Stochastic Systems

ON PATIENT FLOW IN HOSPITALS: A DATA-BASED QUEUEING-SCIENCE PERSPECTIVE

By Mor Armony[*], Shlomo Israelit[†], Avishai Mandelbaum[‡], Yariv N. Marmor[§], Yulia Tseytlin[¶], and Galit B. Yom-Tov[‖]

NYU[*], *Rambam Hospital*[†], *Technion*[‡],
ORT Braude College & Mayo Clinic[§], *IBM Research*[¶], *Technion*[‖]

Hospitals are complex systems with essential societal benefits and huge mounting costs. These costs are exacerbated by inefficiencies in hospital processes, which are often manifested by congestion and long delays in patient care. Thus, a queueing-network view of patient flow in hospitals is natural, for studying and improving its performance. The goal of our research is to explore patient flow data through the lenses of a queueing scientist. The means is exploratory data analysis (EDA) in a large Israeli hospital, which reveals important features that are not readily explainable by existing models.

Questions raised by our EDA include: Can a simple (parsimonious) queueing model usefully capture the complex operational reality of the Emergency Department (ED)? What time scales and operational regimes are relevant for modeling patient length of stay in the Internal Wards (IWs)? How do protocols of patient transfer between the ED and the IWs influence patient delay, workload division and fairness? EDA also underscores the importance of an integrative view of hospital units by, for example, relating ED bottlenecks to IW physician protocols. The significance of such questions and our related findings raises the need for novel queueing models and theory, which we present here as *research opportunities*.

Hospital data, and specifically patient flow data at the level of the individual patient, is increasingly collected but is typically confidential and/or proprietary. We have been fortunate to partner with a hospital that allowed us to open up its data for everyone to access. This enables reproducibility of our findings, through a user-friendly platform that is accessible through the Technion SEELab.

CONTENTS

Keywords and phrases: Queueing Models, Queueing Networks, Healthcare, Patient flow, EDA, Emergency Departments, Hospital wards, Event logs

1. Introduction. Health care systems in general, and hospitals in particular, are major determinants of our quality of life. They also require a significant fraction of our resources and, at the same time, they suffer from (quoting a physician research partner) "a ridiculous number of inefficiencies; thus everybody—patients, families, nurses, doctors and administrators are frustrated." In (too) many instances, this frustration is caused and exacerbated by delays—"waiting for something to happen"; in turn, these delays and the corresponding queues signal inefficiencies. Hospitals hence present a propitious ground for research in Queueing Theory and, more generally, Applied Probability (AP) and Operations Research (OR). Such research would ideally culminate in reduced congestion (crowding) and its accompanying important benefits: clinical, financial, psychological and societal. And for such benefits to accrue, it is critical that the supporting research is data-based.

Unfortunately, however, operational hospital data is accessible to very few researchers, and patient-level data has in fact been publicly unavailable. The reasons span data nonexistence or poor quality, through concerns for patient confidentiality, to proprietorial constraints or attitudes of the data owners. We are thus humbly attempting, in this present work, to change this landscape of data-based hospital OR and, in doing so, introduce a new standard. Specifically, we identify and propose research opportunities and challenges that arise from exploratory analysis of ample hospital data. Just as significantly, we also open up our data and make it universally accessible at the Technion IE&M Laboratory for Service Enterprise Engineering (SEELab): the data can be either downloaded or analyzed online, through a user friendly platform (SEEStat) for Exploratory Data Analysis (EDA). Our goal is thus to provide an entry to and accelerate the learning of data-based OR of hospitals; Interested researchers can reproduce our EDA, and use it as a trigger and a starting point for further data mining and novel research of their own.

1.1. *Patient Flow Focus.* Of particular interest to both researchers and practitioners is *patient flow* in hospitals: improving it can have a significant impact on quality of care as well as on patient satisfaction; and restricting attention to it adds a necessary focus to our work. Indeed, the medical community has acknowledged the importance of patient flow management (e.g. Standard LD.3.10.10, which the Joint Commission on Accreditation of Hospital Organizations (JCAHO, 2004) set for patient flow leadership). This acknowledgment is natural given that operational measures of patient flow are relatively easy to track, and that they inherently serve as proxies for other quality of care measures (see Section 6.1). In parallel, patient flow has caught the attention of researchers in OR in general, and Queueing Theory in particular. This is not surprising: hospital systems, being congestion-prone, naturally fit the framework of Queueing Theory, which captures the tradeoffs between (operational) service quality vs. resource efficiency.

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? To address this question, we study an extensive data set of patient flow through the lenses of a queueing scientist. Our study highlights interesting phenomena that arise in the data, which leads to a discussion of their implications on system operations and queueing modeling, and culminates in the proposal of related research opportunities.

However, patient flow, as highlighted by our title ("On Patient Flow . . ."), is still too broad a subject for a single study. We thus focus on the inter-ward resolution, as presented in the flow chart (process map) of Figure 1; this is in contrast to intra-ward or out-of-hospital patient flow. We further narrow the scope to the relatively isolated ED+IW network, as depicted in Figure 2 and elaborated on in §1.2.1.

1.2. *Rambam hospital.* Our data originates at the Rambam Medical Center, which is a large Israeli academic hospital. This hospital caters to a population of more than two million people, and it serves as a tertiary referral center for twelve district hospitals. The 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), at the flow level of Figure 1, and the resolution level of individual patients. Thus, the data allows one to

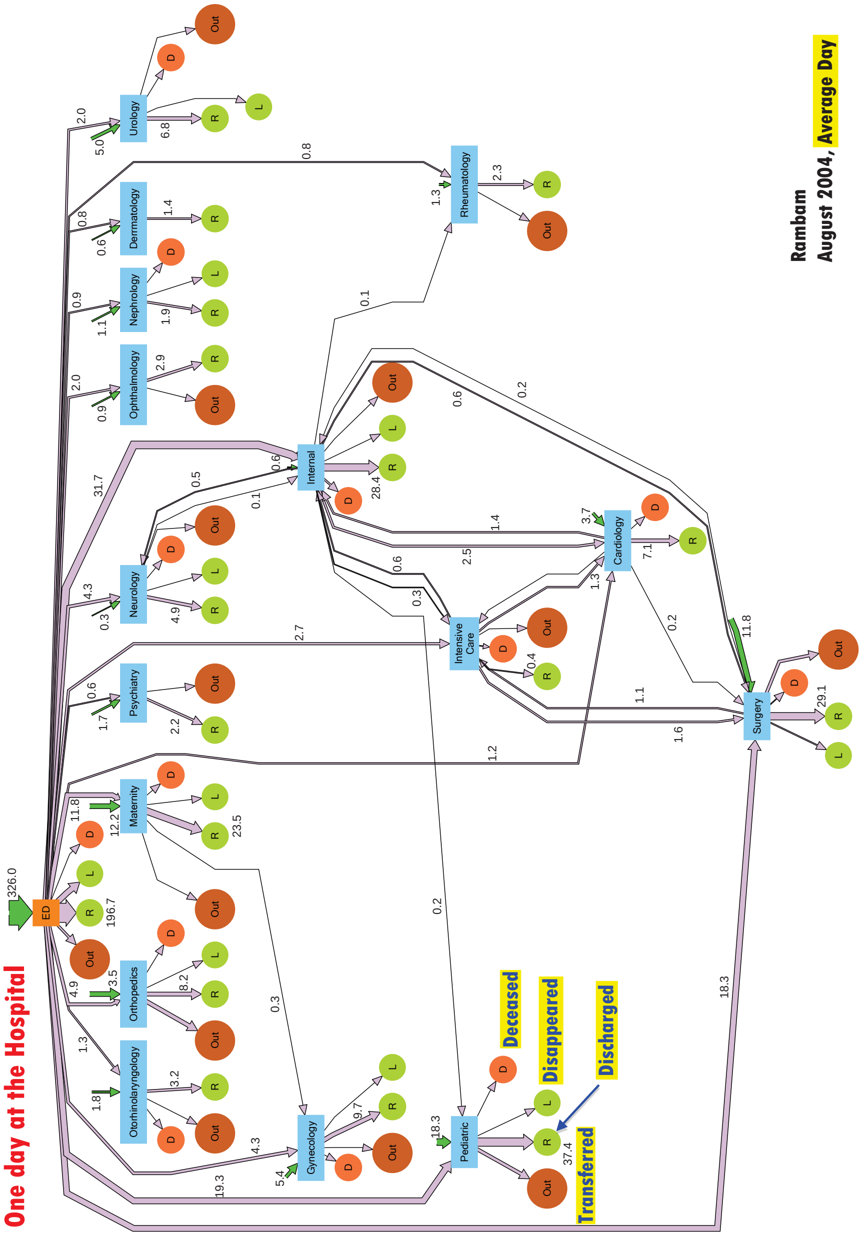


Fig 1: Patient Flow (Process Map) at inter-ward resolution. (Data animation is available at SEEnimations). For example, during the period over which the flow was calculated (August 2004), 326 patients arrived to the ED per day on average, and 18.3 transferred from the ED to Surgery. (To avoid clutter, arcs with monthly flow below 4 patients were filtered out; Created by SEEGraph, at the Technion SEELab.)

follow the paths of individual patients throughout their stay at the hospital, including admission, discharge, and transfers between hospital units.

1.2.1. *The ED+IW network.* 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 lack necessary focus. Instead, and although our data encompasses the entire hospital, we focus on a sub-network that consists of the main Emergency Department (ED) (adult Internal, Orthopedics, Surgery, and Trauma) and five Internal Wards (IWs), denoted by A through E; see Figure 2. This sub-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: According to our data, approximately 53% of the patients entering the hospital remain within this sub-network, and 21% of those are hospitalized in the IWs; indeed, 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 93.5% of the patient transfers

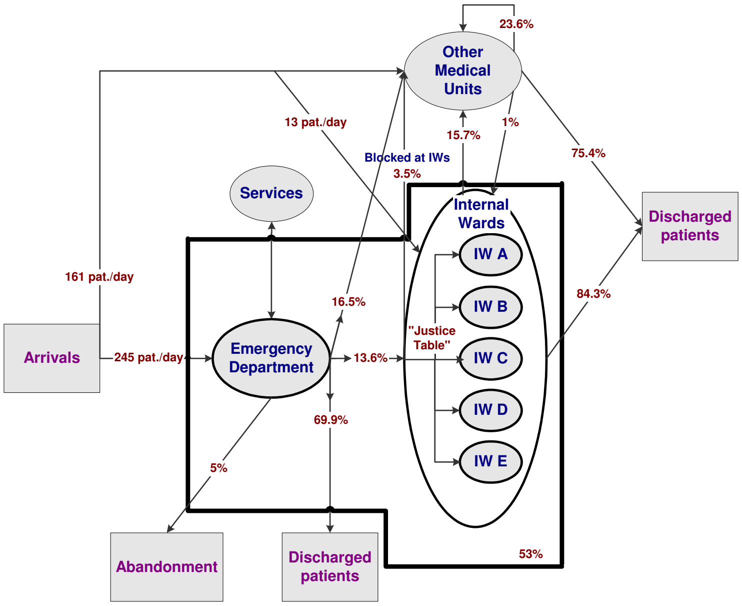


Fig 2: The ED+IW sub-network

into the IWs are either from outside the hospital or from within the ED+IW network.

1.2.2. *Data Description.* Rambam's 2004–2008 patient-level flow data consists of 4 compatible "tables", that capture hospital operations as follows. The first table (Visits) contains records of ED patients, including their ID, arrival and departure times, arrival mode (e.g. independently or by ambulance), cause of arrival, and some demographic data. 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 ward responsible for each patient as, sometimes, due to lack of capacity, patients are not treated in the ward that is clinically most suitable for them; 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.

1.3. *Apologies to the Statistician.* Our approach of learning from data is in the spirit of Tukey's Exploratory Data Analysis (EDA) (Tukey, 1977), which gives rise to the following two "apologies". Firstly, 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, model selection).

Secondly, our data originates from a single Israeli hospital, operating during 2004–2008. This raises doubts regarding the generality of the scientific and practical relevance of the present findings, and rightly so. Nevertheless, other studies of hospitals in Israel (Marmor (2003); Tseytlin (2009) and Section 5.6 of EV) and in Singapore (Shi et al., 2012), together with other privately-communicated empirical research by colleagues, reveal phenomena that are common across hospitals worldwide (e.g. the LOS distributions in Figure 9). Moreover, the present research has already provided the empirical foundation for several graduate theses, each culminating in one or several data-based theoretical papers (see §2.1). All in all, our hope is that reading the manuscript will dispel doubts concerning its broad relevance and significance.

1.4. *Paper structure.* The rest of the paper is organized as follows: We start with a short literature review in Section 2. We then proceed to discuss the gate to the hospital—the ED—in Section 3, followed by the IWs (§4), and the ED+IW network as a whole (§5). We start each section with background information. Next, we highlight relevant EDA, and lastly we propose corresponding research opportunities. In §6, we offer final commentary, where we also provide a broader discussion of some common themes that arise throughout the paper. Finally, the Appendix covers data access instructions and documentation, as well as EDA logistics. We encourage interested readers to refer to EV: a working paper that provides a more elaborate discussion of various issues raised here, and that covers additional topics that are not included here due to focus and space considerations.

2. Some hints to the literature.

Patient flow in hospitals has been studied extensively. Readers are referred to the papers in Hall (2013) and Denton (2013)—both also providing leads to many further references. In the present subsection, we merely touch on published work, along the three dimensions that are most relevant to our study: a network view, queueing models and data-based analysis. Many additional references to recent and ongoing research, on particular issues that arise throughout the paper, will be further cited as we go along. This subsection concludes with what can be viewed as "proof of concept": a description of some existing research that the present work and our empirical foundation have already triggered and supported.

Most research on patient flow has concentrated on the ED and how to improve ED 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 (e.g. operating-room schedules). 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 (e.g. 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 alleviate these delays. Finally, Shi et al. (2012) develops data-based models that lead to managerial insights on the

ED-to-Ward transfer process.

There has been a growing body of research that treats operational problems in hospitals with Operations Research (OR) techniques. Brandeau, Sainfort and Pierskalla (2004) is a handbook of OR methods and applications in health care; the part that is most relevant to this paper is its chapter on Health Care Operations Management (OM). Next, Green (2008) surveys the potential of OR in helping reduce hospital delays, with an emphasis on queueing models. A recent handbook on System Scheduling is Hall (2012)—it includes chapters worth reading and additional leads on OR/OM and queueing perspectives of patient flow. Of special interest is Chapter 8, where Hall describes the challenging reality of bed management in hospitals. Jennings and de Véricourt (2008, 2011) 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, Kolesar and Whitt (2007) survey and develop (time-varying) queueing networks that help determine the number of physicians and nurses required in an ED.

There is also an increased awareness of the significant role that data can, and often must, play in patient flow research. For example, Kc and Terwiesch (2009) is an empirical work in the context of ICU patient flow; it has inspired the analytical model of Chan, Yom-Tov and Escobar (2014) (see also Chan, Farias and Escobar (2014) on the correlation between patient wait and ICU LOS). Another example is Baron et al. (2014) that does both modeling and data analysis for patient flow in outpatient test provision centers. More on patient flow in outpatient clinics and the need for relevant data is discussed in Froehle and Magazine (2013).

2.1. *A proof of concept.* The present research has provided the empirical foundation for several graduate theses and subsequent research papers: Marmor (2010) studied ED architectures and staffing (see Zeltyn et al. (2011) and Marmor et al. (2012)); Yom-Tov (2010) focused on time-varying models with customer returns to the ED (Yom-Tov and Mandelbaum, 2014) and the IWs; Tseytlin (2009) investigated the transfer process from the ED to the IWs (Mandelbaum, Momcilovic and Tseytlin, 2012); Maman (2009) explored over-dispersion characteristics of the arrival process into the ED (Maman, Zeltyn and Mandelbaum, 2011); and Huang (2013) developed scheduling controls that help ED physicians choose between newly-arriving vs. in-process patients, while still adhering to triage constraints (Huang, Carmeli and Mandelbaum, 2011).

3. Emergency Department.

Patient flow in the Emergency Department (ED) is a complex process that involves a multitude of interrelated steps (e.g. Figure 1 in Zeltyn et al. (2011)). This process has been widely investigated, both academically (Hall et al., 2006; Saghafian, Austin and Traub, 2014) and in practice (IHI, 2011; McHugh et al., 2011). We shall hence be content with its empirical *macro* view, which already turns out to be highly informative. Specifically, we view the ED as a black-box, and then highlight interesting phenomena that relate to its patient arrivals, departures, and occupancy counts. Our EDA underscores the importance of including time- and state-dependent effects in the ED—some of these are not readily explained by existing queueing models. Yet, and albeit this dependence, it also reveals that a simple stationary model may provide a good fit for patient-count during periods when the ED is most congested. For limited purposes, therefore, our EDA supports the use of simple stationary models for the ED, which has been prevalent in the literature (e.g. Green et al. (2006) and de Bruin et al. (2009)).

3.1. *Basic facts.*

The main ED has 40 beds and it treats on average 245 patients daily: close to 60% are classified as Internal (general) patients and the rest are Surgical/Orthopedic, excluding a few per day that suffer from multiple trauma. The ED has three major areas: Trauma acute, Internal acute, and Surgical/Orthopedic acute; some of the patients in the latter two are “Walking” patients that do not require a bed. While there are formally 40 beds in the ED, this bed capacity is highly flexible and can be doubled and more. Hence there is effectively no upper bound on how many patients can simultaneously reside within the ED—either in beds or stretchers, chairs, etc. The hospital has other EDs, physically detached from the main one discussed here—these are dedicated to other patient types such as Pediatrics or Ophthalmology. Throughout the rest of our paper, we focus on the main ED and simply refer to it as the ED. Furthermore, within the ED, we focus on Internal (general) patients, in beds or walking: they constitute the majority of ED patients and give rise to most operational challenges.

During weekdays, the average length of stay (ALOS) of patients in the ED is 4.25 hours: this covers the duration from entry until the decision to discharge or hospitalize; it does not include *boarding time*, which is the duration between hospitalization decision to actual transfer. We estimate boarding time to be 3.2 hours on average (See Section 5.2). In addition, 10% (5%) of weekday patients experience LOS that is over 8 (11) hours, and about 3–5% leave on their own (LWBS = left without being seen by a doctor, LAMA = left against medical advice, or Absconded = disappeared throughout the process and are neither LWBS nor LAMA). Finally, out of the 2004–2005 ED patients, around 37% were eventually readmitted; and, overall, 3%, 11%, and 16% of the patients returned within 2, 14, and 30 days, respectively.

3.2. *Exploratory Data Analysis.* In this section we highlight some of our EDA findings that relate to ED patient arrivals and patient-count distribution.

3.2.1. *Time dependency.* As observed also in Green, Kolesar and Whitt (2007), the ED hourly arrival rate varies significantly during the day. In Rambam's ED, it varies by a factor of almost 10; See Figure 3. We also observe a time-lag between the arrival rate and occupancy levels, which is due to the former changing significantly during a patient LOS (Bertsimas and Mourtzinou, 1997). This lag must be accounted for in staffing recommendations (Feldman et al., 2008; Green, Kolesar and Whitt, 2007; Yom-Tov and Mandelbaum, 2014).

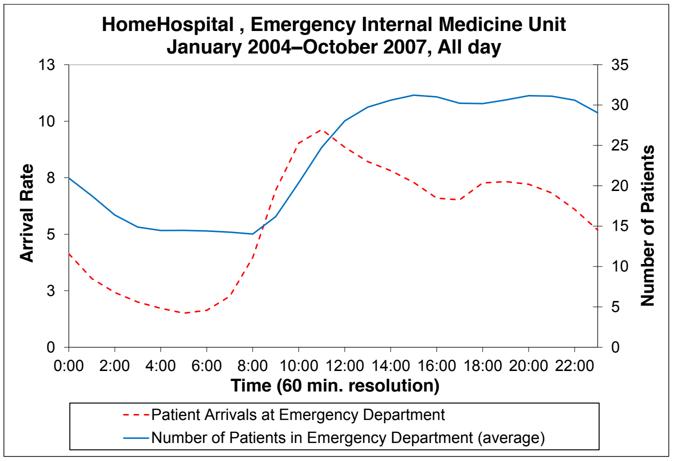


Fig 3: Average number of patients and arrival rate by hour of the day

Analyzing the same data, Maman (2009) found support for the daily arrival process to fit a time-varying Poisson process, but with heterogeneity levels across days such that the *arrival rate* itself must be *random* (slightly over-dispersed). Kim and Whitt (2014) identified similar patterns in a large Korean hospital. The time-varying arrivals contribute to an overall time varying ED environment, which we focus on next.

3.2.2. *Fitting a simple model to a complex reality.* Figure 4 (left) shows 24 patient-count histograms for internal ED patients, each corresponding to a specific hour of the day, with reference (right) to mean patient count, also by hour of the day. (Similar shapes arise from total ED patient count—see Figure 10 in EV.) The figure displays a clear time-of-day behavior: There are two distinct bell-shaped distributions that correspond to low occupancy (15 patients on average) during the AM (3–9AM), and high (30 patients) during the PM (12–11PM); with two transitionary periods of low-to-high (9AM–12PM) and high-to-low (11PM–3AM). We refer to these four periods as the four "occupancy regimes". Interestingly, when asking SEEStat to fit a mixture of three normal distributions to the ED occupancy distribution, the fit algorithm *automatically* detects the low, high and transitionary phases (See Figure 5a).

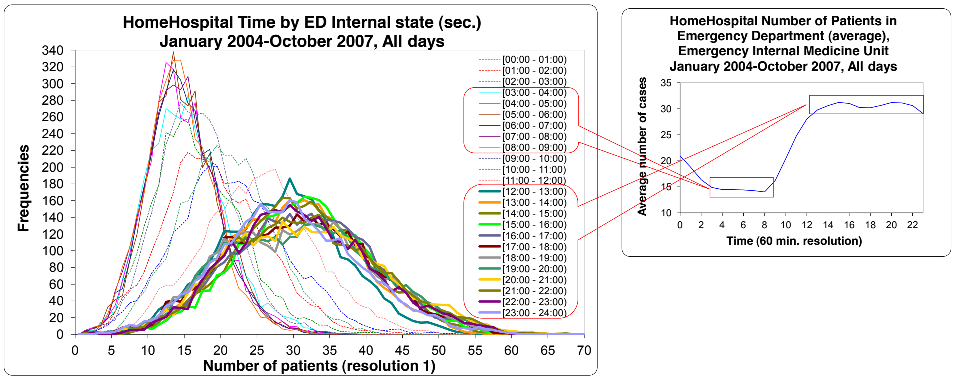
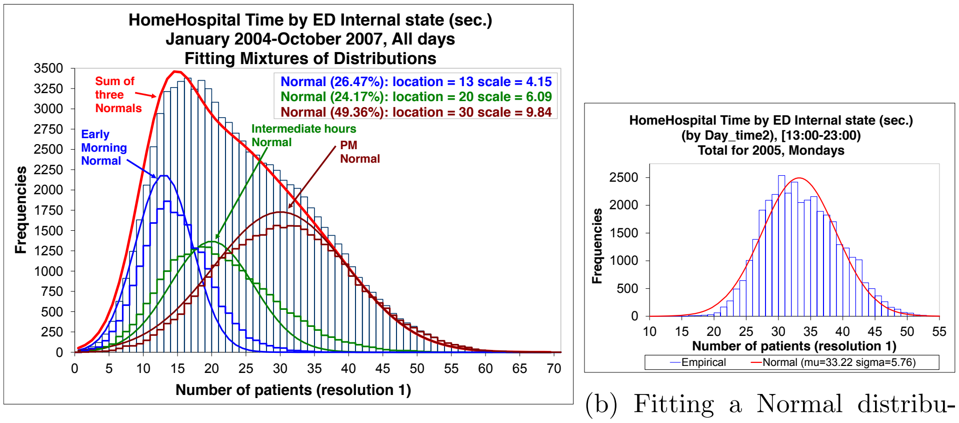


Fig 4: Internal ED Occupancy histogram (left) and Average Census (right), by hour of the day

Further EDA (Figure 5b) reveals that, during peak times (PM), when controlling for factors such as day-of-the-week, patient type and calendar year, one obtains a good fit for the empirical distribution by a "steady-state" normal distribution with equal mean and variance. Hence, one might speculate that the underlying system dynamics can be modeled by an $M/M/\infty$ queue, which has a Poisson steady-state (mean=variance). Alternatively, however, it may also be described as an $M/M/N + M$ model with equal rates of service and abandonment (LWBS, LAMA, or Absconded). It follows that one cannot conclusively select a model through its empirical steady-state distribution—which is a trap that is easy to fall into and from which Whitt (2012) rescued us.

One is thus led to the relevance-boundary of "black-box" ED models: they



(a) Fitting a mixture of three Normal distributions to the Empirical distribution of ED occupancy

(b) Fitting a Normal distribution for a specific year, day of the week, and time of day

Fig 5: Fitting parametric distributions to the Empirical distribution of ED occupancy

may support operational decisions that depend only on total patient count but not on internal dynamics (nor may these decisions alter internal dynamics); or they can model ED sojourn times within a larger hospital model. If in addition, and following Whitt (2012), a birth-death steady-state model is found appropriate for the "black-box", then model reversibility accommodates also applications that *do change* total count: for example, ambulance diversion when total count exceeds a certain threshold, which then truncates the count to this threshold (and the steady-state distribution is truncated correspondingly; see Kelly (1979)). On the other hand, black-box models cannot support ED staffing (e.g. Yom-Tov and Mandelbaum (2014) acknowledges some internal network dynamics), or ambulance diversion that depends on the number of boarding patients (awaiting hospitalization). We discuss this further in Section 3.3.

3.2.3. *State dependency.* In addition to time-dependent effects, we observe that the Internal ED displays some intriguing state-dependent behavior. Specifically, Figure 6 depicts service (or departure) rates as a function of the Internal patient count L (in bed or walking): the left graph displays the total service rate, and the right graph shows the service rate per patient. These graphs cannot arise from commonly used (birth-death) queueing models such as $M/M/N$ (total rate that is linearly increasing up to a certain point and then it is constant) or $M/M/\infty$ (constant rate per patient). In contrast, the per-patient service rate has an interval ($11 \leq L \leq 20$) where it is increasing in L, which is between two intervals of decrease. (The noise at the extremes, $L \leq 3$ and $L \geq 55$, is due to small sample sizes.) Note that Batt and Terwiesch (2012) and Kc and Terwiesch (2009) also found evidence for a state-dependent service rate.

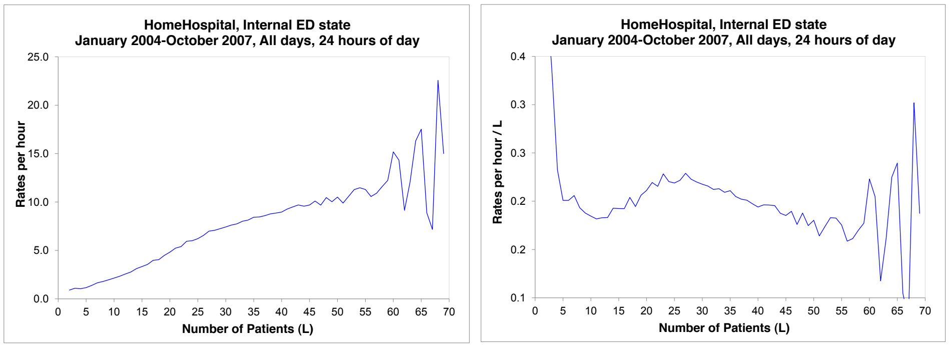


Fig 6: Service rate and service rate per patient as a function of L

What can cause this particular state-dependence of the service rate per patient? We start with the "slowdown" ($L \geq 25$) which, in a congested ED, is to be expected under any of the following scenarios:

- *Multiple resource types with limited capacity:* As the number of occupied beds increases, the overall load on medical staff and equipment increases as well. Assuming a fixed processing capacity, the service rate per bed must then slow down.
- *Psychological:* Medical staff could become emotionally overwhelmed, to a point that exacerbates slowdown (Sullivan and Baghat, 1992).
- *Choking:* Service slowdown may also be attributed to so-called resource "choking": medical staff becomes increasingly occupied with caring for to-be-transferred (boarding) ED patients (who create work while they wait and, moreover, their condition could actually deteriorate), that might end up taking capacity away from the to-be-released patients, thereby "choking" their throughput (see Figure 13 in Section 5.3). The choking phenomenon is well known in other environments such as transportation (Chen, Jia and Varaiya, 2001) and telecommunications (Gerla and Kleinrock, 1980), where it is also referred to as throughput degradation.
- *Time dependency and patient heterogeneity:* Finally, slowdown as well as speedup may be attributed to the combination of time dependent arrivals and heterogenous patient mix (Marmor et al., 2013). We now

expand on this in the context of the speedup effect.

As opposed to the slowdown, the apparent speedup ($10 \leq L \leq 25$) turns out to be an artifact of biased sampling due to patient-heterogeneity and time-variability (as observed in Section 3.2.1). To see this, we further investigate the departure rate per patient, as a function of the patient count, at four different time-of-day intervals (corresponding roughly to the four occupancy regimes identified in Figure 4). For each of these, we observe, in Figure 7, either a constant service rate or a slowdown thereof, but no speedup.

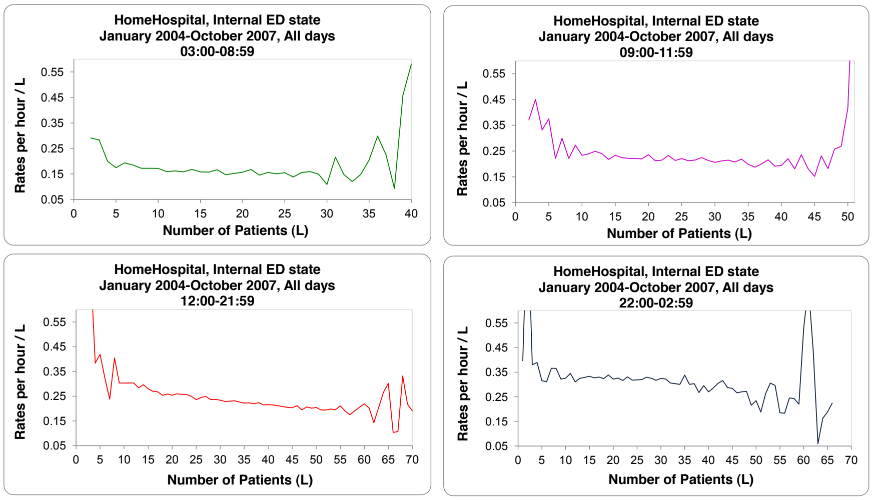


Fig 7: Service rate per patient as a function of L by occupancy regime

Now the rate-per-patient in Figure 6 is a weighted average of the four graphs of Figure 7. But these weights are not constant as a function of the patient count, as seen in Figure 8. Moreover, the service rate as a function of patient count varies at different times of the day. It follows that, what appears to be a speedup (increasing graph), is merely a weighted average of non-increasing graphs with state-dependent weights.

3.3. **Research Opportunities**. Our EDA leaves open questions for further data-based theoretical exploration. For example: What causes the particular shape of time-dependent arrival-rates – the two local peaks in Figure 3 – which is common in many service systems (including hospitals across the globe and call centers)? What is the dominant cause for service-rate slow-down in Figure 6, and what can be done to alleviate this slowdown?

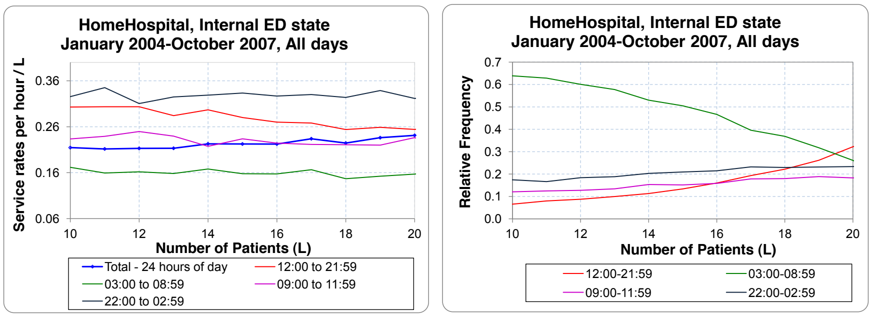


Fig 8: Service rate as a function of $10 \leq L \leq 20$ (left), and Relative frequency (weight) of occupancy regime per L (right)

How does one separate the effects of time- and state-dependency, which one is the more dominant and under what circumstances?

In addition, the observations in this section also raise some broader research directions, within several (somewhat overlapping) model dimensions: granularity, performance metrics, and applications.

3.3.1. *Model granularity.* Our focus in this section has been on overall ED (Internal) patient count. This aggregates ED dynamics into merely arrivals and departures which, as described in Subsection 3.2.2, yields a useful black-box model but with a limited applicability scope. In contrast to this macro level, one could consider a detailed model (most likely simulation, as in Zeltyn et al. (2011)), which acknowledges explicitly micro-events at the level of individual patients and providers (physicians, nurses).

The macro- and micro-models are two extreme cases of model granularity, with a range of levels in between. Such intermediate levels could correspond, for example, to the overall design of ED patient-flow (Marmor et al., 2012); or to queueing models (Huang, Carmeli and Mandelbaum, 2011; Yom-Tov and Mandelbaum, 2014) that focus on phenomena (patients re-visiting physicians) or function (physician staffing and scheduling). The granularity level to be used depends on the target application, data availability and analytical techniques. Choosing the “right” level for an OR/queueing model has been mostly art, which calls for systematizing this choice process. It could start with Whitt (2012) and (Dong and Whitt, 2014) that fit birth-death models, and continue with existing and possibly novel statistical techniques for model selection (Burnham and Anderson, 2002).

3.3.2. *Performance metrics.* There are numerous ED performance metrics that have not been discussed here or have merely been touched upon. Of special importance is time-till-first-consultation, which is often part of triage protocols: e.g. following the Canadian Triage and Acuity Scale (Canadadian-Triage), 90% of Category 3 (Urgent) patients should be seen by a physician within 30 minutes of arrival. Another measure that has gained recent prominence is readmission rates: it is being used as a proxy for clinical quality of care, and we further discuss it in Section 6.1.1.

Additional measures include length of stay (LOS), abandonment (LWBS, LAMA, absconded), workload and offered load, bed utilization, boarding times, staff-to-bed ratios, and customers who are blocked upon their ED arrival (e.g. ambulance diversion). These metrics are mostly related to ED congestion (Hwang et al. (2011) lists over 70), and they have given rise to prevalent crowding indices (e.g. Bernstein et al. (2003); Hoot et al. (2007)). While such indices support daily ED management, they arose from ad-hoc statistical analysis that seeks to summarize (e.g. via regression) the state of ED congestion. OR and queueing models, on the other hand, constitute a natural systematic habitat for congestion indices. The models can thus help validate existing indices or devise new ones, for example by solving control problems of patient flow that yields rigorous state-summaries and sufficient statistics (Huang, Carmeli and Mandelbaum (2011)).

Unfortunately, useful metrics are often difficult or impossible to measure from data. One can then attempt to infer them from the measurable. An example is patients' patience (the *time* a patient is willing to wait before abandoning the ED); while the overall abandonment proportion is observable, exact times till abandonment are not. Specifically, patients are either served, in which case their waiting time provides a lower bound for their patience, or they are discovered missing when called for service, which provides an upper bound. Statistical inference of ED (im)patience therefore requires novel models and methods: these would combine current-status (Sun, 2006) and survival-analysis (Brown et al., 2005)—in the latter, abandonment times are observed, while they are not in the former.

3.3.3. *Applications.* Applications of queueing models to ED patient flow include the following categories: ED design (e.g., Marmor et al. (2012)), capacity sizing, staffing (e.g., Yom-Tov and Mandelbaum (2014)), and flow control (e.g., Allon, Deo and Lin (2013); Dobson, Tezcan and Tilson (2013); Hagtvedt et al. (2009); Huang, Carmeli and Mandelbaum (2011)). Design challenges cover, for example, operational (fast-track) vs. clinical priorities (see also Zeltyn et al. (2011)), physician-led triage vs. the prevalent nurse-led (Burström et al., 2012; Oredsson et al., 2011), and the creation of a dedicated sub-ED (e.g. for patients with chest-pain; Zalenski et al. (1998)). Addressing these challenges, as well as delving into the other above-mentioned categories, would require data beyond our present resolution and hence we do not elaborate further.

4. Internal Wards.

Internal Wards (IWs), often referred to as General Internal Wards or Internal Medicine Wards, are the "clinical heart" of a hospital. Yet, relative to EDs, Operating Rooms and Intensive Care Units, IWs have received less attention in the Operations literature; this is hardly justified. IWs and other medical wards offer a rich environment in need of OR/OM research, which our EDA can only tap: It has revealed multiple time-scales of LOS, intriguing phenomena of scale-diseconomies and coexisting operational-regimes of multiple resource types (beds, physicians). These characteristics are attributed to IW inflow design, capacity management and operational policies (e.g. discharge procedures, physician rounds).

4.1. *Basic facts.*

Rambam hospital has five Internal Wards consisting of about 170 beds that accommodate around 1000 patients per month. Wards A through D are identical from a clinical perspective; the patients treated in these wards share the same array of clinical conditions. Ward E is different in that it admits only patients of less severe conditions. Table 1 summarizes the operational profiles of the IWs. For example, bed capacity ranges from 24 to 45 beds and Average LOS (ALOS) from 3.7 to 6 days.

TABLE 1
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).

IWs B and E are by far the smallest (least number of beds) and the "fastest" (shortest ALOS, highest throughput). The superior operational performance of IW E is to be expected as it treats the clinically simplest cases. In contrast, the “speed” of IW B is not as intuitive because this ward is assigned the same patient mix as IWs A,C, and D.

A shorter ALOS could reflect a more efficient clinical treatment or, alternatively, a less conservative discharge policy. Either must not arise from clinically premature discharges of patients, which would hurt patients clinical quality of care. To get a grasp on that quality, we use its operational (accessible hence common) proxy, namely patient readmission rate (proportion of patients who are re-hospitalized within a pre-specified period of time: one month in our case). In Table 1 we observe that the readmission rate of IW B is comparable to the other wards. Moreover, patient surveys by Elkin and Rozenberg (2007) indicated that satisfaction levels do not differ significantly across wards. We conclude that IW B appears to be operationally superior yet clinically comparable to the other wards. This fact may be attributed to the smaller size of IW B, which we return to in Section 4.3.3.

4.2. *EDA: LOS—a story of multiple time scales.* Next, we examine the distribution of LOS in the IWs. While it is to be expected that clinical conditions affect patients LOS, the influence of operational and managerial protocols is less obvious. It turns out that some of this influence can be uncovered by examining the LOS distribution at the appropriate time scale.

Figure 9 shows the LOS distribution in IW A, in two time scales: days and hours. At a daily resolution, the Log-Normal distribution turns out to fit the data well. When considering an hourly resolution, however, a completely different distribution shape is observed: there are peaks that are periodically 24 hours apart, which correspond to a *mixture* of daily distributions. (We found that a normal mixture fits quite well, as depicted by the 7 normal mixture-components over the range of 0–150 hours in the right diagram of Figure 9.)

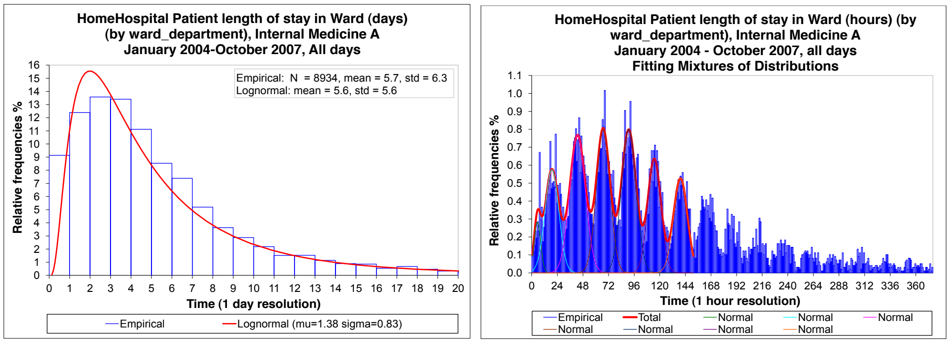


Fig 9: LOS distribution of IW A in two time-scales: daily and hourly

These two graphs reveal the impact of two operational protocols: The daily time scale represents physician decisions, made every morning, on whether to discharge a patient on that same day or to extend hospitalization by at least one more day. The second decision is the hour-of-day at which the patient is actually discharged. This latter decision is made according to the following discharge process: It starts with the physician who writes the discharge letters (after finishing the morning rounds); then nurses take care of paperwork, instructing patients on how to continue medical treatment after discharge, and then arranging for transportation (if needed). The discharge procedure is performed over "batches" of patients and, hence, takes a few hours. The result is a relatively low variance of the discharge time, as most patients are released between 3pm and 4pm—see Figure 10; which provides an explanation for the observed peaks in the hourly LOS distribution that are spaced 24 hours apart. The variation around these peaks is determined by the arrival process: patients are hospitalized in IWs almost exclusively over a 12-hour period (10am–10pm), with a peak in arrival rate between 3pm–7pm (Figure 10).

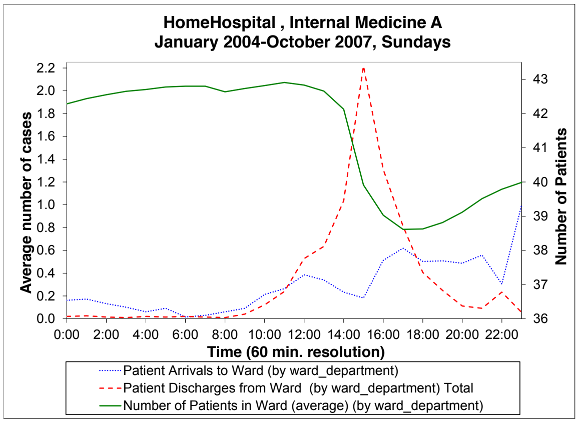


Fig 10: Arrivals, departures, and average number of patients in Internal wards by hour of day

Note that the arrival process to the IWs couples almost exclusively with the departure process from the ED, and hence the timing of its peak (3pm–7pm) is naturally coupled with IW discharge peaks (3pm–4pm). In other words, and as further discussed in Section 5.6, the discharge policy from IWs significantly influences ED congestion. Similar observations in a Singapore hospital led Shi et al. (2012) to model an inpatient ward as a 2-time-scale system, and to consequently propose flow-stabilization as a means of reducing delays.

Who is the Server?: Operational time-resolutions, specifically days/hours and hours/minutes for IWs, correspond to the time scale by which service durations are naturally measured which, in turn, identifies a corresponding notion of "a server". For example, IW LOS resolution in days corresponds to conceptualizing beds as servers. This is the setup in de Bruin et al. (2009) and Bekker and de Bruin (2010) who assume (hyper-) exponential LOS. (Log-Normal service durations are yet to be accommodated by queueing models.) Another IW resolution is hours, which is appropriate with servers being nurses, physicians or special IW equipment. Here service times are measured in minutes or parts of an hour, and offered load (workload) is calculated (from arrival and service data) in units of, say, hours of work that arrive per hour of the day.

4.2.1. **Research Opportunities***.* We discuss here workload characterization, protocol mining via LOS distributions, flow control and why Log-Normal.

Offered Load, or Workload: The offered load is the skeleton around which capacity (staffing in the case of personnel) is dimensioned (Green, Kolesar and Whitt, 2007). Consider nurses as an example. Their time-varying offered load results from both routine and special care, and it varies during the day for at least two reasons (see Equation (1) in Mandelbaum, Momcilovic and Tseytlin (2012)): (a) routine care depends linearly on patient count, which varies over a day (Figure 10), and (b) admission and discharge of patients require additional work beyond routine, and it is more frequent during some hours than others (Figure 10). Combining both of these time variations, it is clear that staffing levels must (and actually do) vary during the day, hence the importance of observing and understanding the system in hourly resolution. As mentioned above, some efforts to develop queueing models for nurse staffing in medical wards have been carried out by Jennings and de Véricourt (2011), Green and Yankovic (2011) and Yom-Tov (2010). However, these works neither explain or incorporate the LOS distribution observed in our data, nor do they distinguish between routine, admission, and discharge workload. Even such a distinction might not be rich enough: indeed, the hospital environment calls for a broader view of workload, which we discuss in Section 5.5.4.

LOS and Protocols: LOS or Delay distributions encapsulate important operational characteristics, and can hence be used to suggest, measure or track improvements. Consider, for example, the *hourly* effect of IW LOS (Figure 9), which is due to IW discharge protocols. It calls for an effort in the direction of smoothing IW discharge rates over the day (Shi et al., 2012). Taking an example from elsewhere at the hospital, consider the differences in shape of LOS distribution between two Maternity wards (§4.2.1 in EV), which result from differing patient mix; it suggests the redesign of routing protocols towards a more balanced workload (Plonski et al., 2013). Queueing models are natural for analyzing the interplay between LOS distributions and operational protocols. This leads to open data-based questions in two directions: first, incorporating protocols (e.g. patient priorities, resource scheduling) in queueing models and validating the theoretical LOS distribution against data (performance); second and conversely, mining protocols from data. We now give two examples, one for each of the two directions.

Flow Control: How will changes in the IW discharge process influence the system? For example, would the balancing of discharges more uniformly over the day benefit the entire hospital? How would such a change influence delays of patients waiting to be transferred into the IW from the ED? This connection between ED boarding and ward discharges was explored by Shi et al. (2012). We return to it in Section 5.6.

Why Log-Normal? A long-standing challenge is to explain the prevalence of Log-Normal as a distribution of service durations (e.g. IW LOS in days here, or durations of telephone calls in Brown et al. (2005)). Is Log-normality due to service protocols? It is perhaps an inherent attribute of customer service requirements? Note that Log-Normal has an intrinsic structure that is both *multiplicative*—its logarithm is a central limit, and *additive*—it is infinitely divisible, being an integral against a Gamma process (Thorin, 1977). Can these properties help one explain the empirical Log-Normal service time distribution?

4.3. *EDA: Operational regimes and economies of scale.* An asymptotic theory of many-server queues has been developed in recent years (Gans, Koole and Mandelbaum (2003) can serve as a starting point), which has highlighted three main operational regimes: Efficiency Driven (ED), Quality Driven (QD) and Quality & Efficiency Driven (QED). The ED-regime prioritizes resource efficiency: servers are highly utilized (close to 100%), which results in long waits for service. In fact, waiting durations in the ED regime are at least in the order of service times. In the QD regime, the emphasis is on the operational quality of service: customers hardly wait for service, which requires that servers be amply staffed and thus available to serve. Finally, the QED regime carefully balances service quality and server efficiency, thus aiming at high levels of both and achieving it in systems that are large enough. Under the QED regime, server utilization could exceed 90%

while, at the same time, possibly half of the customers are served without delay, and those delayed wait one order of magnitude less than their service duration. The QED regime also exhibits economies of scale in the sense that, as the system grows, operational performance improves.

Many-server queueing theory is based on asymptotic analysis, as the number of servers grows indefinitely. Nevertheless, QED theory has been found valuable also for small systems (few servers) that are not exceedingly overloaded. This robustness to system size is due to fast rates of convergence (Janssen, van Leeuwaarden and Zwart, 2011) and, significantly, it renders QED theory relevant to healthcare systems (Jennings and de Véricourt, 2011; Yom-Tov and Mandelbaum, 2014). One should mention that, prior to the era of many-server theory, asymptotic queueing theory was mostly concerned with relatively small systems—that is few servers that are too overloaded for QED to be applicable (e.g. hours waiting time for service times of minutes). This regime is nowadays referred to as conventional heavy-traffic (Chen and Yao, 2001) and, at our level of discussion, it is convenient to incorporate it into the ED-regime.

In the following subsection, we seek to identify the operational regime that best fits the IWs. We then investigate (§4.3.3) the existence of economies-of-scale in the hospital environment. We shall argue that, although IW beds plausibly operate in the QED regime, there is nevertheless evidence for *dis-economies* of scale.

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

We argue that IW beds operate (as servers) in the QED regime. To support this statement, we first note that our system of IWs has many (10's) beds/servers. Next we consider three of its performance measures: (a) bed occupancy levels; (b) fraction of patients that are hospitalized in non-IWs while still being under the medical care of IW physicians (patients who were *blocked* from being treated in IWs due to bed scarcity); (c) ratio between waiting time for a bed (server) and LOS (service time).

Considering data from the year 2008, we find that 3.54% of the ED patients were blocked, the occupancy level of IW beds was 93.1%, and patients waited hours (boarding) for service that lasted days (hospitalization). Such operational performance is QED—single digit blocking probability, 90+% utilization and waiting duration that is one order of magnitude less than service. Preliminary formal analysis, carried out in Section 4.3.1 of EV, demonstrates that QED performance of a loss model (Erlang-B, as in de Bruin et al. (2009)) usefully fits these operational performance measures of the IWs.

Turning to physicians as servers, we argue that they operate in the ED regime (conventional heavy traffic). This is based on the following observation: from 4pm to 8am on the following morning, there is a single physician on duty in each IW, and this physician admits the majority of 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 become available. The admission process by the physician lasts approximately 30 minutes, and waiting time for physicians is plausibly hours (it takes an average of 3.2 hours to transfer a patient from the ED to the IWs; see Section 5.2). Performance of physicians is therefore Efficiency Driven.

4.3.2. **Research Opportunities**. We identified two operational regimes, QED and ED, that coexist within the ED+IW: waiting in the ED for IW service. What queueing models and operational regimes can valuably capture this reality? Note that such models must accommodate three time scales: minutes for physician treatment, hours for transfer delays, and days for hospitalization LOS. Some questions that naturally arise are the following: How do the regimes influence each other? Can we assume that the "bottleneck" of the system is the ED resource (physicians)? Thus, can one conclude that adding physicians is necessary for reducing transfer delays, while adding beds would have only a marginal impact on these delays? How would a change of physician priority influence the system, say giving higher priority to incoming patients (from the ED) over the already hospitalized (in the IWs)? Does the fact that physicians operate in the ED-regime eliminate the economies of scale that one expects to find in QED systems? Empirical observations that will now be presented suggest that this might indeed be the case.

4.3.3. *Diseconomies of scale (or how ward size affects LOS).* Our data (Table 1) exhibits what appears to be a form of diseconomies of scale: a smaller ward (IW B) has a relative workload that is comparable to the larger wards, yet it enjoys a higher turnover rate per bed and a shorter ALOS, with no apparent negative influence on the quality of medical care. The phenomenon is reinforced by observing changes in LOS of IW B, when the number of beds in that ward changes. Figure 11 presents changes in ALOS and the average patient count, in IWs B and D over the years. During 2007, the ALOS of Ward B significantly increased. This was due to a temporary capacity increase, over a period of two months, during which IW B was made responsible for 20 additional beds. We observe that, although the same operational methods were used, they seem to work better in a smaller ward. In concert with the latter observation, we note a reduction in ALOS of IW D, mainly from 2007 when ward size decreased as a result of a renovation. One is thus led to conjecture that there are some drawbacks in operating large medical units—e.g. larger wards are more challenging to manage, at least under existing conditions.

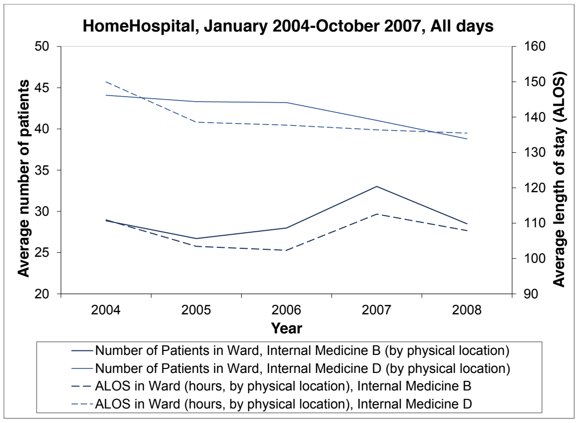


Fig 11: Average LOS and number of patients in Internal wards B and D by year

Several factors could limit the blessings of scale economies:

- *Staffing policy*: It is customary, in this hospital, to assign an IW nurse to a fixed number of beds; then nominate one experienced nurse to be a *floater* for solving emerging problems and help as needed. This setting gives little operational advantage to large units, if at all: the larger the unit the less a single floater can help each nurse. The tradeoff that is raised is between personal care (dedicated servers hence care continuity) vs. operational efficiency (pooling). This tradeoff has been addressed in call centers (Aksin, Karaesmen and Ormeci, 2007; Jouini, Dallery and Aksin, 2009), and in outpatient medical care (Balasubramanian, Muriel and Wang, 2012; Balasubramanian et al., 2010), but inpatient healthcare will surely add its own idiosyncracies. Another natural tradeoff that arises is whether the floater should indeed be an experienced nurse, or is it better to let more junior nurses be floaters so that they can learn from this broader experience.
- *Centralized medical responsibility*: Ward physicians share the responsibility over all patients. Every morning, the senior physicians, resi-

dents, interns, and medical students examine every patient case together (physicians' rounds) and discuss courses of treatment. This is essential as Rambam hospital is a teaching hospital, and one of its central missions is the education and training of doctors. Naturally, the larger the unit the longer its morning round and, consequently, less capacity is available for other tasks (e.g. admissions and discharges).

4.3.4. ***Research opportunities.*** In Section 4.3.2 of EV we provide additional plausible explanations for the observed diseconomies of scale. This phenomenon is important to model carefully and understand, as it can significantly affect decisions on unit sizing and operational strategy. While Queueing Theorist are well equipped to address the operational dimensions of such decisions, they must collaborate with researchers from other disciplines such as organizational behavior for complete comprehension. Now suppose one takes size differences among wards as a given fact (e.g. due to space constraints that cannot be relaxed). Then the following question arises: What protocol should be used to route patients from the ED to the wards, in order to fairly and efficiently distribute workload among them? This challenge is directly related to the process of transferring patients from the ED to the IWs, which is the topic of the next section.

5. The ED+IW Network.

After discussing the ED and IWs separately, in this section we discuss the ED+IW network as a whole. We start with the "ED-to-IW" process of transferring patients from the ED to the IWs. One may think of this process as the "glue" that connects the ED to the IWs. We discuss delays in the transfer process (Sections 5.2-5.4) and fairness in this process towards both patients and medical staff (Section 5.5). We conclude, in Section 5.6, with an integrative view of the interplay between the three components: ED, IWs, and ED-to-IW.

5.1. *ED-to-IW Transfer Process: Basic facts.*

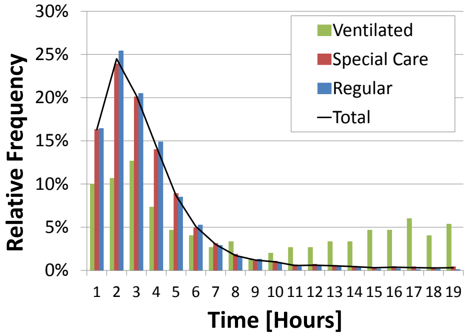
The "ED-to-IW" process covers patient transfers from the ED to the IWs. We view this process in the context of *flow or routing control*. Routing in *hospitals* differs from routing in other service systems, for various reasons including incentive schemes, customers' (patients') limited control (or even helplessness), and the timing of the routing decision. Thus, although the transfer process involves routing-related issues similar to those that have been looked at extensively in the queueing literature, our data indicate that unusual system characteristics significantly affect delays and fairness features in a hospital setting, which creates many research opportunities.

A patient, whom an ED physician decides to hospitalize in an IW, is assigned to one of five wards, according to a certain *routing policy* (described momentarily). If that ward is full, its staff may ask for reassignment with the approval of the hospital's Head Nurse. Once the assigned ward is set, the ward staff prepares for this patient's arrival. In order for the transfer to commence, a bed and medical staff must be available, and the bed and equipment must be prepared for the specific patient (including potential rearrangement of current IW patients). Up to that point, the patient waits in the ED and is under its care and responsibility. If none of the IWs is able to admit the patient within a reasonable time, the patient is "blocked", namely transferred to a non-internal ward. Then the latter undertakes *nursing* responsibilities while *medical* treatment is still obtained from an IW physician.

An integral component of the transfer process is a *routing policy*, or patient assignment algorithm. As described in Section 4.2, Wards A–D provide similar medical services, while Ward E treats only the less severe patients. The similarity between Wards A–D requires a systematic assignment scheme of patients to these wards. Rambam hospital determines the assignment via a round-robin (cyclical) order among each patient type (ventilated, special care, and regular), while accounting for ward size (e.g. if Ward X has twice as many beds as Ward Y, then Ward X gets two assignments per one assignment of Y). This scheme is implemented by a computer software called "The Justice Table". As the name suggests, the algorithm was designed by the hospital to ensure fair distribution of patient load among wards, so that staff workload will be balanced. It is worth noting that a survey among 5 additional hospitals in Israel (EV, Section 5.6) revealed that a cyclical routing policy is very common; yet, some hospitals apply alternative assignment schemes, for example, random assignment by patient ID. Interestingly, only one of the surveyed hospitals uses an assignment that takes into account real-time bed occupancy.

5.2. *Delays in transfer.* As is customary elsewhere, the operational goal of Rambam hospital is to admit ED boarding patients to the IWs within *four hours* from decision of hospitalization. However, the delays are often significantly longer. The waiting-time histogram in Wards A–D, for the years 2006-2008, is depicted in Figure 12. We observe significant delays: while the average delay was 3.2 hours, 25% of the patients were delayed for more than 4 hours.

An interesting phenomenon is observed when analyzing transfer delays by patient type. We note that, on average, ventilated patients wait much longer (8.4 hours) than regular and special care patients (average of 3 and



Patient Type	AVG (STD)	% delay $\leq$ 4 h	% delay $>$ 10 h
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%

Fig 12: Transfer time by patient type, in hours

* 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)

3.3 hours respectively)—see Figure 12. In particular, the delay distribution of these ventilated patients is bi-modal with 41% of such patients delayed by more than 10 hours. Ventilated patients must have the highest priority in transfer but, in reality, many do not benefit from it.

How come do many of the ventilated patients experience such long delays? We observe that the shorter delays of the ventilated patients ($\leq$ 4 hours) have a pattern that resembles that of the other two patient types. The longer delays are harder to decipher. Possible explanations include: (a) Ventilated patients are hospitalized in a *sub-ward* inside the IW (A–D), often referred to as Transitional (intensive) Care Unit (TCU) (Armony, Chan and Zhu, 2013). Each such TCU has only 4–5 beds. The average occupancy rate of the TCUs at Rambam is 98.6%; the combination of high occupancy with a small number of beds results in long waits during overloaded periods. (b) Ventilated patients require a highly qualified staff to transfer them to the ward. Coordinating such transfers takes longer.

5.2.1. **Research Opportunities.** Delays in transfer add opportunities to those arising from protocol mining, as discussed at the end of §4.2.1; relevant here is the specific challenge of deciphering a routing protocol from data such as in Figure 12. In addition, one would like to be able to analyze and optimize patient-flow protocols in queueing models, specifically here fork-join networks (representing synchronization between staff, beds and medical equipment) with heterogeneous customers. Such models, under the FCFS discipline, were approximated in Nguyen (1994). Their control was discussed in Atar, Mandelbaum and Zviran (2012) and Leite and Fragoso (2013).

The discussion above also raises the tension between pooling and continuity-of-care. The fact that Rambam chose to distribute TCU beds among four IWs, instead of having one larger TCU, definitely increases waiting time for a TCU bed. Nevertheless, it is also advantageous from the quality-of-care perspective to have the TCU beds be part of an IW since, when patients' condition improve, they are transferred from the TCU in the IW to a regular room in the same IW, while continuing treatment by the same medical staff (physicians and nurses). This continuity of care reduces the number of hand-offs, which are prone to loss of information and medical errors. The tradeoff between pooling and continuity-of-care is an interesting challenge to navigate using OR methods.

5.3. *Influence of transfer delays on the ED.* Patients awaiting transfer (boarding patients) overload the ED: beds remain occupied while new patients continue to arrive, and the ED staff remains responsible for these boarding patients. Therefore, the ED in fact takes care of two types of patients: *boarding patients* (awaiting hospitalization) and *in-process patients* (under evaluation or treatment in the ED). Both types suffer from delays in the transfer process.

Boarding patients may experience significant discomfort while waiting: the ED is noisy, it is not private and often does not serve hot meals. In addition, ED patients do not enjoy the best professional medical *treatment* for their particular condition, and do not have dedicated attention as in the wards. Moreover, longer ED stays are associated with higher risk for hospital-acquired infections (nosocomial infections). Such delays may increase both hospital LOS and mortality rates, similarly to risks of delays in ICU transfer (e.g. Chalfin et al. (2007); Long and Mathews (2012); Maa (2011)). Hence, the longer patients wait in the ED, the higher the likelihood for clinical deterioration and the lower is their satisfaction.

In-process ED patients may suffer from delays in treatment, as additional workload imposed by transfer patients on ED staff can be significant. Figure 13 shows our estimates of the fraction of time that ED physicians spent caring for the transfer patients, assuming (the Rambam experience) that every such patient requires an average of 1.5 minutes of physician's time every 15 minutes. We observe that transfer patients take up to 11% of physician time in the ED. This extra workload for the ED staff, that occurs at times when their workload is already high, results in "wasted" capacity and *throughput degradation*, as discussed in Section 3.2.3.

To summarize, by improving patient flow from the ED to the IWs, in particular reducing transfer times, hospitals can improve the service and

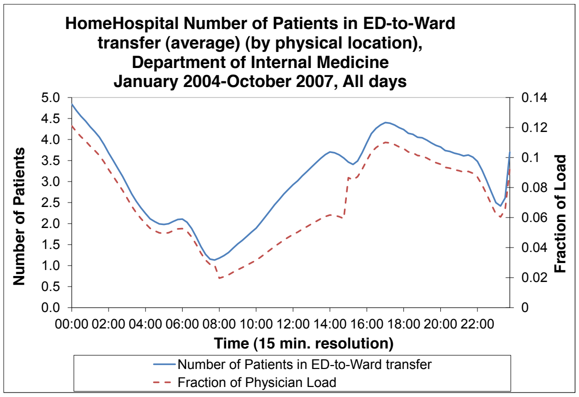


Fig 13: Number of patients in ED-to-IW transfer (A–E) and the fraction of time that ED physicians devote to these patients

treatment provided to both transfer and in-process patients. In turn, reducing the workload in the ED would improve response to arriving patients and could, in fact, save lives.

5.3.1. **Research Opportunities**. The delays in transfer give rise to the following research 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 customers receive service only once they reach a server, and not while waiting in queue. In contrast, here a waiting patient is "served" by both the ED and the IW. In the ED, clinical treatment is provided: according to regulations, transfer patients must be examined at least every 15 minutes. In the ward, "service" actually starts prior to the physical arrival of the patient, when 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 modeling the ED-to-IW process, and it affects staffing, work scheduling, etc. A natural modeling framework here would be queueing networks with signals (Chao, Miyazawa and Pinedo, 1999).
2. *Emergency Department architecture:* As described, ED staff attends to two types of patients: transfer and in-process. Each type has its own 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 service is random, in

nature.

One may consider two options for ED architecture: (a) treat transfer and in-process patients together in the same physical location, as is done at Rambam, or (b) move the transfer patients to a transitional unit (sometimes called "delay room" or "observation room"), where they wait for transfer; this is done, for example, in a Singapore hospital that we were in contact with. Note that using option (b) implies having dedicated staff, equipment and space for this unit. The following question then arises: Under what conditions is each of these ED architectures more appropriate?

Note that the Singapore hospital architecture is even more complicated than (b) above, as the responsibility for the transfer patients is handed over to IW physicians after a two-hour ED boarding delay. This provides the IW medical staff with an *incentive* to transfer the patients to the ward, as soon as possible, where they can be comfortably treated. In EV, Section 5.6, we further discuss how different architectures are related to incentive schemes and, in turn, influence delay times.

5.4. *Causes of delay.* In order to understand the causes of long delays in the ED-to-IW transfer, we interviewed hospital staff, conducted a time and motion study, and further explored our data. We learned that delays are caused not only by bed unavailability; patients often wait even when there are available beds. Indeed, our data shows 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 12 shows, the probability to be admitted to the wards, immediately (or within a short time) after 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 13 plausible causes for delay, which are summarized in the Cause-and-Effect (Fishbone) diagram depicted in Figure 14. We elaborate here on two that have some interesting modeling aspects.

1. *Timing of routing decision: Input-queued vs. Output-queued system.* Recall that preparation for a transfer of a *particular* patient starts in the designated ward, prior to the actual transfer. This forces the hospital to adopt an output-queued scheme (Stolyar, 2005), where each patient is first assigned to an IW and then waits until the ward is able to admit. This is in contrast to a scheme where patients are placed in a "common" queue, then routed to an IW only once at the head of the line and a bed in *any* of the IWs becomes available. The latter

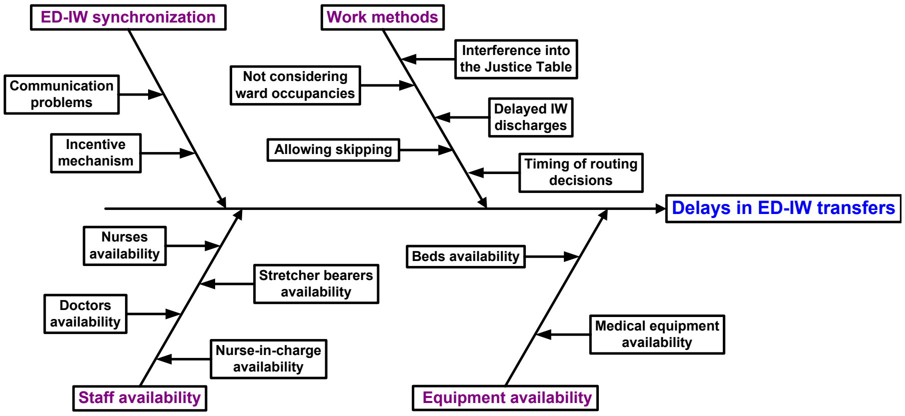


Fig 14: ED-to-IW delays—Causes and effects chart

is referred to as an *input-queued* scheme. Figure 15 depicts the two schemes.

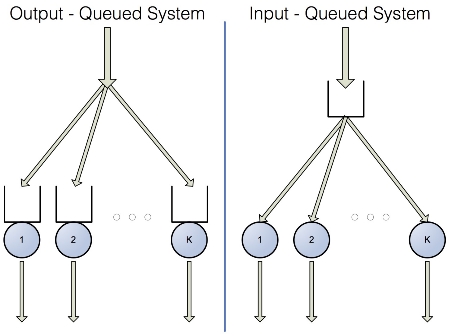


Fig 15: Output- vs. Input-queued scheme

Output-queued schemes are inherently less efficient than their input-queued counterparts, because the routing decision is made at an earlier time with less information. Moreover, the output-queued scheme is inequitable towards patients because FCFS is often violated.
The problem of customer routing in input-queued schemes has received considerable attention in the queueing literature (e.g. Armony (2005); Atar and Shwartz (2008); Gurvich and Whitt (2010); Mandelbaum and Stolyar (2004)). Similar issues in output-queued systems have been generally overlooked. Exceptions include Stolyar (2005) and

Tezcan (2008) who establish that the two systems have asymptotically similar performance, in both the conventional and the many-server heavy traffic regimes. This implies that inefficiencies, which arise in our ED-to-IW process due to the use of an output-queued scheme, become negligible in heavily 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. But how well does that insight translate to an environment such as a medical unit? This should be tested empirically; a first step was taken by Tseytlin and Zviran (2008).

2. *Not considering ward occupancies: The role of information availability in routing.* An additional important aspect of routing schemes, which directly affects patient delays, is the availability of information on the system state, at the moment of the routing decision. On the one hand, hospitals may base the routing on no information, namely use a static routing policy like round robin (surprisingly, our experience suggests that this is a prevalent policy). On the other extreme, a full information policy that takes into account current occupancy levels and projected future dismissals and transfers is feasible, if the information system is accurate and accommodating enough (See Chapter 8 in Hall (2012), which discusses bed management). It is important to understand the effect of information availability on system performance and fairness towards patients and medical staff.

5.5. *Fairness in the ED-to-IW process.* Transfer policies may have ramifications on fairness towards *customers* (patients) and towards *servers* (medical and nursing staff). We investigate both aspects next.

5.5.1. *Fairness towards patients.* In Section 5.4, we pointed out that output-queued schemes lead to diminished patient fairness, as FCFS order is often violated. (For references on the significance of FCFS in customer justice perception, see Mandelbaum, Momcilovic and Tseytlin (2012).) Indeed, our Rambam data indicate that 45% of the ED-to-IW transfer patients were "overtaken" by another patient (see Table 2). Moreover, more than a third of those were overtaken by at least three other patients. Although this figure includes overtaking between patient types, which may be due to clinical priorities, within each patient type there were significant FCFS violations as well. Specifically, 31% were actually overtaken by at least one patient of the same type, most of them not within the same ward, and hence these violations are conceivably due to the output-queued scheme.

While output-queues are inherently inefficient and unfair, they are un-

TABLE 2
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%	

likely to change in Rambam hospital due to the practical/clinical considerations described above, as well as psychological consideration (e.g., early ward assignment reduces uncertainty which in turn reduces anxiety for patients and their families). The use of output-queues in the ED-to-IW process illustrates some idiosyncrasies of flow control in healthcare.

5.5.2. **Research Opportunities**. A natural question is how to best maintain patient fairness in the output-queued scheme: What routing policies will keep the order close to FCFS? Is FCFS asymptotically maintained in heavy-traffic?

What other fairness criteria should be considered? Assuming that patients have preferences (clinical or prior experiences) for a specific ward, fairness may be defined with respect to the number of patients who are not assigned to their top priority. Related to this is the work of Thompson et al. (2009) that looks 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. One may alternatively consider achieving equity in terms of blocking probability (recall the discussion in §4.3.1) or patient delay. For the latter, Chan, Armony and Bambos (2011) show that such fairness may be achieved via Maximum Weighted Matching.

5.5.3. *Fairness towards staff.* In Section 5.4 we discussed the implications of the routing policy on delays in the ED-to-IW process; in addition, routing also has a significant impact on wards' workload. High workload tends to cause personnel burnout, especially if work allocation is perceived as unjust (references can be found in Armony and Ward (2010)). Rambam hospital takes fairness into consideration, as is implied from the name "Justice Table". However, is the patient allocation to the wards indeed fair?

There are many candidates for defining server "fairness". One natural measure is equity in the occupancy level. Since the number of nurses and doctors is typically proportional to the number of beds, equal occupancy levels imply that each nurse/doctor treats the same number of patients, on average. But does this imply that their workload is evenly distributed?

As mentioned in §4.2.1, staff workload in hospitals is not spread uniformly over a patient's stay, as patients admissions/discharges tend to be work intensive and treatment during the first days of a patient's hospitalization require much more time and effort from the staff than in the following days (Elkin and Rozenberg, 2007). Thus, one may consider an alternative fairness criterion: balancing the incoming load, or the "flux"—number of admitted patients per bed per time unit, among the wards. In Table 1 we observe that Ward B has a high average occupancy rate. In addition, as it is both the smallest and the "fastest" (shortest ALOS) ward, then (by Little's law) it has the highest flux among comparable IWs A–D. The workload of Ward B staff is hence the highest. We conclude that the most efficient ward is subject to the highest load—that is, patient allocation appears unfair towards ward staff.

Our data have already motivated some work on fair routing. *Analytical* results for *input-queued* systems were derived in Mandelbaum, Momcilovic and Tseytlin (2012), where both occupancy level and flux are taken into account with respect to fairness. Tseytlin and Zviran (2008) perform a *simulation* study of the *output-queued* system under various routing schemes. They propose a simulation-supported algorithm that balances a weighted function of occupancy and flux to achieve both fairness and short delays.

5.5.4. **Research Opportunities.** In the context of output-queued systems, a more rigorous analytical study is needed to formalize the conclusions of Tseytlin and Zviran (2008). Specifically, how to combine the occupancy and flux criteria into a single effective workload measure, which would be balanced across wards. Even in the context of input-queued systems, it is our view that Armony and Ward (2010); Mandelbaum, Momcilovic and Tseytlin (2012) and Ward and Armony (2013) have just taken the first steps towards staff fairness, as they do not fully account for the *dynamic* nature of workload in healthcare. As patients progress in their hospital stay, their medical needs change (mostly reduce) and the accuracy in which one can predict their LOS increases. This information could be very useful in successfully balancing workload.

The underlying definition of operational fairness, in our discussion thus far, proposed equal workload division across medical staff. A prerequisite for solving the "fairness problem" is then to define and calculate workload appropriately. However, we argue that such calculations must include not only direct time per resource but also emotional and cognitive efforts, as well as other relevant factors. For example, 1-minute of a standard chore does not compare with a 1-minute life-saving challenge (Plonski et al., 2013). Thus, the mix of medical conditions and patient severity should also be included in workload calculation. For the latter, it is not straightforward to determine whether wards would be inclined to admit the less severe patients (who add less workload, and potentially less emotional stress), as opposed to the more severe patients, who would challenge the medical staff, thus providing them with further learning and research opportunities; the latter is especially relevant in teaching hospitals such as Rambam.

5.6. *A system view.* In this Section we underscore the importance of looking at this network of ED, IWs and ED-to-IWs as a whole, as these three components are clearly interdependent. For concreteness, we focus on how the discharge policy in the IW affects ED-to-IW transfer times which, in turn, affect ED workload. We thereby argue that an integrative system view is appropriate.

It is natural to expect that the higher the occupancy in the IWs the longer the delays in transfer, due to limited IW resources. The left diagram in Figure 16 displays 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. The right diagram in Figure 16 shows 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 the daily resolution; other factors, such as the IW discharge process, also play a role.

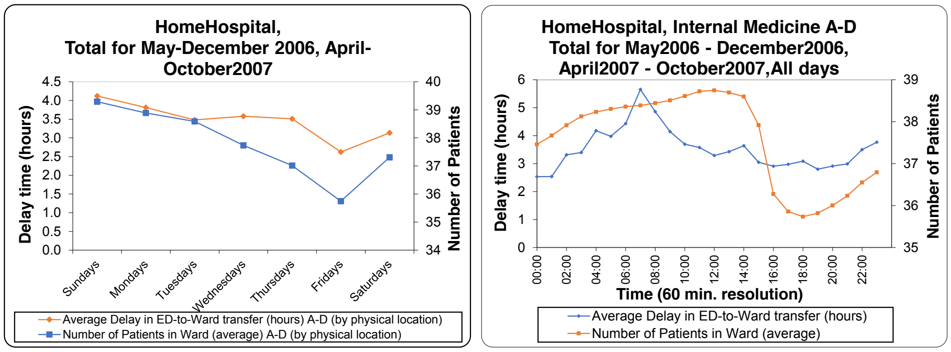


Fig 16: ED-to-IW transfer delays and number of patients in IW

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 note a consistent decline in the transfer delay up until noon. Patients assigned to the IWs during these times are admitted into the IWs between 1–3pm. 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 IW patients around 3–4pm when most of the IWs discharges are complete.

Further data analysis reveals that patients who are transferred to the IWs before 8am experience a significantly shorter LOS; Figure 17 shows that early hospitalization may reduce ALOS by more than 1 day. A correlation between hospitalization time and LOS was also reported by Earnest, Chen and Seow (2006): they observed that patients who are admitted in afternoon/night hours have ALOS that is longer than patients admitted in the morning. In contrast, we differentiate between early- and late-morning admissions. Regardless, in both cases, the plausible explanation for the difference in ALOS is the same: If patients are admitted to the ward early enough, the first day of treatment is more effective, as tests, medical procedures and treatments start earlier, and hence LOS is reduced. Thus, we argue that it is important to shorten the ED-to-IW transfer process and improve the IW admission process so that the first day of hospitalization is not "wasted". More on that in the research opportunities that follow.

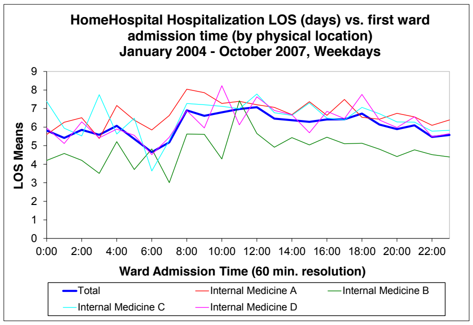


Fig 17: ALOS in IWs A through D, as a function of ward admission-time

In Section 5.3, we discussed how transfer delays impact physician workload in the ED and hence may influence quality of care there. Thus, we observe a chain of events in which the discharge policy in the IWs impacts the delays in transfer, which in turn affects workload in the ED. In particular, a system-view perspective is called for.

5.6.1. **Research opportunities**. Our discussion suggests that daily routines (schedules) in the IWs have significant impact on transfer delays and thereby on ED workload. At the same time, these routines also affect IW LOS. The question arises as to if and how one might wish to change these daily routines in view of these effects. The question fits well within a queueing context. The present daily routine at Rambam may be viewed as a priority scheme where currently-hospitalized IW patients enjoy priority during morning physicians' rounds; these patients become low-priority, as discharged patients obtain priority in the afternoon, then followed by newly-admitted patients. Is it possible to positively affect overall system performance by altering these priorities? (e.g. prioritizing newly-admitted and to-be discharged patient in the morning.) More broadly, the challenge falls within the uncharted territory of designing priority schemes for time-varying queueing networks.

Our discussion here brings us back to the broader issue—that is the need for a system view, in order to understand and relieve delays in patient flow. Consider, for example, the boarding patients in EDs (Figure 16) or in ICUs (Long and Mathews, 2012). Delays in transferring these boarding patients are often due to scarce resources or synchronization gaps (Zaied, 2011), which are rooted in parts of the system that differ from those where the delays are manifested. For example, scarce resources in the IWs exacerbate ED delays, and tardy processing of MRI results can prolong ICU LOS. It follows that a system view is required for the analysis of patient flow in hospitals.

When analyzing ED+IWs flows (§5), the wards operate naturally on a time-scale of days while the ED time scale is hours. Wards thus serve as a random environment for the ED (Ramakrishnan, Sier and Taylor, 2005). Figure 9 (§4.2) reveals that the hourly scale is also of interest for IWs. These empirical observations arise in a service system (hospital) that evolves in multiple time scales, which are all natural and relevant for measuring and modeling its performance. The mathematical manifestation of such scales is asymptotic analysis that highlights what matters at each scale, while averaging out details that are deemed insignificant (e.g., Mandelbaum, Momcilovic and Tseytlin (2012), Shi et al. (2012), Gurvich and Perry (2012) and Zacharias and Armony (2013)).

6. Discussion and concluding remarks. We have described research opportunities that arose from EDA of operation patient flow data. We now discuss the relationship between operational performance measures and overall hospital performance, and conclude with comments on data-based OR research.

6.1. *Operational measures as surrogates to overall hospital performance, or queueing models reach beyond the operational.*

Hospital performance is measured across a variety of dimensions: clinical, financial, operational, psychological (patient satisfaction) and societal. The most important measures are clearly *clinical* but, practically, *operational* performance is the easiest to quantify, measure, track and react upon in real time. Moreover, operational performance is tightly coupled with the other dimensions (e.g. rate of readmissions with quality of clinical care, or LOS and LWBS with financial performance), which explains its choice as a "language" that captures overall hospital performance.

Operational performance measures are often associated with patient flow. Among these, we discussed LWBS (Section 3) and "blocking" (where patients end up being hospitalized in a ward different from that which is medically best for them - Section 4.3.1); boarding (transfer) time from the ED to the appropriate medical unit; and measures related to LOS, in the ED or IWs, such as merely averages (or medians), or fractions staying beyond a desired threshold. Other measures that have not been mentioned require intra-ward data, which is beyond our data granularity. Examples include the time until triage or until a patient is first seen by an ED physician (Zeltyn et al., 2011), the number of visits to a physician during an ED sojourn (Huang, Carmeli and Mandelbaum, 2011) and the time-ingredients of an ED visit (treatment and waiting—for a resource, for synchronization or for a treatment to take its effect; see Zaied (2011) and Atar, Mandelbaum and Zviran (2012)).

6.1.1. *Readmissions.*

As already indicated in Sections 3.1 and 4.1, our data supports the analysis of readmissions (Mandelbaum et al., 2013). We now elaborate on this operational measure of performance since policy makers are increasingly focusing on, as part of efforts to extend quality of care measures from within-hospital processes to after-hospital short-term outcomes (Medicare USA, 2013). As mentioned, the likelihood of readmission to the hospital, within a relatively short time, is a natural indirect measure for quality of care (similarly to first-call-resolution rates in call centers). Consequently, readmission rates are accounted for when profiling hospitals' quality and determining reimbursements for their services.

One should consider readmissions judiciously as some of them could be due to factors outside the hospital control, or they may be an integral part of the treatment regiment. For example, returns within a few months to chemotherapy are typically *planned* and are unrelated to poor quality. But there are also unplanned chemotherapy returns after 1–2 weeks, which arise from complications after treatment. To properly incorporate readmissions in a queueing model (such as in Yom-Tov and Mandelbaum (2014)) one should distinguish between these two readmission types by, for example, modeling planned (unplanned) readmissions as deterministic (stochastic) returns. Also note that readmissions should be measured in their natural time-scale. For example, readmission to an ED should be measured in a time scale of days-weeks, while readmissions to an IW have a natural time-scale of weeks-months.

6.1.2. *Capacity and Cost of Care.* Of utter importance to hospital managers and policy makers is hospital costing. Kaplan and Porter (2011) argue that the mapping of patient / process flow, and the association of its activities with their supporting resources, should constitute the first step in understanding the cost of care. This is nothing but promoting a queueing-network view for understanding and calculating cost-of-care. Indeed, (Kaplan and Porter, 2011) further submit that most hospital costs are mistakenly judged as fixed while they ought to be viewed as variable costs; it follows that the corresponding resource levels are flexible, an observation that renders controllable most resources in a hospital.

This viewpoint naturally connects with the distinction between static and dynamic capacity, which we now explain. Capacity of a hospital or a ward is commonly expressed in terms of the number of beds (or rooms, or physical space). However, 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. (Note that bed capacity plays the dual role of static capacity—capping the number of patients that can be simultaneously hospitalized, and dynamic capacity—serving as a proxy for the processing capacity of medical personnel.). And this distinction connects back to costs 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 its level (and hence also cost) is *flexible (controllable)*.

6.2. *Some concluding comments on data-based research—a great opportunity but no less of a challenge.* The goal of the present work has been two-fold: first, to encourage and strengthen, through data and its EDA, the natural link between queueing theory and its application to patient flow in healthcare; and second, facilitate data-based learning for researchers who seek to reinforce this important link.

While theory has been the comfort zone of Operations Research (OR) and Applied Probability (AP), the situation dramatically differs when (big) data is brought into the picture. Fundamental changes are therefore essential—both within our OR/AP community as well as our potential healthcare partners: changes in accessibility to healthcare data, in education (e.g. concerning the necessity of data-based OR research, importance and need to publish EDA, benefits of research reproducibility) and in funding priorities (e.g. for developing and sustaining the infra-structure that is a prerequisite for a research such as the one reported here.)

But we are cautiously optimistic. Indeed, comprehensive data-collection is becoming increasingly feasible, systematic and cheaper, for example via Real-time Location Systems (RTLS), which will ultimately integrate with Personal-Health and Financial Records. This will enable partnerships with providers of healthcare services, that are based on multidisciplinary (clinical, operational, financial, psychological) tracking of complete care-paths. Also, tracking resolution and scope will be at the level of the individual patient and provider while covering the full cycle of care. The process of data-based OR research in hospitals is thus only beginning.

REFERENCES

Aksin, O. Z., Karaesmen, F. and Ormeci, E. L. (2007). A Review of Workforce Cross-Training in Call Centers from an Operations Management Perspective. In *Workforce Cross Training Handbook* (D. Nembhard, ed.) CRC Press. 4.3.3

Allon, G., Deo, S. and Lin, W. (2013). Impact of Size and Occupancy of Hospital on the Extent of Ambulance Diversion: Theory and Evidence. *Operations Research* **61** 544-562. 3.3.3

Armony, M. (2005). Dynamic Routing in Large-Scale Service Systems with Heterogeneous Servers. *Queueing Systems* **51** 287–329. 1

Armony, M., Chan, C. W. and Zhu, B. (2013). Critical Care in Hospitals: When to Introduce a Step Down Unit? Working paper, Columbia University. 5.2

Armony, M. and Ward, A. (2010). Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems. *Operations Research* **58** 624–637. 5.5.3, 5.5.4

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y. and Yom-Tov, G. B. (2013). Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective. An Extended Version (EV). 1.3, 1.4, 3.2.2, 4.2.1, 4.3.1, 4.3.4, 5.1, 2

Atar, R., Mandelbaum, A. and Zviran, A. (2012). Control of Fork-Join Networks in Heavy Traffic. Allerton Conference,. 5.2.1, 6.1

Atar, R. and Shwartz, A. (2008). Efficient Routing in Heavy Traffic under Partial Sampling of Service Times. *Mathematics of Operations Research* **33** 899–909. 1

Balasubramanian, H., Muriel, A. and Wang, L. (2012). The Impact of Flexibility and Capacity Allocation on the Performance of Primary Care Practices. *Flexible Services and Manufacturing Journal* **24** 422–447. 4.3.3

Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., Wood, D. and Stahl, J. (2010). Improving clinical access and continuity using physician panel redesign. *Journal of General Internal Medicine* **25** 1109–1115. 4.3.3

Baron, O., Berman, O., Krass, D. and Wang, J. (2014). Using Strategic Idleness to Improve Customer Service Experience in Service Networks. *Operations Research* **62** 123–140. 2

Batt, R. J. and Terwiesch, C. (2012). Doctors Under Load: An Empirical Study of State Dependent Service Times in Emergency Care. Working paper. 3.2.3

Bekker, R. and de Bruin, A. M. (2010). Time-Dependent Analysis for Refused Admissions in Clinical Wards. *Annals of Operations Research* **178** 45–65. 4.2

Bernstein, S. L., Verghese, V., Leung, W., Lunney, A. T. and Perez, I. (2003). Development and Validation of a New Index to Measure Emergency Department Crowding. *Academic Emergency Medicine* **10** 938–942. 3.3.2

Bertsimas, D. and Mourtzinou, G. (1997). Transient Laws of Non-stationary Queueing Systems and their Applications. *Queueing Systems* **25** 115–155. 3.2.1

Brandeau, M. L., Sainfort, F. and Pierskalla, W. P., eds. (2004). *Operations Research and Health Care: A Handbook of Methods and Applications*. Kluwer Academic Publishers, London. 2

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005). Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *Journal of the American Statistical Association* **100** 36–50. 3.3.2, 4.2.1

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodal Inference: a Practical Information-Theoretic Approach, 2nd Edition*. Springer. 3.3.1

Burström, L., Nordberg, M., Ornung, G., Castrén, M., Wiklund, T., Engström, M. L. and Enlund, M. (2012). Physician-Led Team Triage Based on Lean Principles May be Superior for Efficiency and Quality? A Comparison of Three Emergency Departments with Different Triage Models. *Scandinavian Journal Trauma, Resuscitation, and Emergency Medicine* **20** 20–57. 3.3.3

Canadadian-Triage, Admission of Paitents to over-capacity inpatient beds. Appendix A, `http://www.calgaryhealthregion.ca/policy/docs/1451/Admission_over-capacity_AppendixA.pdf`. 3.3.2

Chalfin, D. B., Trzeciak, S., Likourezos, A., Baumann, B. M. and Dellinger, R. P. (2007). Impact of Delayed Transfer of Critically Ill Patients from the Emergency Department to the Intensive Care Unit. *Critical Care Medicine* **35** 1477–1483. 5.3

Chan, C., Armony, M. and Bambos, N. (2011). Fairness in Overloaded Parallel Queues. Working paper. 5.5.2

Chan, C., Farias, V. and Escobar, G. (2014). The Impact of Delays on Service Times in the Intensive Care Unit. Working paper. 2

Chan, C., Yom-Tov, G. B. and Escobar, G. (2014). When to use Speedup: An Examination of Service Systems with Returns. *Operations Research* **62** 462?482. 2

Chao, X., Miyazawa, M. and Pinedo, M. (1999). *Queueing Networks: Customers, Signals and Product Form Solutions*. Wiley. 1

Chen, C., Jia, Z. and Varaiya, P. (2001). Causes and Cures of Highway Congestion. *Control Systems, IEEE* **21** 26–33. 3.2.3

Chen, H. and Yao, D. D. (2001). *Fundamentals of Queuing Networks: Performance, Asymptotics, and Optimization*. Springer. 4.3

Cooper, A. B., Litvak, E., Long, M. C. and McManus, M. L. (2001). Emergency Department Diversion: Causes and Solutions. *Academic Emergency Medicine* **8** 1108–1110. 2

de Bruin, A. M., van Rossum, A. C., Visser, M. C. and Koole, G. M. (2007). Modeling the Emergency Cardiac In-Patient Flow: An Application of Queuing Theory. *Health Care Management Science* **10** 125–137. 2

de Bruin, A. M., Bekker, R., van Zanten, L. and Koole, G. M. (2009). Dimensioning Hospital Wards using the Erlang Loss Model. *Annals of Operations Research* **178** 23–43. 2, 3, 4.2, 4.3.1

Denton, B. T., ed. (2013). *Handbook of Healthcare Operations Management: Methods and Applications*. Springer. 2

Dobson, G., Tezcan, T. and Tilson, V. (2013). Optimal Workflow Decisions for Investigators in Systems with Interruptions. *Management Science*. Forthcoming. 3.3.3

Dong, J. and Whitt, W. (2014). On fitted birth-and-death queue models. Working paper, Columbia University. 3.3.1

Donoho, D. L., Maleki, A., Shahram, M., Rahman, I. U. and Stodden, V. (2009). Reproducible Research in Computational Harmonic Analysis. *IEEE Computing in Science & Engineering* **11** 8–18. 6.2

Earnest, A., Chen, M. and Seow, E. (2006). Exploring if day and time of admission is associated with average length of stay among inpatients from a tertiary hospital in Singapore: an analytic study based on routine admission data. *BMC Health Services Research* **6** 6. 5.6

Elkin, K. and Rozenberg, N. (2007). Patients Flow from the Emergency Department to the Internal Wards. IE&M project, Technion (In Hebrew). 4.1, 5.5.3

Feldman, Z., Mandelbaum, A., Massey, W. A. and Whitt, W. (2008). Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science* **54** 324–338. 3.2.1

Froehle, C. M. and Magazine, M. J. (2013). Improving scheduling and flow in complex outpatient clinics. In *Handbook of healthcare operations management: methods and applications* (B. T. Denton, ed.) 9 229–307. Springer. 2

Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review and Research Prospects. *Manufactoring, Services and Operations Management* **5** 79–141. 4.3

Gerla, M. and Kleinrock, L. (1980). Flow Control: A Comparative Survey. *IEEE Transactions on Communcations* **28** 553–574. 3.2.3

Green, L. (2004). Capacity Planning and Management in Hospitals. In *Operations Research and Health Care: A Handbook of Methods and Applications* (M. L. Brandeau, F. Sainfort and W. P. Pierskalla, eds.) 14–41. Kluwer Academic Publishers, London. 2

Green, L. V. (2008). Using Operations Research to Reduce Delays for Healthcare. In *Tutorials in Operations Research* (Z.-L. Chen and S. Raghavan, eds.) 1–16. INFORMS. 2

Green, L. V., Kolesar, P. J. and Whitt, W. (2007). Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. *Production and Operations Management* **16** 13–39. 2, 3.2.1, 4.2.1

Green, L. and Yankovic, N. (2011). Identifying Good Nursing Levels: A Queuing Approach. *Operations Research* **59** 942–955. 2, 4.2.1

Green, L., Soares, J., Giglio, J. F. and Green, R. A. (2006). Using Queuing Theory to Increase the Effectiveness of Emergency Department Provider Staffing. *Academic Emergency Medicine* **13** 61–68. 3

Gurvich, I. and Perry, O. (2012). Overflow Networks: Approximations and Implications to Call-Center Outsourcing. *Operations Research* **60** 996–1009. 5.6.1

Gurvich, I. and Whitt, W. (2010). Service-Level Differentiation in Many-Server Service Systems via Queue-Ratio Routing. *Operations Research* **58** 316–328. 1

Hagtvedt, R., Ferguson, M., Griffin, P., Jones, G. T. and Keskinocak, P. (2009). Cooperative Strategies To Reduce Ambulance Diversion. *Proceedings of the 2009 Winter Simulation Conference* **266** 1085–1090. 3.3.3

Hall, R. W., ed. (2012). *Handbook of Healthcare System Scheduling*. Springer. 2, 2

Hall, R. W., ed. (2013). *Patient Flow: Reducing Delay in Healthcare Delivery*. Springer 2nd edition. 2

Hall, R., Belson, D., Murali, P. and Dessouky, M. (2006). Modeling Patient Flows Through the Healthcare System. In *Patient Flow: Reducing Delay in Healthcare Delivery* (R. W. Hall, ed.) 1 1–45. Springer. 2, 3

Hoot, N. R., Zhou, C., Jones, I. and Aronsky, D. (2007). Measuring and Forecasting Emergency Department Crowding in Real Time. *Annals of Emergency Medicine* **49** 747–755. 3.3.2

Huang, J. (2013). Patient Flow Management in Emergency Departments PhD thesis, National University of Singapore (NUS). 2.1

Huang, J., Carmeli, B. and Mandelbaum, A. (2011). Control of Patient Flow in Emergency Departments: Multiclass Queues with Feedback and Deadlines. Working paper. 2.1, 3.3.1, 3.3.2, 3.3.3, 6.1

Hwang, U., McCarthy, M. L., Aronsky, D., Asplin, B., Crane, P. W., Craven, C. K., Epstein, S. K., Fee, C., Handel, D. A., Pines, J. M., Rathlev, N. K., Schafermeyer, R. W., Zwemer, F. L. and Bernstein, S. L. (2011). Measures of Crowding in the Emergency Department: A Systematic Review. *Academic Emergency Medicine* **18** 527–538. 3.3.2

IHI, (2011). Patient First: Efficient Patient Flow Management Impact on the ED. *Institute for healthcare improvement*. http://www.ihi.org/knowledge/Pages/ImprovementStories/PatientFirstEfficientPatientFlowManagementED.aspx. 3

Janssen, A. J. E. M., van Leeuwaarden, J. S. H. and Zwart, B. (2011). Refining Square-Root Safety Staffing by Expanding Erlang C. *Operations Research* **56** 1512–1522. 4.3

JCAHO, (2004). JCAHO Requirement: New Leadership Standard on Managing Patient Flow for Hospitals. *Joint Commission Perspectives* **24** 13–14. 1.1

Jennings, O. B. and de Véricourt, F. (2008). Dimensioning Large-Scale Membership Services. *Operations Research* **56** 173–187. 2

Jennings, O. B. and de Véricourt, F. (2011). Nurse Staffing in Medical Units: A Queueing Perspective. *Operations Research* **59** 1320–1331. 2, 4.2.1, 4.3

Jouini, O., Dallery, Y. and Aksin, O. Z. (2009). Queueing Models for Full-Flexible Multi-class Call Centers with Real-Time Anticipated Delays. *International Journal of Production Economics* **120** 389–399. 4.3.3

Kaplan, R. S. and Porter, M. E. (2011). How to Solve the Cost Crisis in Health Care. *Harvard Business Review* **89** 46–64. 6.1.2

Karr, A. F. (2009). Secure Statistical Analysis of Distributed Databases, Emphasizing What We Don't Know. *Journal of Privacy and Confidentiality* **1**. http://repository.cmu.edu/jpc/vol1/iss2/5. 6.2

Kc, D. and Terwiesch, C. (2009). Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations. *Management Science* **55** 1486–1498. 2, 3.2.3

Kelly, F. P. (1979). *Markov Processes and Reversibility*. Wiley. 3.2.2

Kim, S. H. and Whitt, W. (2014). Are Call Center and Hospital Arrivals Well Modeled by Nonhomogeneous Poisson Processes? Forthcoming at M&SOM. 3.2.1

Leite, S. C. and Fragoso, M. D. (2013). Diffusion Approximation for Signaling Stochastic Networks. *Stochastic Processes and their Applications* **123** 2957–2982. 5.2.1

Long, E. F. and Mathews, K. M. (2012). "Patients Without Patience": A Priority Queuing Simulation Model of the Intensive Care Unit. Working paper. 5.3, 5.6.1

Maa, J. (2011). The Waits that Matter. *The New England Journal of Medicine*. 5.3

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. 2.1, 3.2.1

Maman, S., Zeltyn, S. and Mandelbaum, A. (2011). Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments. Working paper. 2.1

Mandelbaum, A., Momcilovic, P. and Tseytlin, Y. (2012). On Fair Routing From Emergency Departments to Hospital Wards: QED Queues with Heterogeneous Servers. *Management Science* **58** 1273–1291. 2.1, 4.2.1, 5.5.1, 5.5.3, 5.5.4, 5.6.1

Mandelbaum, A. and Stolyar, S. (2004). Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized $c\mu$-Rule. *Operations Research* **52** 836–855. 1

Mandelbaum, A., Trofimov, V., Gavako, I. and Nadjhahrov, E. (2013). Home-Hospital (Rambam): Readmission Analysis. http://seeserver.iem.technion.ac.il/databases/Docs/HomeHospital_visits_return.pdf. 6.1.1

Marmor, Y. N. (2003). Developing a Simulation Tool for Analyzing Emergency Department Performance. Master's thesis, Technion—Israel Institute of Technology. 1.3

Marmor, Y. N. (2010). Emergency-Departments Simulation in Support of Service-Engineering: Staffing, Design, and Real-Time Tracking. PhD thesis, Technion—Israel Institute of Technology. 2.1

Marmor, Y. N., Golany, B., Israelit, S. and Mandelbaum, A. (2012). Designing Patient Flow in Emergency Departments. *IIE Transactions on Healthcare Systems Engineering* **2** 233–247. 2.1, 3.3.1, 3.3.3

Marmor, Y. N., Rohleder, T., Cook, D., Huschka, T. and Thompson, J. (2013). Recovery bed planning in cardiovascular surgery: a simulation case study. *Health Care Management Science* **16** 314–327. 3.2.3

McHugh, M., Van Dyke, K., McClelland, M. and Moss, D. (2011). Improving Patient Flow and Reducing Emergency Department Crowding. *Agency for healthcare research and quality*. http://www.ahrq.gov/research/findings/final-reports/ptflow/index.html. 3

Medicare USA, (2013). Hospital Compare: 30-Day Death and Readmission Measures Data. http://www.medicare.gov/HospitalCompare/Data/RCD/30-day-measures.aspx. 6.1.1

Nadjhahrov, E., Trofimov, V., Gavako, I. and Mandelbaum, A. (2013). Home-Hospital (Rambam): EDA via SEEStat 3.0 to Reproduce "On Patients Flow in Hospitals". http://ie.technion.ac.il/Labs/Serveng/files/HHD/reproducing_flow_paper.pdf. 6.2, 6.2

Nestler, S. (2011). Reproducible (Operations) Research: A Primer on Reproducible Research and Why the O.R. Community Should Care About it. *ORMS Today* **38**. 6.2

Nguyen, V. (1994). The Trouble with Diversity: Fork-Join Networks with Heterogeneous Customer Population. *The Annals of Applied Probability* 1–25. 5.2.1

Oredsson, S., Jonsson, H., Rognes, J., Lind, L., Göransson, K. E., Ehrenberg, A., Asplund, K., Castrén, M. and Farrohknia, N. (2011). A Systematic Review of Triage-Related Interventions to Improve Patient Flow in Emergency Departments. *Scandinavian Journal Trauma, Resuscitation, and Emergency Medicine* **July 19** 19–43. 3.3.3

Plonski, O., Efrat, D., Dorban, A., David, N., Gologorsky, M., Zaied, I., Mandelbaum, A. and Rafaeli, A. (2013). Fairness in Patient Routing: Maternity Ward in Rambam Hospital. Technical report. 4.2.1, 5.5.4

Ramakrishnan, M., Sier, D. and Taylor, P. G. (2005). A Two-Time-Scale Model for Hospital Patient Flow. *IMA Journal of Management Mathematics* **16** 197–215. 5.6.1

Rambam, Rambam Health Care Campus, Haifa, Israel. http://www.rambam.org.il/Home+Page/. 1.2

RambamData, Rambam Hospital data repositories. Technion SEELab, http://seeserver.iem.technion.ac.il/databases/. 6.2, 6.2

Saghafian, S., Austin, G. and Traub, S. J. (2014). Operations research contributions to emergency department patient flow optimization: review and research prospects. Working paper. 3

SEELab, SEE Lab, Technion—Israel Institute of Technology. http://ie.technion.ac.il/Labs/Serveng/. 1, 1, 6.2, 6.2

SEEServer, Server of the Center for Service Enterprise Engineering. http://seeserver.iem.technion.ac.il/see-terminal/. 6.2

SEEStat, SEEStat documentation, Technion—Israel Institute of Technology. http://ie.technion.ac.il/Labs/Serveng/. 1, 6.2, 6.2

Shi, P., Chou, M. C., Dai, J. G., Ding, D. and Sim, J. (2012). Hopital Inpatient Operations: Mathematical Models and Managerial Insights. Working paper. 1.3, 2, 4.2, 4.2.1, 5.6.1

Stolyar, S. (2005). Optimal Routing in Output-Queued Flexible Server Systems. *Probability in the Engineering and Informational Sciences* **19** 141–189. 1, 1

Sullivan, S. E. and Baghat, R. S. (1992). Organizational Stress, Job Satisfaction, and Job Performance: Where Do We Go from Here? *Journal of Management* **18** 353–375. 3.2.3

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer. 3.3.2

Tezcan, T. (2008). Optimal Control of Distributed Parallel Server Systems under the Halfin and Whitt Regime. *Math of Operations Research* **33** 51–90. 1

Thompson, S., Nunez, M., Garfinkel, R. and Dean, M. D. (2009). Efficient Short-Term Allocation and Reallocation of Patients to Floors of a Hospital During Demand Surges. *Operations Research* **57** 261–273. 5.5.2

Thorin, O. (1977). On the Infinite Divisibility of the Lognormal Distribution. *Scandinavian Actuarial Journal* **1977** 121–148. 4.2.1

Tseytlin, Y. (2009). Queueing Systems with Heterogeneous Servers: On Fair Routing of Patients in Emergency Departments. Master's thesis, Technion—Israel Institute of Technology. 1.3, 2.1

Tseytlin, Y. and Zviran, A. (2008). Simulation of Patients Routing from an Emergency Department to Internal Wards in Rambam Hospital. OR Graduate Project, IE&M, Technion. 1, 5.5.3, 5.5.4

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison Wesley. 1.3

Ward, A. and Armony, M. (2013). Blind Fair Routing in Large-Scale Service Systems with Heterogeneous Customers and Servers. *Operations Research* **61** 228–243. 5.5.4

Whitt, W. (2012). Fitting Birth-and-Death Queueing Models to Data. *Statistics and Probability Letters* **82** 998–1004. 3.2.2, 3.3.1

Yom-Tov, G. B. (2010). Queues in Hospitals: Queueing Networks with ReEntering Customers in the QED Regime. PhD thesis, Technion—Israel Institute of Technology. 2.1, 4.2.1

Yom-Tov, G. B. and Mandelbaum, A. (2014). Erlang-R: A Time-Varying Queue with Reentrant Customers, in Support of Healthcare Staffing. *M&SOM* **16** 283–299. 2.1, 3.2.1, 3.2.2, 3.3.1, 3.3.3, 4.3, 6.1.1

Zacharias, C. and Armony, M. (2013). Joint Panel Sizing and Appointment Scheduling in Outpatient Care. Working paper. 5.6.1

Zaied, I. (2011). The Offered Load in Fork-Join Networks: Calculations and Applications to Service Engineering of Emergency Department Master's thesis, Technion—Israel Institute of Technology. 5.6.1, 6.1

Zalenski, R. J., Rydman, R. J., Ting, S., Kampe, L. and Selker, H. P. (1998). A National Survey of Emergency Department Chest Pain Centers in the United States. *The American Journal of Cardiology* **81** 1305–1309. 3.3.3

Zeltyn, S., Marmor, Y. N., Mandelbaum, A., Carmeli, B., Greenshpan, O., Mesika, Y., Wasserkrug, S., Vortman, P., Schwartz, D., Moskovitch, K., Tzafrir, S., Basis, F., Shtub, A. and Lauterman, T. (2011). Simulation-Based Models of Emergency Departments: Real-Time Control, Operations Planning and Scenario Analysis. *Transactions on Modeling and Computer Simulation (TOMACS)* **21**. 2.1, 3, 3.3.1, 3.3.3, 6.1

Acknowledgements. We dedicate our paper to the memory of the late David Sinreich. As a professor at the 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 led jointly by IBM Haifa Research (headed by Oded Cohn), Rambam Hospital (Director Rafi Beyar) and Technion IE&M Faculty (Former Dean Boaz Golany).

Data analysis, maintenance of data repositories, and continuous advice and support have been cheerfully provided by the Technion SEE Laboratory and its affiliates: Valery Trofimov, Igor Gavako, Ella Nadjharov, Shimrit Maman, Katya Kutsy, Nitzan Carmeli, and Arik Senderovic.

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 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. The joint research of MA and AM was funded by BSF (Bi-national Science Foundation) Grants 2005175/2008480. AM was partially supported by ISF Grant 1357/08 and by the Technion funds for promotion of research and sponsored research. MA was partially funded by the Lady Davis Fellowship as a visitor at the Technion IE&M faculty.

Some of AM's research was funded by and carried out while visiting the Statistics and Applied Mathematical Sciences Institute (SAMSI) of the NSF; the Department of Statistics and Operations Research (STOR), the University of North Carolina at Chapel Hill; the Department of Information, Operations and Management Sciences (IOMS), Leonard N. Stern School of Business, New York University; and the Department of Statistics, The Wharton School, University of Pennsylvania – the hospitality of all these institutions is acknowledged and truly appreciated.

Our paper greatly benefited from feedback by colleagues and a thorough and insightful refereeing process. Ward Whitt discovered a flaw in our empirical analysis that led to the rewriting of the ED section. Jim Dai read carefully the first version and provided significant and helpful editorial feedback. Last but certainly not least, we are grateful to the editor of *Stochastic Systems*, Peter Glynn, for leading and guiding us patiently and safely through the revision process.

Appendix 1: A framework for OR/AP data-based research. The traditional still prevalent model for data-based OR/AP research has been one where an *individual* researcher, or a small group, obtains and analyzes data for the sake of an *isolated* research project. Our experience is that such a model cannot address today's empirical needs. For example, hospital data is typically large, complex, contaminated and incomplete, which calls for a professional inevitably time-consuming treatment. Next, using data in a single project, or a few for that matter, is wasteful—on the other hand, data-reuse and sharing, across student generations or research groups, requires infrastructure, documentation, maintenance and coordination. Finally, healthcare data is often confidential and proprietary, and that prevents reproducibility and slows down progress.

Towards a culture of reproducible research in empirical OR/AP. Data-based OR/AP research must strive for reproducibility of research outcomes—a fundamental principle in the traditional sciences. Reproducibility enables scrutiny of analysis and recommendations. This yields credibility and trust, which is an absolute prerequisite for influencing hospital practices.

Reproducible (Operations) Research is discussed in Nestler (2011), which is also a source for additional references and links. There have been some systematic attempts to establish a reproducibility culture in research (Donoho et al., 2009). It ought to start with funding agencies and journal policies: e.g. the Editorial Statement of the Finance Department in *Management Science* reads: "Authors of empirical and quantitative papers should provide or make available enough information and data so that the results are reproducible." It can advance with research such as Karr (2009), that aims at statistical analysis of distributed (unsharable) databases (e.g. hospital data); it will ideally culminate in a multitude of research labs, each providing free access to its data and serving its own research community and beyond.

A feasible model. A model for such a lab is the Technion SEELab, where readers can access RambamData. Little effort will be then required to reproduce our present EDA and going beyond it. In fact, most of our figures were created by SEEStat—a SEELab-developed user-friendly platform for online (real-time) EDA—and readers can recreate this process by following Nadjhahrov et al. (2013).

Appendix 2: Accessing data repositories and EDA tools at the SEELab. SEELab is a data-based research laboratory, residing at the IE&M Faculty of the Technion in Haifa, Israel. (SEE stands for "Service Enterprise Engineering".) SEELab maintains a repository for transaction-level operational data (log-files) from large service operations. This data is collected and cleaned, thus preparing it for research and teaching. Currently, SEELab databases include call-by-call multi-year data from 4 call centers, an internet academic website, 8 emergency departments (mainly their arrivals data) and 4 years of data from the Rambam Hospital—the latter is the empirical foundation for the present paper.

The EDA environment of SEELab is SEEStat—a software platform that enables real-time statistical analysis of service data at seconds-to-months time resolutions. SEEStat was used to create most of our figures. It implements many statistical algorithms: parametric distribution fitting and selection, fitting of distribution mixtures, survival analysis and more—with all algorithms interacting seamlessly with all the databases. SEEStat also interacts with SEEGraph, a pilot-environment for structure-mining, on-demand creation, display and animation of data-based process maps (e.g. Figure 1, and the animation of its underlying data).

Three SEELab data-bases are publicly accessible at the SEELab server SEEServer: two from call centers and one from the Rambam hospital. For example, data from a U.S. banking call center covers the operational history of close to 220 million calls, over close to 3 years; 40 million of these calls were served by (up to 1000) agents and the rest by a VRU (answering machine). The Rambam data is described in §1.2. Our analysis of the current data greatly benefitted from our call-center experience. Moreover, the completeness of call-center data provides an ideal to strive for, with the typically partial hospital data - this gap is now narrowing with the increasing prevalence of real-time locating systems (RTLS) data.

SEEStat Online: The connection protocol to SEEStat, for any research or teaching purpose, is simply as follows: go to the SEELab webpage
`http://ie.technion.ac.il/Labs/Serveng`;
then proceed, either via the link **SEEStat Online**, or directly through
`http://seeserver.iem.technion.ac.il/see-terminal`, and complete the registration procedure. Within a day or so, you will receive a confirmation of your registration, plus a password that allows you access to SEEStat, SEELab's EDA environment, and via SEEStat to the above-mentioned databases. Note that your confirmation email includes two attachments: a trouble-shooting document and a self-taught tutorial that is based on call

center data and the Rambam hospital data. We propose that you print out the tutorial, connect to SEEStat and then let the tutorial guide you, hands-on, through SEEStat basics—this should take no more than 1.5 hours.

On data cleaning and maintenance. There were plenty of records that were flawed due to archiving or simply 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 led to the discovery of a high fraction of "transfers" from the ED to a virtual ward, all occurring precisely at 11:59pm; subsequent analysis managed to associate 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 the tip of the iceberg.

Reproducing our EDA and beyond. Rambam data is publicly available, either for downloading (RambamData consists of records per individual customers) or through SEEStat, as discribed above. The download link includes data documentation. To facilitate reproducibility, the document Nadjhahrov et al. (2013) provides a detailed description of the creation process of our EDA, which includes all figures (except for Figure 12) in the present paper.

Mor Armony
Stern School of Business, NYU
44 West 4th Street
New York, NY 10012
E-mail: marmony@stern.nyu.edu

Shlomo Israelit
Director, Emergency Trauma Department
Rambam Health Care Campus (RHCC)
6 Ha'Aliya Street
Haifa, Israel 31096
E-mail: s_israelit@rambam.health.gov.il

Avishai Mandelbaum
Faculty of Industrial Engineering and Management
Technion—Israel Institute of Technology
Technion city, Haifa, Israel, 32000
E-mail: avim@ie.technion.ac.il

Yariv N. Marmor
Department of Industrial Engineering and Management
ORT Braude College
Karmiel, Israel
Health Care Policy and Research Department
Mayo Clinic
200 First Street SW
Rochester, MN, USA, 55905
E-mail: myariv@braude.ac.il

Yulia Tseytlin
IBM Haifa Research Lab
Haifa University Campus
Mount Carmel, Haifa, Israel, 31905
E-mail: yuliatse@gmail.com

Galit B. Yom-Tov
Faculty of Industrial Engineering and Management
Technion—Israel Institute of Technology
Technion city, Haifa, Israel, 32000
E-mail: gality@tx.technion.ac.il