case in the Oncology ward, where returns occur after a short time and an independence assumption may be harder to justify. In addition, the Oncology ward represents a planned return environment in which bed capacity must be reserved for patients who are in the midst of a series of treatments. In fact, these patients must typically have a higher priority over new admits, who can be transferred to another facility if needed. Hence, Oncology ward planning is closer to that of a medical clinic (see Green et al. (2007b)), in contrast to an IW.

Patient returns here call for a separation between hospitalized patients and those who are currently on "leave" (i.e. at home). This may be done by modeling Oncology patient flow using the Semi-open Erlang-R queue (Yom-Tov (2010), Part II), introduced in Figure 23. (Note that this semi-open version is different from the open version introduced earlier in Section 3.2.) In this case, "Needy" patients are those who are currently hospitalized in the ward, "Content" patients are the ones who are in the midst of a series of treatment but are currently at home. The model enables one to separate the two streams of incoming patients: new and returning, and set each one its own service goals. The probability of blocking (P(Block)) measures the proportion of new patients that cannot be admitted to the ward, while the probability of waiting measures waiting for beds of in-process patients. We set s to be the number of beds allocated to the ward, while n is the maximal number of patients treated by the ward. Returning patients should not wait long for a bed. The target value for P(Block) may be determined by revenue considerations, as each patient blocked is treated in a different facility.

Yom-Tov (2010) proposes an indicator, the *offered load ratio*[5] B, for when returning patients must be acknowledged explicitly, as opposed to being absorbed within the arrival process of new arrivals. $B = \frac{\frac{1}{(1-p)\mu}}{\frac{p}{(1-p)\delta}}$ where μ is the service rate in the Needy station, δ is the service rate in the Content station, and p is the return probability. The value of the offered load ratio for Oncology wards in Rambam hospital equals $B = \frac{1/22}{0.817 \cdot 1/3.42} = 0.19$. The analysis in Yom-Tov (2010), suggests that with such an offered load ratio, the influence of returning patients is very important. Therefore,

[5] The offered load ratio is the ratio between the offered load of the Needy station and the offered load of the Content station.

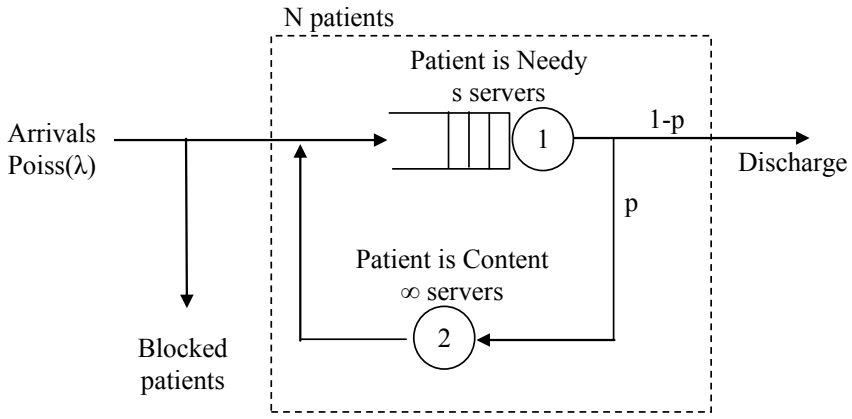
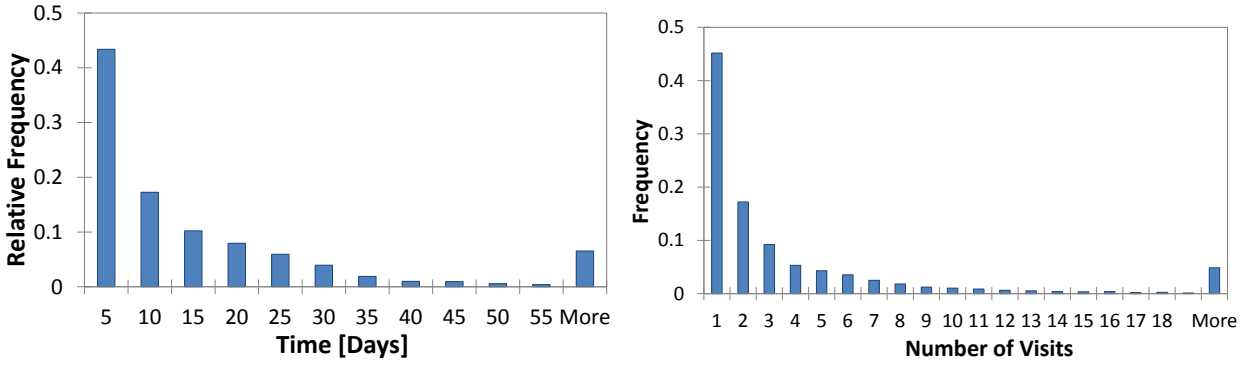


Figure 23 **Semi-open Erlang-R queueing network**

we argue that, the semi-open Erlang-R model is more suitable than the Erlang-B model (of Bekker and de Bruin (2010)) for determining bed allocation for Oncology wards.

Lastly, we would like to raise two questions concerning unplanned and planned returns:

- *Modeling planned returns:* Oncology treatments typically have fixed protocols with deterministic times between returns and a fixed number of planned returns. The question arises as to whether it is appropriate to use the proposed queueing model of Figure 23, with random returns, to determine bed allocation? To answer this question, we note that although each patient may have a deterministic protocol, there are many such protocols and, therefore, a mix of a large number of patients may result in a random-like distribution of the time between returns and the number of returns. This is exactly the case in Rambam Hospital as demonstrated using Figure 24.



Time between successive returns The number of returns

Figure 24 **Distributions of returns to Oncology wards**

- *Modeling unplanned returns:* When unplanned returns reflect poor quality of service, there is an additional complication as the rate of returns may be endogenously determined by clinical and operational decisions. Such erroneous decisions could be an outcome of ward overload; For example, misdiagnosis and blocking may occur more frequently when the system is overloaded. We argue that this aspect should be explicitly modeled when making staffing and capacity decisions, as offered load affects capacity, which in turn affects rate of returns, which in turn affects the offered load. In particular, one may wish to identify staffing and capacity levels that obtain an equilibrium in this cycle.

ICUs represent an example of endogenous returns, as verified by Kc and Terwiesch (2009). They showed, using patient flow data of ICU units in a US hospital, that the readmission probability can increase as a result of ICU overload. In this case, operational decisions such as whether to speedup patients LOS and move them to step down units during overloaded periods, clearly influence medical outcome and secondary LOS. The returns in ICU occur within days, and therefore assuming independence, between first and second visits, is not reasonable. One needs to explicitly model the endogenous nature of these returns, as is done by Chan et al. (2014), who capture the dependencies between ICU overload, LOS, and readmission probability, using a state-dependent queueing model.

4.3. Economies of scale

When applying the theory of many-server queues to call centers, it has become customary to distinguish between three main operational regimes: The Efficiency-Driven regime (ED-regime), the Quality-Driven regime (QD), and the Quality- and Efficiency-Driven regime (QED). The ED-regime emphasizes the efficiency of resources: servers are highly utilized (close to 100%), and hence customers typically suffer a relatively long (though still tolerable) wait for service. In the QD regime, the emphasis is on the operational quality of service: servers are available for service for ample time, and consequently customers hardly wait for service. The QED regime is somewhere in between: the emphasis is on carefully balancing service quality and servers efficiency, aiming at high levels of both. In large systems, that operate in the QED regime, we find that server utilization could reach 90% and more while, at the same time, around half of the customers are served immediately upon arrival. The QED regime also exhibits economies-of-scale: for example, as the system grows, a fixed occupancy level results in better performance.

In the following section, we explore the question of which of these regimes best fits the IWs operations. We discuss this in relation to beds as well as physicians. In addition, we question the existence of the economies-of-scale phenomenon in the hospital environment. We shall argue that, although in general the IW beds operate in the QED regime, there is nevertheless evidence for diseconomies of scale.

4.3.1. In what regime do the Internal wards operate? Can QED- and ED-regimes co-exist? We start by identifying the operational regimes that are relevant to our system of IWs. As the system consists of multiple types of servers (beds, nurses, physicians), each must be considered separately. Here we focus on beds and physicians.

We argue that bed capacity of the IWs is managed in the QED regime. To support this statement, and in view of the fact that our data does not include staffing levels of nurses nor does it track nurses service times, we propose to fit a loss model (Erlang-B, as in de Bruin et al. (2009)) to help estimate theoretical value for the probability of blocking, namely the probability that a to-be-hospitalized patient does not find an available bed in any of the IWs.

Consider an M/M/N/N queue (Erlang-B). In the QED regime ($N \approx R + \beta\sqrt{R}$), $P(Block) \approx \frac{h(-\beta)}{\sqrt{N}}$, where $h(\cdot)$ is the hazard rate of the standard normal distribution. It follows that the occupancy level $\rho \approx 1 - \frac{\beta + h(-\beta)}{\sqrt{N}}$, hence both P(Block) and $1-\rho$ are $O(1/\sqrt{N})$. Considering Rambam hospital's from 2008, we find that the average LOS in all IWs is 5.12 days; there are 186 beds in all the IWs, and the total arrival rate is 34.4 patients per day. Thus $\beta \approx \frac{N-R}{\sqrt{R}} = \frac{186-34.4*5.12}{\sqrt{34.4*5.12}} = 0.4$, and $P(Block) \approx 2.9\%$. According to our data, the fraction of patients that were physically hospitalized in other wards, but were still under the medical responsibility of IW physicians, is 3.54% of the patients, which is quite close to the theoretical value of 2.9%. In addition, the approximation for the occupancy level is $\rho \approx 91.7\%$, which is again very close to the actual data-based value of 93.1%. These two facts support our hypothesis that IW beds operate in the QED regime.

Turning to physicians as servers, we argue that they operate in the ED-regime. This is based on the following observation: from 4pm to 8am the following morning, there is only one physician on duty in each IW. This physician admits most of the new patients of the day. Therefore, patients that are admitted to an IW (only if there is an available bed) must wait until both a nurse and the physician on call are available. The admission process by the physician takes on average 30 minutes. Thus, a plausible model for the waiting of a patient in the ED, until transferred to one of the IWs, is an $M_t/G/1$ model. We hypothesize that physicians operate in the ED-regime since mean service time is 30 minutes, while the waiting time of a patient to be transferred from the ED to an IW is 3.1 hours, on average (see Section 5.2). This is a characteristic of the ED-regime, where waiting time is about an order of magnitude more than the service duration.

Implications to queueing modeling: We identified two operational regimes, QED- and ED-regime, that coexist in a single system. Which queueing model and operational regime (scaling) can describe this phenomenon? Note that such model must accommodate three time scales: minutes for physician treatment, hours for transfer delays, and days for hospitalization LOS. Focusing on the physicians (the ED-regime resource), the following questions naturally arise: How do the regimes influence each other? Can we assume that the "bottleneck" of the system is the "ED-regime"

resource? Can we conclude that adding physicians would reduce waits for transfer while adding beds will have a marginal impact on these delays? How would a change of priority in a physician's operation influence the system, say giving higher priority to incoming patients over the already hospitalized? Does the fact that physicians operate in the ED-regime eliminate the economies-of-scale that one expects to find in QED systems? The following discussion suggests that this indeed might be the case.

4.3.2. Diseconomies of scale (or how does size affect LOS) Our data reveals, in some cases, diseconomies of scale, namely the smallest ward has relative workload that is comparable to the larger wards, yet it enjoys higher turnover rate per bed and shorter ALOS, with no apparent influence on medical care quality. To understand the reasons behind this phenomenon, we interviewed physicians and nurses. We present here their interpretation, and raise questions regarding the significance of incorporating such aspects into models of medical wards. We start with describing the diseconomies of scale phenomenon itself.

Wards A–D provide similar medical services. However, they do differ in their operational measures: Capacity and ALOS. Ward capacity is measured by its number of beds (fixed capacity) and the number of service providers – physicians, nurses, administrative staff and support staff (processing capacity). Generally (and in Rambam hospital IWs in particular), processing capacity is determined proportionally to fixed capacity (see, however, discussions on the deficiencies of such "proportional" staffing in Jennings and de Véricourt (2008)). As an example, in Rambam Hospital, an IW nurses-to-beds ratio in morning shifts is 1:5 or 1:6 (depending on the size of the Transitional Care Unit (TCU) within each Internal ward; up to 5 beds). It follows that a ward's operational capacity can be characterized by its number of beds only – denoted as its *standard* capacity. The *maximal* capacity, on the other hand, stands for the standard capacity plus extra beds, which can be placed outside the usual designated areas during overloaded periods.

Medical units are further characterized by various performance measures: operational - average bed occupancy level, Average Length of Stay (ALOS), waiting times for various resources, number of patients admitted or released per bed per time unit (flux); and quality of care, captured by patients' return rate, patients' satisfaction, mortality rate, etc. Note that occupancy levels and flux are calculated relatively to a ward's standard capacity. (Thus, occupancy can exceed 100%.)

Comparing the two basic measures, ward capacity and ALOS, we observe in Table 2 that the wards differ in both. Indeed, Ward B and E are significantly the smallest and the "fastest" (shortest ALOS) among Wards A–E. It is intuitive that Ward E will have shorter ALOS, as Ward E treats the clinically simplest cases. The numbers associated with Ward B are surprising, however, because that ward is assigned the same patient mix as IWs A,C, and D.

Table 2 Internal wards: operational profile

	Ward A	Ward B	Ward C	Ward D	Ward E
Average LOS (days) (STD)	6.0 (7.9)	3.9 (5.4)	4.9 (10.1)	5.1 (6.6)	3.7 (3.3)
Mean occupancy level	97.7%	94.4%	86.7%	96.9%	103.2%
Mean # patients per month	206.3	193.5	209.7	216.5	178.7
Standard (maximal) capacity (# beds)	45 (52)	30 (35)	44 (46)	42 (44)	24
Mean # patients per bed per month	4.58	6.45	4.77	5.16	7.44
Readmission rate (within 1 month)	10.6%	11.2%	11.8%	9.0%	6.4%

Data refer to period May 1, 2006–October 30, 2007 (excluding the months 1–3/2007, when Ward B was in charge of an additional 20-bed sub-ward).

Short ALOS could result from a superior efficient clinical treatment, or a liberal (as opposed to conservative) release policy; a clinically too-early discharge of patients is clearly undesirable. As mentioned in Section 4.2.2, one possible (and accessible) quality measure of clinical care is patients return rate (proportion of patients who are re-hospitalized in the IWs within a certain period of time; in our case, three months). In the same table, we observe that the return rate in Ward B does not differ much from the other wards. Also, according to Elkin and Rozenberg (2007), patient satisfaction level does not differ in Ward B from the other wards. We conclude that Ward B is operationally superior yet clinically comparable to the other wards.

In order to understand why ward B seems to have such superior operational measures, we interviewed the management of the Internal wards, to see if there are differences in policies and management methods. The most significant difference we found was that ward B dedicated one of its physicians to the task of reducing delays caused by other units of the hospital. This physician coordinates medical tests and specialist-services, and ensures that requests for such services will be accommodated fast, possibly by overriding hospital rules. For example, on occasion Ward B has sent patients to Echocardiogram tests, in excess of what is allowed per day by hospital regulations. These regulations were made to guarantee equal access to medical care by all wards, thus this practice will not work if all wards exercise it. Nevertheless, reducing delays is crucial at the beginning of a patient stay, when a treatment plan is outlined. It can reduce ALOS, since the right treatment is administered earlier. But this explanation does not tell the whole story. Such central control of patient and test flow in a ward is easier to exercise in a smaller ward. As a supporting evidence, see Ward B ALOS and average number of patients in 2007, in Figure 25.

We observe that, during 2007, the ALOS of Ward B was significantly increased. This was due to a temporary capacity increase, over a period of 2 month; a time when IW B was made responsible, temporarily, for 20 additional beds. We observe that, although the same policies were used, they

seemed to work better in a smaller ward. In addition, we note a reduction in the LOS of IW D, mainly from 2007 when ward size decreased as a result of ward renovation. This raises the conjecture that there is also an effect of the size of the ward, namely diseconomies of scale in our case: larger wards are more challenging to manage, which has a negative effect on the ALOS of patients.

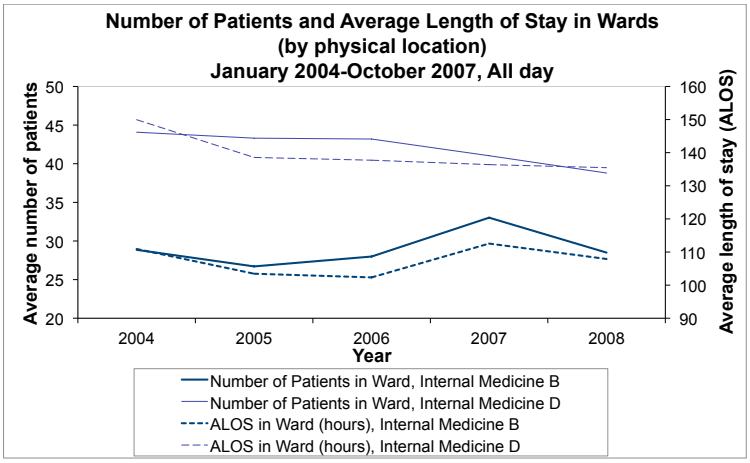


Figure 25 Average LOS and number of patients in Internal wards B and D by year

A few factors could limit the economies-of-scale and might explain why we actually observe diseconomies of scale:

1. *Staffing policy*: It is customary to assign an IW nurse to a fixed number of beds, and nominate one experienced nurse to be a *floater* for solving problems, and help as needed. This setting does not give any advantage to large units; perhaps on the contrary, in larger units, a single floater will have less time to help each nurse. Therefore, larger units need not work better. Note that there exists a rich literature on the subject of server flexibility (Aksin et al. (2007a), Jouini et al. (2009)); The issues that we raise regarding economies-of-scale could well be relevant to that literature.

2. *Centralized medical responsibility*: The physicians in the ward share responsibility over all patients. Every morning, the senior physicians, interns, and medical students examine every case together (physicians' rounds), and discuss the treatment needed. This is essential as Rambam hospital is a university hospital, and one of its central missions is educating and training new physicians. Naturally, in larger units, these morning rounds take longer and consequently, less capacity is allocated to other tasks - this could lead to prolonged ALOS.

3. *External resources allocation*: In some cases, external resources (such as ultrasound for echocardiogram tests) are allocated as a fixed number of tests per IW unit per day. This gives an advantage to smaller units.

It is important to further investigate this form of diseconomies of scale, and model it carefully. It can significantly affect the perception of what is the optimal unit size. In addition, if we take size differences among wards as a given fact (which is natural, for example due to space constraints that cannot be resolved), the following question arises: What patient routing scheme should be used to route patients from the ED to the wards, in order to fairly distribute workload among them? This ED to IWs routing challenge is the subject of the next section.

5. Transfer from the ED to IWs

The present section focuses on patient transfer from the ED to the IWs ("ED-to-IW process"). We focus on two operational problems that commonly arise in the ED-to-IW process: long patients' waiting times in the ED for a transfer to the IWs; and possible lack of fairness towards patients and wards staff. Our analysis is supported by empirical data from Rambam hospital, five additional Israeli hospitals, and one hospital from Singapore we have contact with. We view the process in the context of *flow control*, where patients are *routed* from the Emergency Department to the Internal wards.

There is a rich literature on flow control, in healthcare and beyond. Originated from the telecommunications field, the topic is also relevant in software, traffic, manufacturing and more. As discussed in Section 1, the issue of patient flow control in hospitals is of great importance.

Clearly, routing of *people*, as opposed to data or physical goods, gives rise to different operational issues. However, routing in *hospitals* also differs from routing in other service systems for various reasons including incentive schemes, customers' lack of control, and the timing of the routing decision. Thus, although the transfer process may seem to involve routing related issues similar to those that have been looked at extensively in the queueing literature, our data indicates that these unusual system characteristics significantly affect delays and fairness features in a hospital setting. Thus, studying the transfer process in this setting provides many new research opportunities.

This section is organized as follows. We begin by providing some background on the ED-to-IW process in Rambam hospital - in Section 5.1. Then, in Section 5.2, we describe our empirical findings with respect to delays in the transfer process. In Section 5.3, we analyze the causes for those delays. In Section 5.4, we explore the relationship between delays in the ED-to-IW transfer process and workload in the IWs. Next, in Section 5.5, we discuss the issue of *fairness*, distinguishing between patient fairness and ward fairness, and how routing influence both types. Finally, in Section 5.6, we extend our viewpoint, via an inter-hospital study, and investigate how other hospitals tackle the problem of routing.

5.1. ED-to-IW background

We begin with some background on the patient transfer process from the ED to the IWs in Rambam hospital - as is illustrated in Figure 26. A patient, whom an ED physician decides to hospitalize in an IW, is assigned to one of the five wards, according to a certain *routing policy* (described later). In case that the ward refuses to admit the patient (usually for reasons of overloading), a Head Nurse may approve a so-called "skip"; in which case, a reassignment occurs. Once a ward agrees (or is forced) to admit the patient, the ward staff prepares for this patient's arrival. In order for the transfer to start, a bed and medical staff must be available, and the bed and equipment must be prepared for the patient (including potential rearrangement of current IW patients). Up to that point, the patients wait in the ED and are under the ED's care and responsibility. If none of the wards is able to admit the patient within a reasonable time, the patient is "blocked", which implies a physical transfer to a non-internal ward that takes responsibility for their nursing (the patient's medical treatment is still obtained from an IW physician).

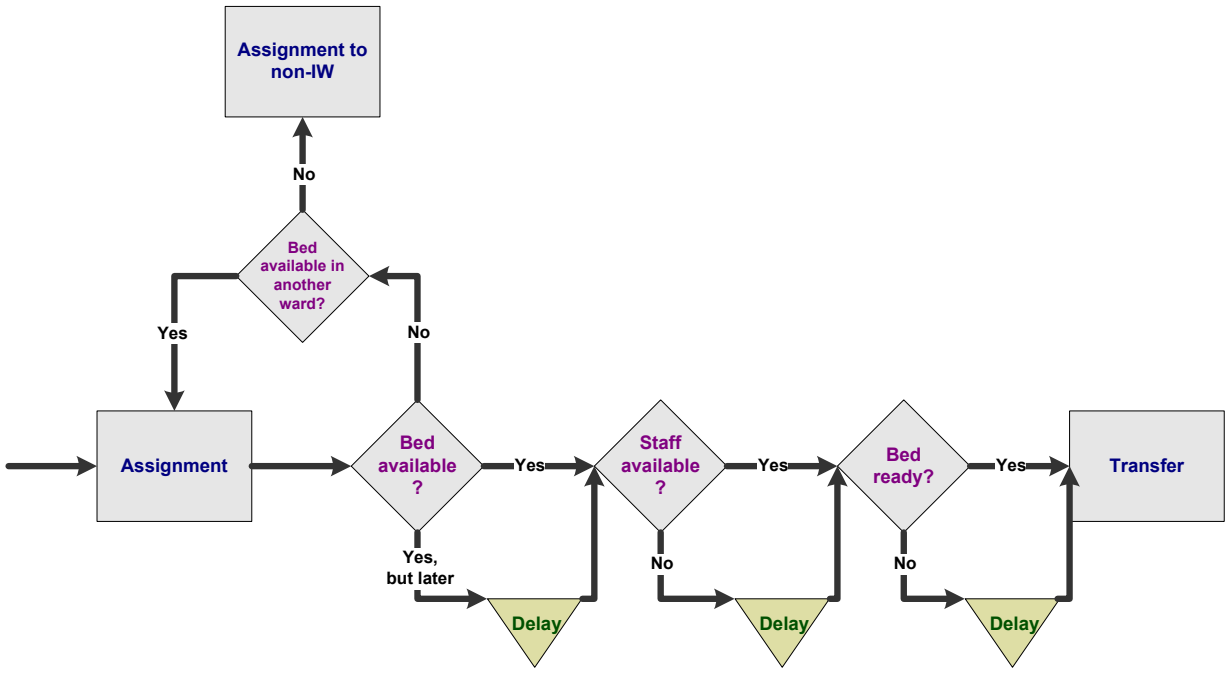


Figure 26 **Abridged process flow diagram for the ED to IW transfer process**

An integral component of the transfer process is a *routing policy*, or an assignment algorithm of patients to the Internal wards. As described in Section 4.1, Wards A-D provide similar medical services, while Ward E only treats "walking" patients. The similarity between Wards A–D requires an assignment scheme of patients to wards. Rambam hospital determines the assignment based on a round-robin (cyclical) order among each patient type (ventilated, special care, and regular), that

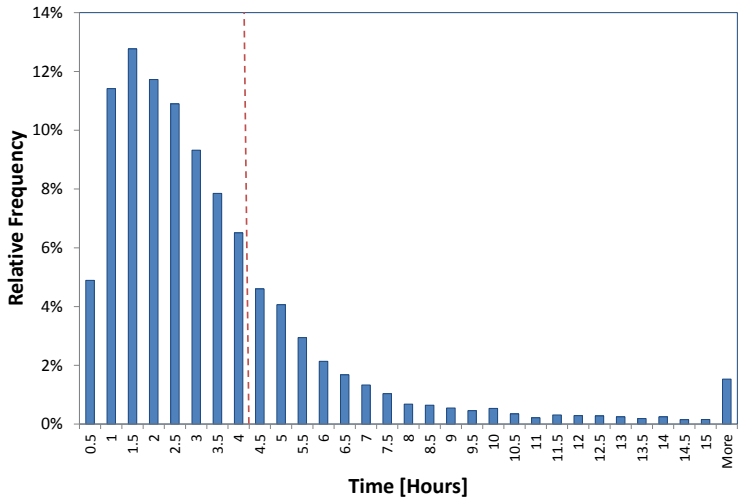
also accounts for the ward size (e.g. if Ward X has twice as many beds as Ward Y, then Ward X gets two assignment per one assignment of Y). This routing scheme is implemented by a computer software called "The Justice Table". Other hospitals apply alternative assignment schemes, as will be discussed later (§5.6).

5.2. Delays in transfer

As mentioned earlier, patients to-be-hospitalized in an IW wait in the ED until their transfer is carried out. There exists a hospital policy obliging the Internal wards to admit patients within *four hours* from the decision of hospitalization, but in certain cases these delays end up being significantly longer. In this section, we estimate those delays in Rambam hospital based on empirical data, discuss their implications on patient health, the unique operational features of this queueing system, and how does the presence of such delays relate to ED architecture.

As exact data on patient delays are not kept in the hospital information systems; Instead, we estimate these delays based on the time that has elapsed from patient assignment to an IW until receiving the first treatment in an IW. This is clearly an overestimate of the actual delay, because the patient might be present at the IW for some time before the first treatment. However, in a time and motion study, Elkin and Rozenberg (2007) found that indeed a significant portion of these times is spent in the ED. We, therefore, use this estimate as our proxy for the actual delay in transfer. The waiting-time histogram for the years 2006-2008 in Wards A–D is depicted in Figure 27. We observe that the delays are significant: for example, the average delay was 3.2 hours (for Wards A-D), with 23% of the patients being delayed for more than 4 hours.

Waiting distribution by patient type: An interesting phenomenon is observed when analyzing delays in transfer by patient types. We note that different types of patients wait different amount of time. Specifically, on average, ventilated patients wait much longer (8.4 hours) than regular and special care patients (average of 3 and 3.3 hours respectively) - see Table 3. We also observe higher standard deviation for ventilated patients. Figure 28 shows the delay time distribution by patient type. The delay histograms for Regular and Special-care patients are similar and they resemble the overall delay histogram of Figure 27. Indeed, it is not surprising that their delays are similar to one another, as they share the same resources and do not differ in priority. Moreover, these two types of patients amount to over 98% of the overall transfer patient population. Ventilated patients have, in theory, a higher priority. However, in reality, they do not seem to benefit from it. We see that, for a significant proportion of the ventilated patients, the delays are very high; 41% of the patients wait 10 hours or more. On the other hand, there is still a significant population that has reasonable delays; 41% of these patients wait less than 4 hours. The shape of the distribution is very different from the other two patient types, with an apparent bi-modal shape.



* Data refer to period 5/1/06 - 10/30/08 (excluding the months 1-3/07 when Ward B was in charge of an additional sub-ward)

Figure 27 Waiting time histogram (Average = 3.22 hours)

Table 3 Average transfer delays per patient type (hours)

Patient Type	Average Delay	Standard Deviation	% delayed up to 4 hours	% Delayed more than 10 hours
Regular	3.00	2.53	77%	2%
Special Care	3.33	3.16	74%	5%
Ventilated	8.39	6.59	41%	41%
All Types	3.22	2.98	75%	4%

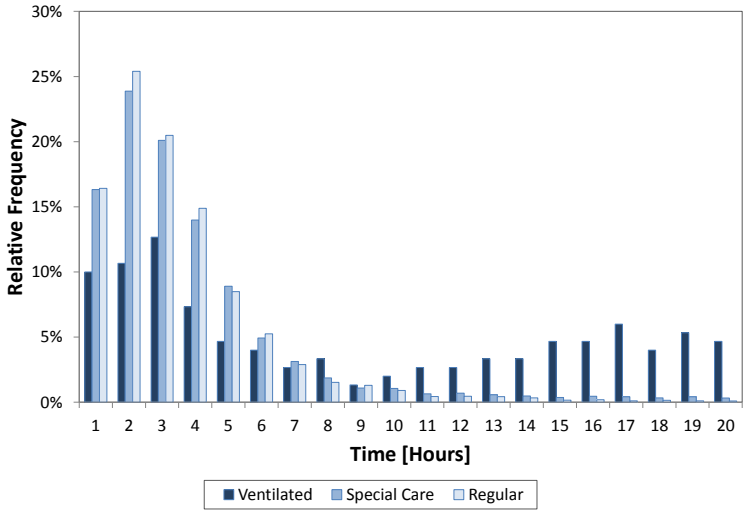


Figure 28 Waiting time histogram by patient type

The question arises as to how come the ventilated patients experience such long delays with a distinctly different distribution. The shorter delays (< 4 hours) have similar pattern as the other two patient types have. The longer delays are harder to decipher. Possible explanations include: 1) Ventilated patients are hospitalized in a *sub-ward* inside of the IW (A-D), often referred to as Transitional (intensive) Care Unit (TCU). Each such TCU has only 4-5 beds. The average occupancy rate of the TCUs at Rambam hospital is 98.6%; the combination of high occupancy with a small number of beds results in much longer waits in overloaded periods. 2) Ventilated patients impose significantly higher load on ward staff than other patients. Thus, a ward may strive to admit a ventilated patient at a more convenient time, especially if it is in a state of over-loading. 3) Ventilated patients require a highly qualified staff to transfer them to their ward (especially since they carry an oxygen tank). Coordinating such transfers takes longer. Also, preparation in the IW might take longer, as sometimes special equipment is needed.

In general, the ventilated patients are delayed until the IWs can provide them with the special conditions they require. These patients add significant load on the ED when they wait there due to their clinical severity. Hence, one should strive to make those long delays significantly shorter.

Clinical consequences of long delays: Patients waiting for transfer overload the ED, as beds remain occupied while new patients continue to arrive, and the ED medical staff remains responsible for transfer patients as long as they are within the ED[6]. Therefore, ED actually takes care of two types of patients: *transfer patients* (patients awaiting hospitalization), and *in-process patients* (patients in evaluation or treatment in the ED). Both types may suffer as the consequence of the transfer patients' delays.

Transfer patients may experience significant discomfort while waiting: the ED is noisy, it is not-private, and does not serve hot meals for patients. In addition, the patients do not enjoy the best professional medical treatment for their situation, and do not have dedicated attention as in the wards. Moreover, they may be exposed to more germs and diseases; hence the longer the patients wait in the ED, the lower their satisfaction and the higher the likelihood for clinical deterioration (Maa (2011)).

In-process patients may suffer from delays in treatment, as the additional workload imposed by transfer patients can be significant. In Section 2, we showed that the additional workload on ED physicians inflicted by transfer patients can be high as 10% (see Figure 7), and that overload contributes to throughput degradation in the ED. Certainly, that may impair patient safety.

Hence, improving the efficiency of patient flow from the ED to the IWs, while reducing waiting times in the ED, will improve the service and treatment provided to patients. In addition, reducing the load on the ED will lead to a better response to arriving patients and is likely to save lives.

[6] In contrast, a Singapore hospital has transfer patients become the responsibility of the IW physicians, after two hours of delay in the ED (see Section 5.2).

Implications on queueing research: The delays in transfer give rise to two interesting questions:

1. *Modeling transfer queue:* Transfer patients may be viewed as customers waiting in queue to be served in the IW. Traditionally in queueing theory, it has been assumed that the customers receive service only once they reach the service station, and not while waiting in queue. In contrast, here while a patient waits s/he is "served" by both the ED and the IW. In the ED, clinical treatment is provided: according to regulations, transfer patients need to be examined at least as frequently as every 15 minutes. In the ward, the "service" actually starts prior to the physical arrival of the patient, as the ward staff, once informed about a to-be-admitted patient, starts preparing for the arrival of this *specific* patient. The above has implications on the ED-to-IW process modeling, and affects staffing, work scheduling, etc.

2. *Emergency Department architecture:* In Section 3.3, we discussed ED architecture (see Figure 18), accounting for different types of patients arriving to the ED. Taking into consideration delays in transfer may also influence the ED structure. We noted above that the ED staff takes care of two types of patients: transfer and in-process patients. Each type has different service requirements, leading to differing service distributions and differing distribution of time between successive treatments. While transfer patients receive periodic service according to a nearly-deterministic schedule (unless complications arise), in-process patients service is much more random.

One may consider two options: (a) treat transfer and in-process patients together in the same physical location, as done in Rambam hospital, or (b) move the transfer patients to a transitional unit (sometimes called "delay room" or "observation room"), where they wait for their transfer; as done for example, in the Singapore hospital we have contact with. Note that using option (b) implies having dedicated stuff, equipment and space for this unit.

The following question naturally arises here is: how to characterize the conditions under which each of these ED architecture is more appropriate. In addition, how shall we model each system? For architecture (a) one can develop a multi-class variation of the Erlang-R model; for (b) *tandem queues* or multiple-stages queues are more appropriate, where the transfer room serves as a "buffer" between the ED and the IWs.

Note that the Singapore hospital architecture is even more complicated than (b), as the responsibility for the transfer patients is handed over to IW physicians after a two-hour delay. The IWs physicians need to come from the ward to take care of the patients in that transfer room . This provides the IW medical staff an *incentive* to transfer the patients to the ward as soon as possible to comfortably treat them. We shall discuss later, in Section 5.6, how different incentive schemes may influence delay times.

5.3. Causes of the delays

In order to understand the causes of the long delays in the transfer between the ED and the IWs, we interviewed hospital personnel, conducted a time and motion study, and further analyzed our data. We have learned that delays are not only caused by bed unavailability; patients often wait even when there are available beds. Indeed, our data show that the fraction of patients who had an available bed in their designated ward, upon their assignment time, was 43%, 48%, 76%, 55%, for Wards A-D, respectively. However, as Figure 27 shows, the probability to be admitted to the wards, immediately (or within a short time) after the hospitalization decision, was much smaller. In fact, over the same period of time, only 4.9% of the patients were admitted to an IW within 30 minutes from their assignment to this ward. Our findings identify additional causes for delay, which are summarized in the cause-and-effect (fishbone) diagram depicted in Figure 29.

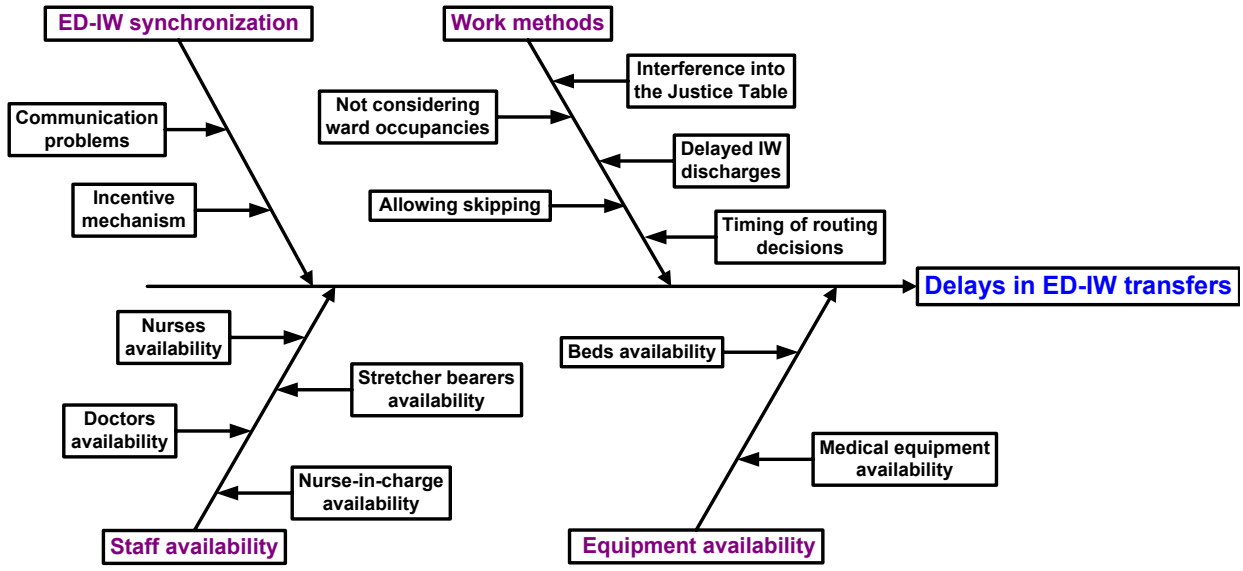


Figure 29 **ED-to-IW delays: Causes and effects chart**

Implication to queueing models: Figure 29 is mostly a result of our time and motion study. However, to understand some more subtle causes, we turn to queueing theory. For example, the timing of the routing decision is related to the question of using *Input-queued* vs. *Output-queued* system, when routing patients from the ED to the IWs. Another important aspect is how *information availability and its utilization* during the routing influence transfer delays.

1. *Input-queued system vs. Output-queued system:* In our transfer process, each patient is first assigned to an IW and then waits until the ward is able to admit him/her. This is in contrast to a scheme in which patients are placed in a "common" queue, and are routed to an IW only once at the head of the queue, and a spot at *any* of the IWs becomes available. Stolyar (2005) refers to

the former as an *output-queued* system, in contrast to the latter, which is an *input-queued* system. Figure 30 depicts the two schemes.

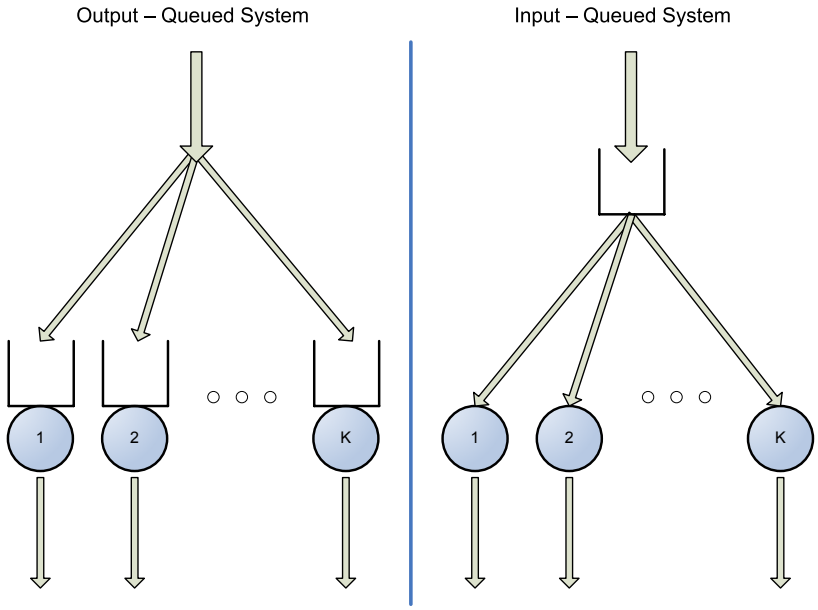


Figure 30 **Input-queued system vs. Output-queued system**

Output-queued systems are inherently less efficient than their input-queued counterparts, because the routing decision is made at an earlier time when less information is available to the decision maker. Moreover, the output-queued system is inequitable towards patients, because FCFS is often violated (we return to this in Section 5.5.1). The reasons for the hospital to adopt the former and not the latter (as learned from interviewing the key people in the hospital) are administrative and medical (the ward must prepare for admitting a *particular* patient, including customized medical equipment and administrative treatment), as well as psychological (uncertainty with respect to the IW assignment should be resolved for the patient at an early stage) (Elkin and Rozenberg (2007)).

The problem of customer routing in input-queued systems has received considerable attention in the queueing literature (e.g. Mandelbaum and Stolyar (2004), Armony (2005), Atar and Shwartz (2008), Gurvich and Whitt (2010)). The same issue in output-queued systems has received much less attention. Nevertheless, Stolyar (2005) and Tezcan (2008) establish that, asymptotically the two systems have similar performance under both conventional and the Halfin-Whitt heavy traffic regimes, respectively. This suggests that inefficiences that arise in our ED-to-IW process due to the use of output-queued instead of input-queued systems become negligible in highly loaded systems. More generally, insights gained from studying the input-queued systems, as in the above references, may carry over to the output-queued systems.

2. *The role of information availability in routing and its influence on transfer delays:* An additional important aspect of routing schemes that directly affects patient delays, is availability of information in the system at the moment of the routing decision. We identify three categories:

(a) *No information*, meaning that at the moment of routing decision only static parameters are known (number of beds, ALOS), and the system state is unknown. Routing policies operating under no information are round-robin schemes or random assignment according to some constant probabilities.

(b) *Full information*, meaning that dynamic system parameters, such as occupancy rate, are known. Here we recognize several levels of information availability: full mode, e.g. via RFID or smart-beds that record continuously and reliably patients' status (where and in what process step they are at) and beds' status (idle / occupied / cleaned)[7]. In this case routing policies commonly used in call centers (like Longest Idle Server First (LISF)) may apply. A lower level of information availability is knowing occupancy status in each ward. Then routing schemes like Randomized Most Idle (RMI) can be utilized (more on these policies - in Section 5.5.2).

(c) *Partial information*, of two forms: either information is revealed only in certain cases (e.g. when a patient is blocked one learns that the ward is full); or information is revealed periodically (e.g. bed census every morning). In such cases, one can estimate necessary information at the time of the routing decision based on the latest update.

Currently, the ED-to-IW routing at Rambam hospital is determined according to the Justice Table and follows a cyclical order taking into account only patient type and ward size. This is a "no information" scheme; in particular, the algorithm does not take into account ward occupancies at the time of the routing decision. Thus, a patient may be routed to a full ward while another ward has available space, which is certainly undesirable. To remedy the situation, the process allows an IW to skip its turn, but that requires an ad-hoc negotiation between the IW and the ED and, possibly, a reassignment. From our data we observe that, on average, 13.2% of the patients were reassigned (sometimes more than once) and the average time between reassignments was about 1 hour. Clearly, the routing policy directly affects patients' waiting times.

Note that Rambam hospital can utilize routing of Category (c), since the information, that is available in the ED about IW beds occupancy, is updated once a day (in the morning). The exact information on the number of available beds is unavailable at the moment of routing. A hospital with beds management unit may have such information. The question is, how the use of such information will affect the patients' delays? We return to this question in Section 5.5.2.

[7] Note that additional information on *future* occupancy may also be obtained, such as the number of patients expected to be released in the next few hours and the number of elective patients scheduled to arrive.

5.4. Delays in transfer versus load in IW

In this section, we explore the relationship between delays in the ED-to-IW transfer process and workload in the IWs. We show how workload in the IWs affects transfer delays, and thereby argue that the workload information should be incorporated in routing decisions. This elaborate on system view discussion on in Section 2.

It is natural to expect that the higher the workload in the IWs, the longer the delays in transfer. Figure 31 shows the average delay in transfer alongside the average number of patients per ward - in IWs A-D, by day of the week. We observe that, as expected, the two measures have a similar weekly pattern.

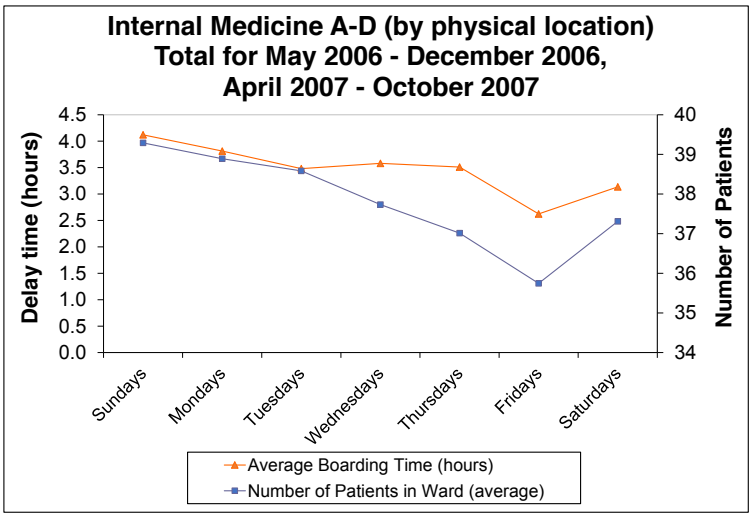


Figure 31 **ED-to-IW transfer delays and number of patients in IW**

Figure 32 displays delays in the transfer process and the average number of patients in the IWs as they vary throughout the day. The correlation here is not as apparent as in Figure 31; other factors, such as the *IW discharge process*, play a role. We observe that the longest delays are experienced by patients assigned to the IWs in early morning (6am-8am) - these patients need to wait on average 5 hours or more. This is due to the fact that IW physicians perform their morning rounds at this time and cannot admit new patients. Then we see a consistent decline in the transfer delay up until noon. Patients assigned to the IWs during these times, are admitted into the IWs between 1pm-3pm. This is about the time when the physicians' morning rounds are complete; staff and beds are starting to become available. Indeed, there is a sharp decline in the number of patients around 3pm-4pm, when most of the IWs discharges occur - as shown in Figure 6. In Section 2.4, we also discussed how those delays impact physician workload in the ED. We note here that further data analysis reveals that patients that are transfered to the IWs before 9am have significantly shorter

LOS; early hospitalization reduces ALOS by 1 day. Thus, we argue that it is extremely important to shorten the ED-to-IW transfer process and improve the admission process in the IWs so that the first hospitalization day is not wasted.

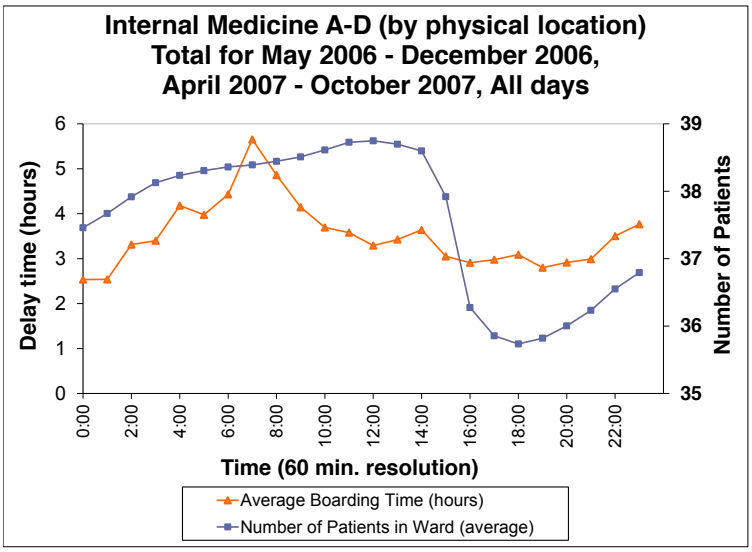


Figure 32 **ED-to-IW transfer delays and number of patients in IW**

Operational implications and research opportunities: The backwards ripple effect that the IW discharge policy has on the transfer delays and ED productivity implies that redesign of that process might affect patient's flow dramatically. Major delays occur due to the fact that during the IW physicians' rounds no patients are admitted. We conjecture that, if the discharge time and physician rounds' time were more uniformly spread over the duration of the day, the delay in transfer could be reduced significantly. To test the impact of such a change one would need to estimate the specific time a patient is available for discharge, which is not included in our data.

The hourly connection between the delays in transfer and the IW workload suggests the need to use a *closed-loop* routing control, that takes into account real-time wards' workload, as well as future information on *planned* discharges. We propose a more rigorous statistical analysis to study the correlation between these measures, to determine exactly what information is the most useful in making these routing decisions.

5.5. Fairness in the ED-to-IW process

Service systems may encounter issues related to fairness towards *customers* (patients) and fairness towards *servers* (medical and nursing staff). This section examines both in the context of the ED-to-IW process.

5.5.1. Patients - fairness There is ample literature on measuring and controlling for fairness in queues from the customer point of view (see references in Mandelbaum et al. (2012)). Various aspects are investigated (for example, single queue vs. multi-queues, or FCFS vs. other queueing disciplines), but all agree that the FCFS policy is typically key in justice perception. Consequently, customer satisfaction in a single queue is higher than in multi-queues (Larson (1987)); and waiting in a multi-queue system produces a sense of injustice even when no objective discrimination exists (Rafaeli et al. (2002)). When discussing input / output queues in Section 5.3, we noted clinical and psychological reasons for using multi-queues (output-queues) in the ED-to-IW process. Besides longer delays, output-queues lead to diminished patients-fairness because FCFS is often violated.

Our data indicate that 45% of the transfer patients were "overtaken" by another patient in being transferred from the ED to the IW. That is, 45% of those patients had at least one "slip" (in the terminology of Larson (1987)). Moreover, more than a third of those were overtaken by at least three other patients. This number indicates a significant FCFS violation in the process. However, this figure includes overtaking between patient types, which may be due to clinical considerations. Table 4 summarizes our findings with respect to FCFS violations per patient type. The last row shows that, even within each patient type, there were significant FCFS violations, with 31% Regular and Special Care patients being overtaken by at least one patient of the same type. In order to examine what fraction of these FCFS violation may be attributed to the output-queued scheme, we look for FCFS violations within each ward. Such violations may not be attributed to this scheme because each IW has its own queue of patients waiting to be admitted. The numbers show that for both Regular and Special-Care patients, roughly 7% of the patients were overtaken by other patients of the same type within the same ward[8]. So overall, for these two patient types, about 24% were overtaken by other patients of their own type, likely due to the output-queued scheme. This is a significant number. It would be interesting to test whether factors such as diagnosis codes affect FCFS violations. It is worth noting that Ventilated patients suffered much fewer FCFS violations. This might be due to the fact that there are much fewer such patients (2% of the overall IW patients).

While output-queues are inherently inefficient and unfair, they are also unlikely to change, at least not in Rambam hospital. The use of output-queues in the ED-to-IW process illustrates the uniqueness of flow control in healthcare - it is an example of how constraints that are dictated by clinical and psychological considerations affect the process and lessen its efficiency and fairness.

Research opportunities: Since the output-queued scheme is inherently unfair, one may naturally ask whether it is possible to maintain relative fairness even if FCFS is not preserved. Specifically, are there some output-queued policies that are more likely to maintain FCFS than others?

[8] These violations may be due to policies such as separating rooms by gender.

Table 4 Percentage of FCFS violations per type within each IW

IW \ Type	Regular	Special care	Ventilated	Total
Ward A	7.57%	7.33%	0.00%	7.37%
Ward B	3.86%	5.72%	0.00%	4.84%
Ward C	7.09%	6.62%	0.00%	6.80%
Ward D	8.18%	7.48%	2.70%	7.81%
Total within wards	6.91%	6.80%	0.67%	6.80%
Total in ED-to-IW	31%	31%	5%	

What are some other fairness criteria that one may consider? Patient severity? Patient type? What are some relevant performance measures beyond delays? Conceivably, patients may have preferences as to what IW to be assigned to. In such cases, fairness could be defined as minimizing the fraction of patients who are not assigned to their top priority (assuming consistency of patients priorities with clinical priorities). Related to this is the work of Thompson et al. (2009) that considers short term allocation of patients to medical wards during demand surges, where each patient type has their clinically ideal ward and less desirable wards. While Thompson et al. (2009) look into minimizing the *cost* that reflects the number of non-ideal ward assignments, we propose to also look at the *equity* between patients in this context. Another fairness criteria that one may consider are achieving equity in terms of blocking probability (recall that blocking arises occasionally as part of the transfer process (see Figure 26, and discussion in Section 4.3.1)) or achieving "weighted" equity in patients delay during overloaded periods. For the latter, Chan et al. (2011) has shown that one may obtain such fairness criteria using load balancing in a particular way. The main question here is how to assign the right weights depending on the patient type.

5.5.2. Staff - fairness In many large hospitals, managerial and physical considerations limit ward size, creating multiple wards that serve a similar purpose. This is also the case of the Internal wards in Rambam hospital. The multi-ward structure necessitates a routing mechanism to determine which ward should a patient be assigned to, like the Justice Table in our process. In Section 5.3, we discussed the implications of routing on delays in transfer from ED to IW. In addition, the routing mechanism has a significant effect on wards workload. High workload tends to cause personnel burnout, especially if the workload division is perceived as unjust (references to studies on the importance of perceived justice among employees can be found in Armony and Ward (2010)). Before arguing that patients' allocation to the IWs in Rambam hospital does not appear to be fair, we should first understand what is meant by "fairness".

We denote by ρ_i the average occupancy level and by γ_i - the bed turnover rate, or *flux* in ward i. When discussing the notion of fairness with IW staff, the consensus is that each nurse/doctor should have the same workload (Elkin and Rozenberg (2007)). Seemingly, this is the same as

saying that each nurse/doctor should take care of an equal number of patients, per unit of time. As the number of nurses and doctors is usually proportional to standard capacity, this criterion is equivalent to keeping bed *occupancy* level equal among the wards. However, by Little's law, $\rho = \gamma \times \text{ALOS}$, thus, if one maintains ward occupancies equal then wards with shorter ALOS will have a higher turnover rate – they will admit more patients per bed – which gives rise to additional fairness concerns. Indeed, the load of staff is not spread uniformly over a patient's stay, as treatment during the first days of hospitalization requires much more time and effort from the staff than in the following days (Elkin and Rozenberg (2007)); in addition, patient admissions and discharges consume doctors' and nurses' time and effort as well. Thus, even if occupancy among wards is kept equal, the staff workload in wards that admit more patients per bed ends up being higher. Hence, a natural alternative fairness criterion is balancing the turnover rate, or the *flux* – namely, the number of admitted patients per bed per unit of time (for example, per month), among the wards.

Rambam hospital takes fairness into consideration, as is implied from the name of the "Justice Table". But is the patient allocation to the wards indeed fair? Let us examine the proposed fairness criteria, with the help of Table 2 (in Section 4.3.2). Consider Ward B which is both the smallest and the "fastest" (shortest ALOS) out of the four wards (A-D). We observe that the average occupancy rate in this ward is high. In addition, the number of patients hospitalized per month in this ward equals about 90% of the number of patients hospitalized per month in the other wards, although its size is only about 2/3 of the others. And indeed, the flux in Ward B (6.38 patients per bed per month) is significantly higher than in the other wards. Hence, by the discussion above, the load of its staff is the highest. (In Section 4.3.2 we have discussed possible reasons for Ward B to be able to discharge patients quicker than the other wards while still maintaining high quality of care). We see that the most efficient ward, instead of being rewarded, is exposed to the highest load; hence, the patients' allocation appears to be unfair, as far as the wards are concerned.

Accounting for patient categories: Fairness is further diminished when we examine admissions separately for each patient category. As mentioned earlier, prior to routing, patients are classified into three categories: ventilated, special-care and regular. In Table 5 we see patient allocation by category. We observe that the fraction of ventilated and special-care patients allocated to Ward B (out of the total number of patients allocated to Ward B) is significantly higher than to the other wards. Load inflicted on the ward's staff by such patients is higher than by regular patients, besides, their LOS is generally longer (which makes the fact that ALOS in Ward B is the shortest even more surprising).

As a last observation, we mention that the above alleged unfairness is caused not only by the algorithm of the Justice Table itself, but also by interventions with its allocations. Indeed, for 13.2% of the patients that were routed via the Justice Table during the period 01/05/2006 -

Table 5 Justice table allocations by patient categories

IW\Type	Regular	Special-care	Ventilated	Total
Ward A	2,316 (50.3%)	2,206 (47.9%)	83 (1.8%)	4,605 (25.2%)
Ward B	1,676 (43.0%)	2,135 (54.7%)	90 (2.3%)	3,901 (21.4%)
Ward C	2,310 (49.9%)	2,232 (48.2%)	88 (1.9%)	4,630 (25.4%)
Ward D	2,737 (53.5%)	2,291 (44.8%)	89 (1.7%)	5,117 (28.0%)
Total	9,039 (49.5%)	8,864 (48.6%)	350 (1.9%)	18,253

* Data refer to period May 1, 2006 - Sep. 1, 2008 (excluding the months 1-3/07)

01/09/2008, the ward chosen originally by the program was actually changed. Furthermore, 19.2% of the patients (over the same time period) were eventually admitted to a ward, different from the one recorded in the Justice Table (*after* the changes described above). Our data show that patients assigned to IW B were less likely to end up in another IW.

Obtaining server-fairness via routing: The question arises as to how to route patients to the various wards, in a way that maintains fairness with minimal compromise on delays in the transfer process. Our data have already motivated several works in this area, which we outline here. We then discuss further research opportunities.

Analytical results for input-queued systems: The question of fair routing has been studied by Mandelbaum et al. (2012). The authors propose a closed-loop input-queued routing scheme named RMI (Randomized Most-Idle) in which a patient (customer) is routed to an IW (server pool) with a probability that is proportional to the number of idle servers (beds) in this ward out of total number of idle servers in the system. They show that, under this routing policy, pools with slower servers have on average higher occupancy and lower flux, which appears fair as faster servers are "rewarded" by having lower utilization. Indeed, the authors *prove* that RMI routing is fair (relating to balancing both occupancy levels and flux among the wards/pools as the criteria for fairness), in the many-server heavy-traffic QED asymptotic regime. The authors in Mandelbaum et al. (2012) also show that RMI asymptotically achieves the same fairness as the LISF (Longest-Idle-Server-First) algorithm proposed in Atar (2008), which is commonly used in call centers and considered fair. Weighted versions of the RMI (WMI) and LISF (LWI) are studied by Mandelbaum et al. (2012) and Ward and Armony (2013), respectively. The paper by Ward and Armony (2013) establishes that not only these simple routing algorithms achieve fairness, but they also maintain delays which are comparable to the asymptotically optimal delays which are obtained through a threshold policy identified by Armony and Ward (2010).

Simulation study - for output-queued systems: While the results in Mandelbaum et al. (2012) are valuable as a benchmark for hospital performance, the question remains as to how to control for fairness via routing in *Output-queued* systems. To address this question, the authors in

Tseytlin and Zviran (2008) perform a detailed simulation of the output-queued system under various routing schemes while accounting for *availability of information* in the system (recall discussion at Section 5.3). They study performance of the system under different routing policies, with respect to resource utilization, transfer delays, and equity among the IWs. The same data set as in the current paper is used as the basis for their simulation. The authors in Tseytlin and Zviran (2008) first consider simple routing schemes that utilize no information, such as Round-Robin algorithms, which are widely used in hospitals, including Rambam hospital (see §5.6). They analyze then more complex closed-loop policies, such as Occupancy Balancing scheme which aims at balancing pools occupancies at each moment of routing. They conclude that an algorithm that balances a weighted function of occupancy and flux achieves both fairness and short delays. Moreover, they show that this algorithm can be implemented in a partial information environment (such as hospitals) without significant compromise in terms of performance.

Research opportunities: Several research opportunities arise naturally from the discussion above. First, with respect to input-queued routing, all current research assumes that the service rates depend on the servers only. The question of how to guarantee both fairness and short delays when service rates depend on both patient types and servers remains open. Also it might be interesting to account for potential patient blocking due to lack of space. In the context of output-queued systems, a more rigorous analytical study is needed to formalize the conclusions of Tseytlin and Zviran (2008).

In the current section we discussed two possible fairness criteria: occupancy and flux, or some combination of the two. A natural question is how to rigorously and usefully combine the two criteria into a single effective workload measure; fair routing would maintain this measure equal across wards. As discussed above, Tseytlin and Zviran (2008) proposed a weighted algorithm that balanced a linear combination of occupancy and flux; it assigned similar *weights* to both criteria. The latter may be generalized (as in Section 7.2 of Tseytlin (2009)) by weighing flux and occupancy proportionately to the time it takes to process admission and discharge of patients, and routine care for patients, respectively. Specifically, consider aiming to balance the following measure across wards: $\frac{\gamma_i N_i T_i^{\gamma} + \rho_i N_i T_i^{\rho}}{n_i}$. Here γ_i is the flux, ρ_i is the occupancy level, N_i is the number of beds in the ward, T_i^{γ} is the average amount of time required from a nurse to complete one admission plus discharge, T_i^{ρ} is the average time of treatment required by a hospitalized patient per unit of time, and n_i is the number of nurses in the ward; all quantities refer to a given ward i. Considering a constant nurse-to-bed ratio, the latter translates into balancing $\gamma_i T_i^{\gamma} + \rho_i T_i^{\rho}$.

The underlying definition of operational fairness in our discussion thus far is having equal workload across medical staff. Thus, a prerequisite for solving the "fairness problem" is to be able to define and calculate workload appropriately. As we argue in Section 6.3, such calculation must

include not only direct time per resource but also considerations such as emotional and cognitive efforts, as well as other relevant factors. For example, one would argue that the mix of medical conditions and the patient severity should also be included in the workload calculation. For the latter, it is not straightforward to determine whether wards would be inclined to admit the less severe patients (and thereby be subjected to less workload, and potentially less emotional stress), or more patients with more severe conditions who would challenge the medical staff, and provide them with further learning and research opportunities.

5.6. Discussion on routing: Beyond Rambam hospital

To further investigate the issue of routing and its effect on delays and fairness, we examine the process of transfer from ED to IW in several other hospitals. The results of a survey responded to by five Israeli hospitals are summarized in Table 6.

Table 6 Hospitals comparison

	Hospital 1	Hospital 2	Hospital 3	Hospital 4	Hospital 5	Rambam
Number of IWs	9	2	3	4	6	5
Total # IW beds	327	45	108	93	210	185
Average weekly # of arrivals to Internal ED	1050	350	637	630	1050	995
Average weekly # of transfers from ED to IWs	525 (50%)	49 (14%)	266 (42%)	168 (26%)	469 (45%)	233 (23%)
Average weekly # of transfers per IW bed	1.606	1.089	2.463	1.806	2.233	1.259
IW Occupancy*	107.5%	118%	106.5%	116.4%	110%	93.8%
ED ALOS (hours)	2.2	6	2.83	6.8	2.5	4.2
IW ALOS (days)	3.9	3.9	3.5	6.1	3.5	5.2
Average delay in ED-to-IW tranfer (hours)	?	4	1	8	0.5	1.5-3
Wards differ?**	yes	yes	no	yes	no	yes
Routing Policy	cyclical order	digit of id	cyclical order	vacant bed	cyclical order***	cyclical order***

* Based on an Internet article ynet (2009).

** Differ in their capacities and ALOS.

*** Accounting for different patient types and ward capacities.

The hospitals differ in their functionality and geographical location. In Table 6 we see that they also differ in *size* (number of IWs and beds in them; number of treated patients), in the *load* IWs

are subjected to (average number of transfers to IWs per bed; average occupancy rate and by ALOS in the ED and in the IWs. Despite these differences, we observe similar issues in the transfer from ED-to-IW process in all hospitals. First, all these hospitals have an extremely high occupancy level. Second, all of them route patient to multiple IWs that provide similar clinical services. In particular, the issue of fairness appears to be relevant. Finally, in all hospitals except for Hospital 5, delays in ED-to-IW transfer are significant. Indeed, the situation in Rambam hospital is not the worst. Note that none of the hospitals, with the exception of Hospital 5, measures these delays - the numbers provided to us were rough estimates.

The routing policies used in the hospitals are intuitive and simple - open-loop cyclical-order and randomized policies are prevalent. Although in four out of the six hospitals the wards are heterogeneous - that is, they differ in their capacities and ALOS - those differences play no role in routing policies; no hospital accounts for differences in ALOS and only Rambam hospital accounts for ward capacities. In addition, Hospitals 1–4 do not take patient category into account in the routing decisions. Surely this cannot be fair, as load inflicted on the ward staff by patients of different categories varies significantly. We next provide more details on the routing schemes used by some of these hospitals.

In Hospital 2 routing is performed according to the second-to-last digit of a patient identity number: if it is odd then the patient is assigned to Ward A, if even - to Ward B. This is analogous to an open-loop random assignment: each ward is chosen with probability 1/2. But ward capacities differ: the size of one ward is 2/3 of the other's. This scheme must be suboptimal with respect to delay minimization and fairness.

Closed-loop schemes introduce their own set of problems. For example, in Hospital 4 patients are sent, on a FCFS basis, to an IW that has a vacant bed (we were told that the wards were *always* full). Indeed, the load of the IWs in this hospital (average occupancy and flux) is very high. Under this scheme, the wards have complete control on when to admit their patients. In particular, unless the right incentive mechanism is put in place, these wards may prolong patients LOS in order to reduce their workload. Indeed, waiting times and ALOS in the ED and in the IWs are *by far* the longest in this hospital.

Hospital 5 has a remarkable performance - patients to be hospitalized in the IWs are transferred from the ED with almost no delay, and fairness is maintained due to cyclical routing to wards of similar size and ALOS, with separate cycles per patients category. However, from a conversation with the ED receptionist, we learnt that, even in this hospital, perception of fairness towards the wards is negative. Specifically, the routing process is managed manually by the receptionist, who receives frequent complaints from the wards' staff about unfairness of the allocations.

Operational implications: Our inter-hospital study raises several interesting operational issues.

"Push" vs. "Pull" routing scheme: An important aspect that comes up in our discussion is the impact of incentives and control ownership over transfer times on delays. The incentive issue was discussed above in the context of Hospital 4; the scheme used in this hospital also raises the issue of who is in charge of determining the transfer assignment and timing. In Hospital 4, the wards effectively decide on the patient routing and the timing of the transfer. We refer to this scheme as a "pull" scheme, as the wards "pull" patients out of the ED. An alternative scheme is a "push" scheme, in which the ED determines the patient assignment and transfer time. Table 6 suggests that the push scheme results in significantly shorter transfer delays. This is also supported by the historical account of how routing was done in Rambam hospital. When the hospital moved from a pull system to a push system transfer delays shrunk from an average of 10.5 hours (with 12% of the transfer patients forced to wait more than 24 hours (!)) to only a few hours (see details in Tseytlin (2009)).

ED operational architecture in the hospitals: From Table 6 we can also learn how different ED policies affect hospitalization processes (continuing the discussion on ED operational architectures in Sections 3.3 and 5.2). Consider Hospitals 1, 3 and 5 versus Hospitals 2 and 4. The hospitals in the first group have ED LOS significantly shorter than the hospitals in the second group (2-3 hours vs. 6-7 hours), and percentage of hospitalizations in IWs - significantly larger. This suggests *ER structure* for the hospitals in the first group versus *ED structure* for the hospitals in the second group. Namely, in the Emergency Departments of Hospitals 1, 3 and 5 patients go through triage and are then transferred to the wards for treatment; in the EDs of Hospitals 2 and 4 patients receive more extensive medical treatment. In this analysis Rambam hospital falls somewhere in the middle.

We also observe that IWs occupancy in Hospitals 1, 3 and 5 is lower than in Hospitals 2 and 4; and transfer delays - shorter. This is seemingly in contradiction with the characterization of ER versus ED architecture. To wit, one would expect hospitals with a higher fraction of IW hospitalization to have higher IW occupancy and longer delays. Even more surprising is the IW ALOS in Hospital 4, which is much longer than IW ALOS in the other hospitals. Since in this hospital, the patients obtain some treatment in the ED, one would expect the ALOS to be shorter. To ascertain the exact causes of these differences, one would need more data across the hospitals which, from our experience, is not easy to get. One plausible explanation of these numbers may have to do with the fact that a hospital with a triage-focused ED is more likely to invest in speeding up the ED-to-IW transfer process (e.g. by providing sufficient IW capacity, or by an efficient IW patient discharge

procedures), so that patients can be cared for promptly. Other explanations may include patients mix, push versus pull, and incentive schemes, as discussed earlier.

Note: the discussion presented here is necessarily superficial, as it is based on questionnaires and interviews solely - no observations or data collection were performed in those hospitals. Collecting detailed patient-level data from multiple hospitals can be beneficial to our understanding of the various findings described above. However, based on the time and effort involved in collecting the detailed data from Rambam hospital, we suspect that collecting such data from multiple hospitals would be very difficult.

Research opportunities: The topic of server incentive in queueing systems has received some attention in the queueing literature, when service has monetary compensation and servers choose their service effort level manifested through their service rate (e.g. Gilbert and Weng (1998), Cachon and Zhang (2007)). In a hospital setting, this approach is typically inappropriate. In Tseytlin (2009) two game-theoretic approaches are explored to study this issue. One is a non-cooperative game-theoretic approach, in which the players (the wards) select their discharge rate in order to achieve a target workload level. The other is a cooperative game theoretic approach which computes the Shapley value of the game of splitting delay costs among the internal wards. The latter approach was also explored by Anily and Haviv (2009), where it is shown that the core is non-empty in the cooperative pooling game.

The aspect of *push versus pull* schemes is tightly related to the incentive structure and provides interesting modeling challenges. Generally, in queueing models, the service start of a customer waiting in a particular queue is determined according to a central queueing discipline (FCFS, priorities, idling versus non-idling, etc.). However, as is evidenced by our observations, system performance is significantly dependent upon whether the ED "pushes" patients to the IWs or the IWs "pull" patients from the ED. How should one model this effect? How does one provide the right incentive mechanism to make either scheme perform well?

One can also provide incentives to the wards to increase their efficiency and shorten transfer delays by creating a centralized system of *rewards and fines*, managed by hospital administration. Possible ways to "reward" a ward for its efficiency (as pointed out to us by one of the IWs' head physicians) could be increasing the ward's staff numbers, allowing it to perform more "skips" during the assignment process, increasing the ward's budget, buying newer equipment, and more. Naturally, the feasibility and impact of such reward/fine system should be further investigated.

To conclude, this section has dealt with the issue of routing patients from the ED to the IW; Seemingly, this is analogous to other routing problems that have been studied extensively in the queueing literature. However, our study has revealed some unusual features that raise new challenges and provide an avenue for new research opportunities.

6. A broader view

We have analyzed several-years of data from a tertiary hospital, concentrating on its main patient flows: within the ED and IWs, and the transfers from ED-to-IWs. Our analysis was further based on traditional time and motion studies, as well as on interviews with hospital personnel and management when called for. The analysis revealed operationally-significant phenomena, such as those related to patient arrivals, departures, waiting times and LOS. These, in turn, offer research opportunities for the Operations Research community at large, and for queueing scientists in particular.

In this conclusion to our paper, we would like to highlight some common themes. Most have already been discussed or touched upon, but the viewpoint here is somewhat broader.

6.1. The effect of patient flow on overall hospital performance

The mission statement of Mayo Clinic (Mayo Clinics (2011)) reads: "Mayo will provide the best care to every patient every day through integrated clinical practice, education, and research." Johns Hopkins Hospital (Johns Hopkins Hospital (2011)) strives "To improve the health of the community and the world by setting the standard of excellence in patient care." These two mission statements are representative in that patient-flow performance is left out of them, and rightly so: hospital systems are here to deliver *health care*, and it is our responsibility, as part of the OR/OM community, to relate *patient care* to *patient flow*, the latter being the focus of our study.

The measurement of hospital performance is a deep non-consensual subject (Shaw (2003), Shih and Schoenbaum (2007), McKee et al. (1997)). Simplifying it for present purposes, we relate to hospital performance along the following dimensions: *clinical, financial, societal, psychological and operational*. Each such dimension has attributes, some general and some perceived as specific to the dimension. Examples of general attributes are *quality* and *risk*; examples of specific attributes are measures associated with patient flow, such as *congestion* and *accessibility* measures, which have been traditionally associated with the operational dimension. But an emerging theme, converging from various disciplines, is that the significance of patient flow spans beyond operations - and this has also been our contention, while still acknowledging that efficient and effective patient flow is not the goal, but rather a prerequisite means for proper care. In support of our claim, we point to the major effects that patient flow has on the other dimensions mentioned above.

- *Clinical*: There has been a growing literature that relates quality of care with patient flow. For example, Chalfin et al. (2007) argue that overload and long waits in transferring patients to ICUs contribute to higher mortality rates; and Kc and Terwiesch (2009) demonstrated that patients, in overloaded ICUs, are likely to be discharged too early, in which case they tend to return, more frequently, for additional treatment (in a more severe state).

The clinical-operational relation also triggered the JCAHO Standard LD.3.10.10. (JCAHO (2004)). Indeed, in addition to the news-catching problems of ED congestion and equitable accessibility to healthcare, the JCAHO standard is no less concerned with the fact that "Problems with patient flow can lead to sentinel events due to delays in treatment." (JCAHO's definition of a sentinel event is "an unexpected occurrence involving death or serious physical or psychological injury, or the risk thereof. ... Such events are called "sentinel" because they signal the need for immediate investigation and response." JACHO (2011)).

- *Financial*: The Financial-Operational connection is widely acknowledged. For example, patient LOS and idleness of resources have long been the target of aggressive reduction efforts in hospitals, mainly to reduce hospitalization costs and increase efficiency. Significantly, though, there is evidence against using LOS as the main cost-proxy: Taheri et al. (2000) argue that, for most patients, LOS reduction has minimal financial impact as most costs are incurred during the *early stages* of hospitalization. Hence cost-reduction efforts must "deemphasize LOS and focus instead on process changes that better use capacity and alter care delivery."

At a higher level, healthcare plays a key role in world economies: for example, it accounts for 6-12% of GDP in member states of the Organization for Economic Cooperation and Development (OECD), with the U.S. being an outlier at 17% (OECD (2011)). This latter figure is considered a "cost crisis" that is begging for operational remedies. Kaplan and Porter (2011) identify process-flow maps as a major prerequisite for properly understanding how much it costs to deliver patient care. Specifically, the analysis of patient flow is the first step in Time-Driven-Activity-Based-Costing (TDABC), which is their proposed remedy for the cost crisis in U.S. heathcare. (Note that both Kaplan and Porter are accepted as world-class leaders in their domain of expertise - Kaplan in Accounting and Porter in Strategy - their support of our premise is thus not to be taken lightly.)

- *Societal*: Article 35 of the "Fundamental Rights of the European Union" states that "Everyone has the right of access to preventive health care and the right to benefit from medical treatment under the conditions established by national laws and practices. A high level of human health protection shall be ensured in the definition and implementation of all Union policies and activities" (EU (2000)). Similar sentiments are expressed by the above-quoted and other hospital mission statements. Thus, healthcare is considered a national resource that is to be equitably shared by everyone (Goddard and Smith (2001), Serban (2011), Oliver and Mossialos (2004)).

No wonder, then, that problems in patient flow, for example excessive ED congestion that lead to ambulance diversion, are often taken to be violations of a fundamental citizen's right, which brings flow problems to the forefront of public attention. This attention is further amplified via increased competition of hospitals over patients (Kessler and McClellan (2005)), and the natural association of poor service levels with high levels of congestion. In other words, poor management

of patient flow in a hospital leads to low levels of patient satisfaction, hence negative perception of this hospital which results in low demand for its services.

- *Psychological*: Congestion and waiting are typically associated with poor service levels. These cause negative customer emotions and consequent low levels of satisfaction (Rafaeli et al. (2002)). Major reasons are the uncertainly and helplessness that too often go hand in hand with waiting for service (Maister (1985), Norman (2009)). The above chain reaction is further exacerbated during hospital waiting where negative emotions could run especially high, in view of what is at stake and the special circumstances (helplessness levels) that customers and relatives are finding themselves in. Environmental pressures affect negatively also medical and support staff, for example job satisfaction (Aiken et al. (2002)), and the latter has been found positively correlated with customer dissatisfaction (Schneider and Barbera (2011)). In summary, patient flow naturally affects the psychological state of both the patients and the staff that experience it.

It follows from the above discussion that patient flow, or rather its operational performance measures, can be naturally associated with clinical, financial, societal (e.g. accessibility) and psychological (e.g satisfaction) performance. We next discuss an important implication of this association.

6.2. Operational measures as surrogates to overall hospital performance performance

We have covered a wide spectrum of operational performance measures that are associated with patient flow. Starting with some directly related to *delayed medical treatment*, we mention the time until a patient is first seen by an ED or IW physician, the fraction of patients who leave without being seen (LWBS) and various measures of ambulance diversion. These measures have analogs elsewhere in a hospital, for example waiting time until an MRI test or being admitted to surgery, and the fraction of patients who end up being hospitalized in a ward that differs from that which is medically best for them. Then we touched on measures associated with unplanned *feedback* flows, for example the fraction of unplanned readmissions to IWs within, say, 3 months, which is analogues to re-admission proportions to ICUs. And, finally, there are measures related to *LOS*, in the ED or IWs, such as merely averages (or medians), or fractions staying beyond a desired threshold (for example, a national "4 hours ED target" in the UK, with similar goals elsewhere.)

Interestingly, Jones and Schimanski (2010) suggest that this "4 hours ED target" is related more to patients' satisfaction than their clinical state. Regardless, this is an example that suggests an important useful principle: when compared with other dimensions, operational performance is easier to quantify, measure, track online and react upon.

We thus envision that operational measures, through their high correlation with other measures and their relative accessibility, serve as surrogates for the *full* spectrum of hospital performance.

This, we believe, opens up a wide uncharted territory for future research, stemming from *real-time status-tracking of individual patients*, as they flow through hospital resources; here "status" is broadly construed as all the above-mentioned performance dimensions, mostly clinical, financial, psychological and operational. (Note, however, that real-time patient-tracking, considered technologically feasible, is still beyond the reach of all but scarcely few hospital measurement systems.)

6.3. Workload

Operational performance of a service-system is determined by the gap, positive or negative, between its workload and the capacity assigned to process it. Both concepts have been discussed ample times throughout our paper, yet they are central and subtle enough to deserve further scrutiny: Workload will be discussed here and Capacity will be the subject of the next subsection.

Workload is associated with a resource and, as such, used to gauge the requirements from that resource: for example, workload of nurses helps determine appropriate nurse staffing levels. Workload has been also shown to affect staff satisfaction level (Aiken et al. (2002)) and patients quality of care (Kc and Terwiesch (2009)). Our use of the term workload has been intentionally imprecise. This allowed us, for example, to conveniently discuss how workload in the IWs is related to workload in the ED (§2.4&5.4), how workload affects LOS in ED (§3.1), and how workload relates to fair routing from the ED to IWs (§5.5). For our purpose now, it helps to turn more concrete at two levels: first, define the *workload* of a *resource*, at a given *time*, to be the amount of work, measured in units of resource-time, that is required from that resource at that time; and second, we restrict attention to workload of a single *nurse*, which captures well the issues that we are seeking to bring up.

In Queueing models, workload is typically an *average* quantity that is defined in steady-state: if λ is the arrival-rate of patients and S is the service time required from a nurse by an arrival, then $R = \lambda \times E[S]$ is the workload of the nurse, which is commonly referred to as *offered load*. Note that this definition of R is consistent with the definition of workload above. Moreover, R can be interpreted as a form of Little's formula, via which it is more natural to conceptualize the workload as an average nominal number of required nurses (e.g. nurse-hours required per hour). This "Little interpretation" reveals the way in which the offered-load $R(\cdot)$ is to be defined in *time-varying* environments (Green et al. (2007a), Reich (2011)), such as hospitals: simply apply the time-varying Little's formula (Bertsimas and Mourtzinou (1997), Green et al. (2007a))

$$R(t) = \int_0^t \lambda(u) P\{S > t-u\} du, \quad t \geq 0,$$

in which $\lambda(\cdot)$ is the arrival rate, as a function of time (and under the mild assumptions required for the integral to make sense).

The above definitions in fact disguise some subtleties. For example, due to interaction among resources, the resource associated with a workload need not be uniquely determined. For example, consider ED physicians that attend to patients after a nurse's triage; assume also that if the physicians are overloaded (e.g. their queue is long) then the nurses help out by taking over some of the physician tasks. It follows that a nurse's workload depends on the physicians' queue: the longer the queue the higher is the workload; and this queue-length depends on the physicians' workload which, in turn, is fed by the nurses.

As workload is matched against capacity, it must be measured in operational units. However, the workload of a nurse is affected by various factors beyond the mere time-content of nurse tasks. For example, 1-minute of a standard chore does not compare with a 1-minute life-saving challenge; and attending to a pregnant woman with her fetus in stress could be demanding both cognitively (two patients in parallel) and emotionally. The calculation of nurses workload must therefore accommodate operational, emotional and cognitive factors, yet the outcome must be in standardized units that are "translatable" into staffing levels. This is attempted in an ongoing research Plonski et al. (2013), already hinted at in Section 4.2.1: here one aims at balancing the "sum of operational and emotional" workload between two maternity wards, one in charge of complications prior to birth and the other after birth; a sought-out allocation of normal births will balance the total workload between the two wards.

It follows from the above that a comprehensive definition of a nurse workload is inevitably complex, which raises a need for its practical approximation. In the context of a medical ward, a natural one is relative *occupancy*, namely the number of hospitalized patients in a ward, divided by the number of its beds. Assuming constant beds-to-nurse ratio (Jennings and de Véricourt (2011)), homogeneous patient mix and even distribution of workload over a patient stay, occupancy is then well correlated with the *daily-routine* workload; it captures, in particular, fixed activities (e.g. feeding or bathing a patient), which constitute a significant fraction of a nurse workload. However, as already discussed in Section 5.5.2, the level of medical attention that patients require declines during their stay. In addition, routine chores are less work intensive than *admissions and releases* of patients, and the latter are naturally associated with patients *turnover* or *flux*, as opposed to bed occupancy. A proxy for workload must, therefore, acknowledge both occupancy and turnover (§ 5.5.2), and possibly be also sensitive to the "age" of hospitalized patients.

6.4. Capacity

Capacity of a hospital or a ward is commonly expressed in terms of number of beds (or rooms, or physical space). But it is also necessary to associate with a ward its *processing capacity*, which is determined by its human and equipment resources: nurses, physicians, support personnel and

medical apparatus. One thus distinguishes between *static* capacity (e.g. beds) and *dynamic* (processing) capacity of a resource. This distinction has operational and financial implications in that static capacity is thought of as *fixed* over the relevant horizon, hence its cost is fixed; processing capacity, on the other hand, is considered *variable* in that it is *flexible (controllable)*, both level- and hence cost-wise. For some purposes, as in Section 3.2, it is convenient to associate the processing capacity of a ward, expressed in terms of, say, average number of patients per day, with the processing capacity (or maximal turnover rate) of its beds. This greatly simplifies modeling by conceptualizing beds as "servers" in a queueing system.

The association of flexible capacity with variable costs plays an important role in the Accounting/Strategic view of a hospital. For example, the above mentioned Kaplan and Porter (2011) argues that most hospital costs are mistakenly judged as fixed while they ought to be viewed as variable costs, which means that the corresponding resource levels are flexible. This is an important observation, with which we modestly concur, as it renders controllable most resources in a hospital - in other words, these resources are able to actively participate in flow control or performance improvement efforts.

A practical proxy for the processing capacity of a ward is the product of its number of beds divided by its Average LOS. This suffices for a black-box model of a ward, as in Sections 3.2 and 5.3. But for more refined purposes, such as balancing load across Maternity Wards (§4.2.1), one must start with mapping the activities of individual nurses and proceed bottom up until establishing overall ward capacity. This calls for data-based research on the *anatomy of capacity* in hospitals, specifically how it is determined by its constituents (clinical personnel, support staff and equipment).

A natural theoretical environment for capacity research is provided by queueing models. To start, such models facilitate the identification of static (fixed) vs. processing (variable) capacity. They also clarify the interrelations between capacity and offered load and hence with flow control, staffing and resource allocation. Consider, for concreteness and simplicity, an $M/M/B/N$ queue in steady-state, with arrival-rate λ and service-rate-per-server μ. Such a queue can serve as a model of an IW with B beds, maximal turnover-rate-per-bed μ, and physical capacity of N patients ($N-B$ could be waiting, say in the ED or another IW, for hospitalization). This IW is thought to operate over a part of the day during which arrival rates do not vary significantly, and under an admission policy that prohibits waiting (blocks arrivals) when the number of its patients reaches N. For this queueing model of an IW, its fixed capacity is N, processing capacity is $B \times \mu$, and the offered-load is $R = \lambda/\mu$. Then the relation between R and B determines whether the IW is under-, over- or critically-loaded; and each of these three operational regimes enjoys its distinct operational characteristics, possibly after further refinements due to the relation between N and

B. It is worthwhile noting that, alternatively, one could justify either $M/M/B/B$ or $M/M/B$ as models of an ED or an IW, each under the appropriate circumstances and protocols (de Bruin et al. (2009), Green (2004)).

Finally, a queueing view of patient-flow raises additional capacity-related issues that are worthy of research. Consider, for example, determining the capacity of the Ophthalmology ward in Rambam hospital, which happens to admit overflow patients from Internal wards. This entails allocation of appropriate equipment, training of Ophthalmology ward personnel to be able to cater to IW patients and developing protocols for overflow of IW patients to Ophthalmology. It is much the same as cross-training multi-skilled agents in a call center and then designing skills-based routing (Aksin et al. (2007a)).

6.5. Fairness and incentives

Fairness is an attribute that arises across many facets of a hospital. One could associate it with fair balancing of workload across hospital personnel, as in routing patients from the ED to IWs (§ 5.5) or pregnant woman to maternity wards (Plonski et al. (2013)); or with fair waiting of patients, for example when comparing input- vs. output-queue architectures (§5.3); or with fair access to hospital resources (e.g. waiting for admission (§5.2) in a way that matches one's clinical priority).

Healthcare is, in fact, an area where deep fairness issues naturally arise, but which are beyond the scope of patient flow, for example those related to fair allocation of public resources (medication) or ethics. Fairness can be hurt or promoted, depending on the corresponding incentive systems. In short, fairness in healthcare, in particular its interaction with operations and patient flow, is an area that calls for empirically-based research. (Section 2.2 in Mandelbaum et al. (2012) offers a short discussion on fairness and relevant references.)

6.6. Time-scales

When analyzing ED-to-IWs flow (§ 5), the wards operate naturally on a time-scale of days while the ED time scale is hours. It follows that the wards serve as a random environment for the ED (Ramakrishnan et al. (2005)). Figure 19 (§ 4.1) reveals that the hourly scale is also of interest for IWs. In fact, as is indicated in Figure 33, the IW patient-count fluctuates significantly, and on an hourly basis, during a typical *LOS excursion* of an IW patient - about one week from admission until release.

The empirical examples above arise from a service system that evolves in multiple time scales, which are all natural for measuring and modeling its performance. The mathematical manifestation of such scales is asymptotic analysis that highlights what matters at that scale, while averaging out details that are deemed insignificant. Consider, for example, Mandelbaum et al. (2012): ED-to-IWs routing gives rise there to 3 counting scales (bed, room, sub-ward), each with its corresponding time

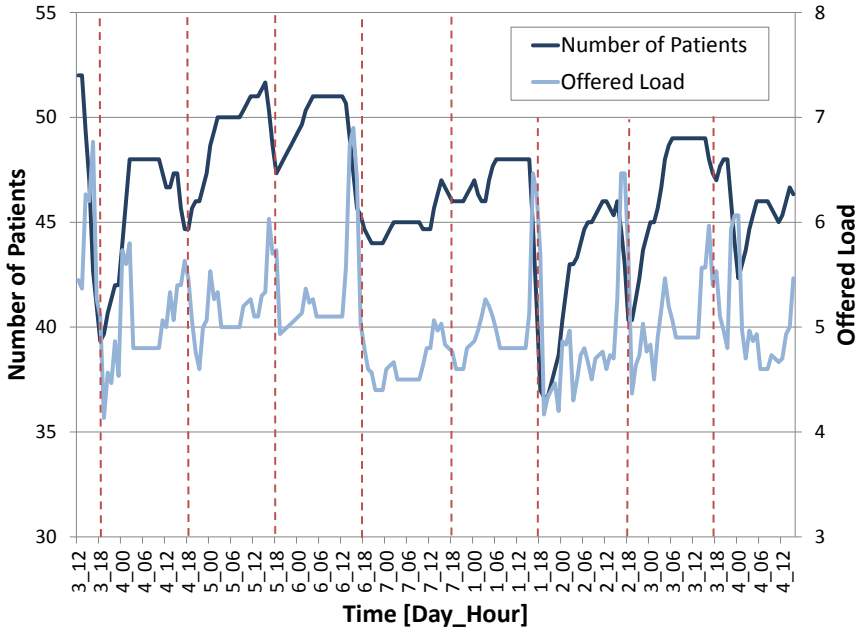


Figure 33 **Changes in Number of Patients During an IW LOS (Excursion)**

scale (hour, day and week, respectively). Then each of the three pairs of (counting, time) scales is associated with a certain stochastic process, with all three processes modeling the (same) number of *idle* IW beds, measured (scaled and centered) appropriately: a queueing process (counting individual beds as they evolve in hours), sub-diffusion (rooms, days) and diffusion (wards, weeks).

More specifically, fluctuations in individual bed-count are naturally described in the *finest* time scale of hours, in the sense that say minute- or daily-fluctuations would obscure useful details: the former (minutes) is too short for observing noticeable changes in bed count, and the latter (days) smoothes out these changes. Similarly, the coarsest scale is (wards, weeks): it measure the total number of available beds in a hospital (in 10's of beds), which fluctuates over a week (corresponding to a "typical" LOS). The intermediate sub-diffusion scale is more subtle, and readers are referred to Section 5.1 of the above paper for its motivation and application in balancing bed-idleness across IWs. (Note that Mandelbaum et al. (2012) carries out only steady-state asymptotics, but it falls short of establishing functional limit theorems, at the process level, which is the "language" used above.)

The analysis, theoretical as well as empirical, of hospital units that operate under multiple time-scales offers significant research opportunities. Hierarchical control of a hospital (e.g. strategic, tactical and operational) is one such example, in which the strategic level generates constraints for tactical decisions which, in turn, percolate down to the operational level (e.g. Zeltyn et al. (2011)).

6.7. System view - beyond Rambam hospital

The ED+IW network is relatively isolated from the rest of the hospital (§1.2). Furthermore, we advocated a system-view of this network since its three components (ED, IWs, transfers) are strongly interdependent (§2.4). Our contention now, data permitting, is that it would have been beneficial to expand the horizon beyond hospital limits. By this we mean that processes and policies of external healthcare organizations could exert significant influence on patient flow within our hospital.

As an example, consider the interaction of hospitals with surrounding nursing homes. These are both feeding and receiving hospital patients, at rates that could vary with the season: specifically, during flu seasons, it is conceivable that the ED+IW network, integrated with the relevant nursing homes, jointly operate as an almost *closed* queueing sub-network. It follows that any interruption in the operations of these nursing homes would immediately impact the ED+IW network. And this actually happened in Rambam hospital, when the largest Israeli HMO refused to cover patients in two relevant major nursing homes, due to financial disagreements. The result was blocked IWs, which rippled to block the ED, thus causing severe disruptions in patient flow within and outside: for example, increased ambulance diversion and hence overloading of other hospitals in the city and elsewhere. The above example would be interesting to analyze, empirically as well as theoretically, but its data needs are beyond our scope.

6.8. Some concluding words on data-based (evidence-based) research

Evidence based medicine is the standard in *patient care*. Our premise is that a similar approach is to be taken with respect to the analysis of hospital *operations*. (For further support, we mention the book on "Measuring Health Care for Operational, Financial and Clinical Improvements" Dlugacz (2006), and recall the HBR article on "The Cost Crisis in Health Care" by Kaplan and Porter (2011)). More specifically, we have proposed a model, perhaps a framework, for evidence-based operational analysis of *patient flow in hospitals*. Our research has been based on Exploratory Data Analysis (EDA), performed from the view-point of a Queueing Scientist, and hopefully creating further inviting and significant research opportunities.

Acknowledgements

We would like to dedicate our paper to the memory of the late David Sinreich. As a professor at Technion IE&M, David was a pioneer and an example-to-follow, in acknowledging the necessity and importance of data- and simulation-based research in healthcare. In particular, the ED data collection that he orchestrated in 8 Israeli hospitals served as a proof-of-feasibility for our EDA.

Our research, and its seed financial backing, started within the OCR (Open Collaborative Research) Project, funded by IBM and lead jointly by IBM Haifa Research (headed by Oded Cohn), Rambam Hospital (Director Rafi Beyar) and Technion IE&M Faculty (Dean Boaz Golany).

Data analysis, maintenance of data repositories, and ample advice and support, have been cheerfully provided by the Technion SEE Laboratory: Valery Trofimov, Igor Gavako, Ella Nadjharov, Shimrit Maman, and Katya Kutsy.

The authors, and the research presented here, owe a great deal to many additional individuals and institutions. This long list, which we can here acknowledge only in part, clearly must start with our host Rambam hospital - its capable management, dedicated nursing staff and physicians, and especially: Rafi Beyar, Rambam Director and CEO, who invited us at the outset to "use" the healthcare campus as our research laboratory; we gladly accepted the invitation, and have been since accompanied and advised by Zaher Azzam (Head of Rambam Internal Ward B), Sara Tzafrir (Head of Rambam Information Systems), and the OCR steering committee: Hana Adami (Head of Rambam Nursing), Yaron Barel (Rambam VP for Operations), Boaz Carmeli (IBM Research), Fuad Basis (Rambam ED), Danny Gopher (Technion IE&M), Avi Shtub (Technion IE&M), Segev Wasserkrug (IBM Research), Amir Weiman (Head of Rambam Accounting) and Pnina Vortman (IBM Research).

We are thankful to Technion students Kosta Elkin, Noga Rozenberg and Asaf Zviran: their projects and cooperation helped us to analyze the ED-to-IWs process.

Financially, the research of YM, GY and YT was supported by graduate fellowships from Technion's Graduate School and the Israel National Institute for Health Services and Health Policy Research. AM's research was supported by ISF (Israeli Science Foundation) Grant 1357/08-11, and Technion funds for the promotion of research and sponsored research. The joint research of MA and AM was funded by BSF (Binational Science Foundation) Grants 2005175/2008480.

References

Adler, P.S., A. Mandelbaum, V. Nguyen, E. Schwerer. 1995. From project to process management: An empirically-based framework for analyzing product development time. *Management Science* **41** 458–484.

Aiken, L.H., S.P. Clarke, D.M. Sloane, J. Sochalski, J.H. Silber. 2002. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA* **288** 1987–1993.

Aksin, O.Z., F. Karaesmen, E.L. Ormeci. 2007a. A review of workforce cross-training in call centers from an operations management perspective. D. Nembhard, ed., *Workforce Cross Training Handbook*. CRC Press.

Aksin, Z., M. Armony, V.M. Mehrotra. 2007b. The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 655–688.

Anily, S., M. Haviv. 2009. Cooperation in service systems. *Operations Research* **33** 899–909.

Armony, M. 2005. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* **51**(3-4) 287–329.

Armony, M., S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, G. Yom-tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. Forthcoming in Stochastic Systems.

Armony, M., A. Ward. 2010. Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research* **58**(3) 624–637.

Atar, R. 2008. Central limit theorem for a many-server queue with random service rates. *Annals of Applied Probability* **18** 1548–1568.

Atar, R., A. Shwartz. 2008. Efficient routing in heavy traffic under partial sampling of service times. *Mathematics of Operations Research* **33** 899 – 909.

Bekker, R., A.M. de Bruin. 2010. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research* **178** 45–65.

Bertsimas, Dimitris, Georgia Mourtzinou. 1997. Transient laws of non-stationary queueing systems and their applications. *Queueing Systems* **25** 115–155.

Brandeau, M.L., F. Sainfort, W.P. Pierskalla, eds. 2004. *Operation Research and Health Care: A Handbook of Methods and Applications*. Kluwer Academic Publishers, London.

Brillinger, D. 2002. John wilder tukey (1915-2000). *Notices of the American Mathematical Society* 193–201.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* **100** 36–50.

Cachon, G., F. Zhang. 2007. Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Science* **53**(3) 408–420.

Camerer, C.F, G. Loewenstein, M. Rabin, eds. 2003. *Advances in Behavioral Economics*. Princeton University Press.

Chalfin, D. B., S. Trzeciak, A. Likourezos, B. M. Baumann, R. P. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35** 1477–1483.

Chan, C., M. Armony, N. Bambos. 2011. Maximum weight matching with hysteresis in overloaded queues with setups. Working paper.

Chan, C., G.B. Yom-Tov, G. Escobar. 2014. When to use speedup: An examination of service systems with returns. *Operations Research* **62**(2) 462–482.

Chen, C., Z. Jia, P. Varaiya. 2001. Causes and cures of highway congestion. *Control Systems, IEEE* **21**(4) 26–33.

Chen, H., J.M. Harrison, A. Mandelbaum, A. van Ackere, L. Wein. 1988. Empirical evaluation of a queueing network model for semiconductor wafer fabrication. *Operations Research* **36** 202–216.

Cooper, A.B., E. Litvak, M.C. Long, M.L. McManus. 2001. Emergency department diversion: Causes and solutions. *Academic Emergency Medicine* **8** 1108–1110.

de Bruin, A.M., R. Bekker, L. van Zanten, G.M. Koole. 2009. Dimensioning hospital wards using the erlang loss model. *Annals of Operations Research* **178** 23–43.

de Bruin, A.M., A.C. van Rossum, M.C. Visser, G.M. Koole. 2007. Modeling the emergency cardiac in-patient flow: An application of queuing theory. *Health Care Management Science* **10**(2) 125–137.

Dlugacz, Y.D. 2006. *Measuring Health Care: Using Quality Data for Operational, Financial, and Clinical Improvements*. Jossey-Bass, A Wiley Imprint.

Edie, L.C. 1954. Traffic delays at toll booths. *Operations Research* **2**(2) 107–138. doi:10.1287/opre.2.2.107. URL http://or.journal.informs.org/cgi/content/abstract/2/2/107.

Elkin, K., N. Rozenberg. 2007. Patients flow from the emergency department to the internal wards. IE&M project, Technion (In Hebrew).

EU. 2000. Charter of fundamental rights of the european union. Http://www.europarl.europa.eu/charter/pdf/text_en.pdf.

Feldman, Z., A. Mandelbaum, W.A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54**(2) 324–338.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review and research prospects. *Manufactoring, Services and Operations Management* **5** 79–141.

Gerla, M., L. Kleinrock. 1980. Flow control: A comparative survey. *IEEE Transactions on Communcations* **28**(4) 553–574.

Gilbert, S. M., Z. K. Weng. 1998. Incentive effects favor nonconsolidating queues in a service system: The principle-agent perspective. *Management Science* **44**(12) 1662–1669.

Goddard, Maria, Peter Smith. 2001. Equity of access to health care services: Theory and evidence from the uk. *Social Science & Medicine* **53**(9) 1149–1162.

Green, L. 2004. Capacity planning and management in hospitals. M.L. Brandeau, F. Sainfort, W.P. Pierskalla, eds., *Operation Research and Health Care: A Handbook of Methods and Applications*. Kluwer Academic Publishers, London, 14–41.

Green, L., J. Soares, J.F. Giglio, R.A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68.

Green, L., N. Yankovic. 2011. Identifying good nursing levels: A queuing approach. *Operations Research* **59**(4) 942–955.

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

Green, L.V., P.J. Kolesar, W. Whitt. 2007a. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16**(1) 13–39.

Green, L.V., S. Savin, M. Murray. 2007b. Providing timely access to care: What is the right patient panel size? *The Joint Commission Journal on Quality and Patient Safety* **33**(4) 211–218.

Gurvich, I., W. Whitt. 2010. Service-level differentiation in many-server service systems via queue-ratio routing. *Operations research* **58**(2) 316–328.

Hagtvedt, R., M. Ferguson, P. Griffin, G.T. Jones, P. Keskinocak. 2009. Cooperative strategies to reduce ambulance diversion. *Proceedings of the 2009 Winter Simulation Conference* **266**(8) 1085–1090.

Hall, R., D. Belson, P. Murali, M. Dessouky. 2006. Modeling patient flows through the healthcare system. R.W. Hall, ed., *Patient Flow: Reducing Delay in Healthcare Delivery*, chap. 1. Springer, 1–45.

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

Herman, Robert. 1992. Technology, human interaction, and complexity: Reflections on vehicular traffic science. *Operations Research* **40**(2) 199–211.

Huang, J., B. Carmeli, A. Mandelbaum. 2015. Control of patient flow in emergency departments: Multiclass queues with feedback and deadlines. Forthcoming in Operations Research.

JACHO. 2011. Sentinel event. Http://www.jointcommission.org/sentinel_event.aspx.

JCAHO. 2004. JCAHO requirement: New leadership standard on managing patient flow for hospitals. *Joint Commission Perspectives* **24**(2) 13–14.

Jennings, O.B., F. de Véricourt. 2008. Dimensioning large-scale membership services. *Operations Research* **56**(1) 173–187.

Jennings, O.B., F. de Véricourt. 2011. Nurse staffing in medical units: A queueing perspective. *Operations Research* **59**(6) 1320–1331.

Johns Hopkins Hospital. 2011. Mission statement of johns hopkins hospital. Http://www.hopkinsmedicine.org/the_johns_hopkins_hospital/mission.html.

Jones, P., K. Schimanski. 2010. The four hour target to reduce emergency department waiting time: A systematic review of clinical outcomes. *Emergency Medicine Australasia* **22** 391–398.

Jouini, O., Y. Dallery, O.Z. Aksin. 2009. Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics* **120**(2) 389–399.

Kaplan, R.S., M.E. Porter. 2011. How to solve the cost crisis in health care. *Harvard Business Review* **89**(9) 46–64.

Kc, D., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55** 14861498.

Kessler, Daniel P., Mark B. McClellan. 2005. Is hospital competition socially wasteful? *Quarterly Journal of Economics* **115**(2) 577–615.

Larson, R.C. 1987. Perspectives on queues: Social justice and the psychology of queueing. *Operations Research* **35**(6) 895–905.

Leland, W.E., M.S. Taqqu, W. Willinger, D.V. Wilson. 1994. On the self-similar nature of ethernet traffic (extended version). *Operations Research* **2**(1) 1–15.

Maa, J. 2011. The waits that matter. *The New England Journal of Medicine* .

Maister, D. 1985. The psychology of waiting lines. J.A. Czepiel, M.R. Solomon, C.F. Surprenant, eds., *The Service Encounter: Managing Employee/Customer Interaction in Service Businesses*. Lexington, MA: D.C. Heath and Company, Lexington Books.

Maman, S. 2009. Uncertainty in the demand for service: The case of call centers and emergency departments. Master's thesis, Technion - Israel Institute of Technology.

Maman, S., S. Zeltyn, A. Mandelbaum. 2011. Uncertainty in the demand for service: The case of call centers and emergency departments. Working paper.

Mandelbaum, A., Y.N. Marmor, Y. Tseytlin, G.B. Yom-Tov, N. Yuviler. 2011. Exploratory data analysis of patient flow data in an israeli hospital. In preparation.

Mandelbaum, A., P. Momcilovic, Y. Tseytlin. 2012. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.

Mandelbaum, A., A. Sakov, S. Zeltyn. 2000. Empirical analysis of a call center. URL http://iew3.technion.ac.il/serveng/References/ccdata.pdf. Technical Report.

Mandelbaum, A., S. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research* **52**(1) 836–855.

Marmor, Yariv N., Thomas Rohleder, Todd Huschka, David Cook, Jeffrey Thompson. 2011. Cardio vascular surgery simulation in support of system engineering decision making. In preparation.

Marmor, Y.N. 2003. Developing a simulation tool for analyzing emergency department performance. Master's thesis, Technion - Israel Institute of Technology.

Marmor, Y.N. 2010. Emergency-departments simulation in support of service-engineering: Staffing, design, and real-time tracking. Ph.D. thesis, Technion - Israel Institute of Technology.

Marmor, Y.N., B. Golany, S. Israelit, A. Mandelbaum. 2012. Designing patient flow in emergency departments. *IIE Transactions on Healthcare Systems Engineering* **2** 233–247.

Mayo Clinics. 2011. Mission statement of mayo clinic. Http://www.clinicmayo.us/.

McKee, M., A. Rafferty, L. Aiken. 1997. Measuring hospital performance: Are we asking the right questions? *Journal of the Royal Society of Medicine* **90** 187–191.

Norman, D.A. 2009. Designing waits that work. *MIT Sloan Management Review* **50**(4) 23–28.

OECD. 2011. Oecd health data 2011 - frequently requested data. Http://www.oecd.org/document/16/0,3343,en_2649_34631_2085200_1_1_1_1,00.html.

Oliver, Adam, Elias Mossialos. 2004. Equity of access to health care: Outlining the foundations for action. *Journal of Epidemiology and Community Health* **58**(8) 655–658.

Plonski, O., D. Efrat, A. Dorban, N. David, M. Gologorsky, I. Zaied, A. Mandelbaum, A. Rafaeli. 2013. Fairness in patient routing: Maternity ward in rambam hospital. Technical report.

Rafaeli, A., G. Barron, K. Haber. 2002. The effects of queue structure on attitudes. *Journal of Service Research* **5**(2) 125–139.

Ramakrishnan, M., D. Sier, P.G. Taylor. 2005. A two-time-scale model for hospital patient flow. *IMA Journal of Management Mathematics* **16** 197–215.

Reich, Michael. 2011. The workload process: Modelling, inference and applications. Master's thesis, Technion - Israel Institute of Technology.

Schneider, Benjamin, Karen M. Barbera. 2011. Driving customer satisfaction through hr: Creating and maintaining a service climate. Society for Human Resource Management (SHRM) and the Society for Industrial and Organizational Psychology (SIOP).

SEELab. 2011. See lab, technion - israel institute of technology. Http://ie.technion.ac.il/Labs/Serveng/.

Serban, Nicoleta. 2011. A space-time varying coefficient model: The equity of service accessibility. *Annals of Applied Statistics* **5**(3) 2024–2051.

Shaw, C. 2003. How can hospital performance be measured and monitored? URL http://www.euro.who.int/document/e82975.pdf.

Shi, P., M.C. Chou, J.G. Dai, D. Ding, J. Sim. 2014. Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science* **24**(2) 13–14.

Shih, A., S.C. Schoenbaum. 2007. Measuring hospital performance: The importance of process measures. *The Commonwealth Fund: Commission on a High Performance Health System* URL http://www.commonwealthfund.org/usr_doc/1046_Shih_measuring_hosp_performance_process.pdf?section=4039.

Stolyar, S. 2005. Optimal routing in output-queued flexible server systems. *Probability in the Engineering and Informational Sciences* **19** 141–189.

Sullivan, S.E., R.S. Baghat. 1992. Organizational stress, job satisfaction, and job performance: Where do we go from here? *Journal of Management* **18** 353–375.

Taheri, P.A., D.A. Butz, L.J. Greenfield. 2000. Length of stay has minimal impact on the cost of hospital admission. *Journal of the American College of Surgeons* **191**(2) 123–130.

Tezcan, T. 2008. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Math of Operations Research* **33** 51–90.

The OCR Project (IBM, Technion), Rambam. 2011. Service science in hospitals: A research-based partnership for innovating and transforming patients care. Http://ie.technion.ac.il/Labs/Serveng/files/Project_summary_for_SRII.pdf.

Thompson, S., M. Nunez, R. Garfinkel, M. D. Dean. 2009. Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research* **57**(2) 261–273.

Tseytlin, Y. 2009. Queueing systems with heterogeneous servers: On fair routing of patients in emergency departments. Master's thesis, Technion - Israel Institute of Technology.

Tseytlin, Y., A. Zviran. 2008. Simulation of patients routing from an emergency department to internal wards in Rambam hospital. OR Graduate Project, IE&M, Technion.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison Wesley.

Ward, A., M. Armony. 2013. Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research* **61**(1) 228–243.

ynet. 2009. What hospital is the most crowded in Israel? *http://www.ynet.co.il/ articles/0,7340,L-3656123,00.html* .

Yom-Tov, G.B. 2010. Queues in hospitals: Queueing networks with reentering customers in the QED regime. Ph.D. thesis, Technion - Israel Institute of Technology.

Yom-Tov, G.B., A. Mandelbaum. 2014. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *M&SOM* **16**(2) 283–299.

Zeltyn, S., Y.N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, D. Schwartz, K. Moskovitch, S. Tzafrir, F. Basis, A. Shtub, T. Lauterman. 2011. Simulation-based models of emergency departments: Real-time control, operations planning and scenario analysis. *Transactions on Modeling and Computer Simulation (TOMACS)* **21**(4).