Accepted Manuscript

Needy boarding patients in emergency departments: An exploratory case study using discrete-event simulation

Kim De Boeck, Raïsa Carmen, Nico Vandaele

PII: S2211-6923(18)30009-2
DOI: <https://doi.org/10.1016/j.orhc.2019.02.002>
Reference: ORHC 180

To appear in: *Operations Research for Health Care*

Received date : 17 January 2018
Revised date : 19 February 2019
Accepted date : 22 February 2019

Please cite this article as: K.D. Boeck, R. Carmen and N. Vandaele, Needy boarding patients in emergency departments: An exploratory case study using discrete-event simulation, *Operations Research for Health Care* (2019), <https://doi.org/10.1016/j.orhc.2019.02.002>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Needy boarding patients in emergency departments: An exploratory case study using discrete-event simulation

Kim De Boeck¹, Raïsa Carmen, Nico Vandaele

Research Centre for Operations Management,

Faculty of Economics and Business,

KU Leuven, Naamsestraat 69, Leuven, Belgium

E-mail: kim.deboeck@kuleuven.be, raisa.carmen@kuleuven.be, nico.vandaele@kuleuven.be

Abstract. Boarding patients and the extra workload they introduce are a major concern in emergency departments. Currently, research on boarding has focused on the beds that these patients occupy while they wait for admission in the inpatient ward. Little attention has been given to the check-ups these patients will inevitably need while waiting. These check-ups also confront the physicians with a challenging task: prioritizing between boarding patients and patients currently under treatment in the emergency department. This article firstly explores the limited previous research on boarding patients. Secondly, this paper quantitatively demonstrates that needy boarding patients can significantly impact system performance and hence should be accounted for in the analysis and planning of an emergency department through discrete-event simulation. Next, three static priority policies (first-come, first-served and always prioritizing either boarding patients or the other patients) and one dynamic priority policy (using accumulating priorities) are evaluated on various performance measures. We find, for our case study, that system performance is optimized by applying a priority policy that favours patients currently under treatment in the emergency department.

Key words: Needy boarding patients, emergency department, discrete-event simulation, prioritization, accumulating priority queue

¹ Corresponding author

1. Introduction

A serious issue often encountered in hospital emergency departments (EDs) is crowding. Crowding occurs when the demand for emergency resources exceeds the resources available in the ED (Moskop, Sklar, Geiderman, Schears & Bookman, 2009). This can cause several negative effects including longer waiting time and length of stay (LOS), lost hospital revenue, increased ambulance diversion, more patients that leave without being seen (LWBS), and even higher mortality rates (Hoot & Aronsky, 2008).

One of the main contributors to crowding is inpatient boarding which results from the inability to transfer patients from the ED to the inpatient ward (IW) (Moskop et al., 2009). As a consequence, these boarding patients need to wait in the ED. While waiting, they require treatment from the ED staff and occupy beds which, as a consequence, cannot be used by newly arriving patients. In this way, boarding patients congest the ED.

Although the importance of inpatient boarding is stated multiple times in the operations research (OR) and operations management (OM) literature, very little research focuses on this topic (Saghafian, Austin & Traub, 2015). Furthermore, when boarding is included in the analysis, it is often assumed that boarding patients occupy beds in the ED, but do not require any other resources. However, in reality, boarding patients may still need additional check-ups by ED physicians, require time from ED nurses or occupy other ED equipment than beds (Armony et al., 2015).

This paper intends to investigate the effect of this simplification by incorporating the additional check-ups required by boarding patients (needed boarding patients) into the analysed system. Adding these required check-ups increases the workload of the physicians, but also raises an additional question. That is, after finishing a treatment, physicians need to decide whether they will see a boarding patient or one of the other patients in the system. Since the treatment characteristics of these two kinds of patients differ, this impacts ED system performance.

The remainder of this paper is organized as follows. Section 2 provides an overview of the currently existing literature on this topic and Section 3 presents the methodology and case study. Section 4 describes, analyses and compares the numerical results of the scenarios investigated. Finally, Section 5 presents the main conclusions and opportunities for future research.

2. Literature review

Inpatient boarding is also referred to in the literature as “access block” (Au et al., 2008; Khanna, Boyle, Good & Lind, 2012; Luo, Cao, Gallagher & Wiles, 2013) or “bed block” (Rashwan, Abo-Hamad, & Arisha, 2015; Saghafian et al., 2015). It is one of the key factors contributing to overcrowding and has a significant influence on patients’ LOS in the ED (Carmen, Defraeye & Van Nieuwenhuyse, 2015; Shi, Chou, Dai, Ding & Sim, 2016). Furthermore, these longer ED stays increase hospital LOS, risk of hospital-acquired infections and mortality rates (Armony et al., 2015). Additionally, inpatient boarding is defined as the main cause of ambulance bypass which occurs when ambulances are diverted to other

hospitals because the ED is too crowded; e.g. Au et al. (2008) observe that access block is responsible for 76% of all ambulance bypass. Also non-boarding patients may suffer from delays in the transfer process; since boarding patients still require treatment while waiting to be admitted to the IW, the additional workload they put on ED staff may be significant as boarding patients take up to 11% of physician time in the ED (Armony et al., 2015).

Surprisingly, however, when analysing the ED, inpatient boarding is often neglected or not the main focus of the study. Indeed, articles in the OR/OM literature concentrating on this specific topic are notably scarce (Saghafian et al., 2015). Relevant inpatient boarding articles are categorized according to two distinguishing features: the articles' focus and methodology (Table 1). Firstly, although ED boarding occurs at the intersection of the ED and the IW, articles that take into account ED boarding usually focus either on the ED or on the IW. Secondly, given the wide range of OR/OM tools currently available, articles are categorized according to the methodology used in the analysis.

Table 1: Overview of relevant inpatient boarding articles.

Article	Focus		Methodology		
	ED	IW	Simulation	Queueing theory	Statistical methods
Bair et al. (2010)	X		X		
Carmen et al. (2015)	X				
Crawford et al. (2014)	X		X		
Khare et al. (2009)	X		X		
Saghafian et al. (2012)	X		X	X	
Au et al. (2008)	X			X	
Carmen et al. (2018)	X			X	
Khanna et al. (2012)	X		X		X
Bagust et al. (1999)		X	X		
Mustafee et al. (2012)		X	X		
Rashwan et al. (2015)		X	X		
Shi et al. (2016)		X	X	X	
Luo et al. (2013)		X			X
Armony et al. (2015)		X			X

Root causes of inpatient boarding are usually sought in the IW. Shortage in the number of IW beds is often identified by hospital staff to be the number one cause of boarding (Mustafee et al., 2012). Other reasons for inpatient boarding are the inability to discharge IW patients in a timely manner (Crawford, Parikh, Kong & Thakkar, 2014; Luo et al., 2013; Rashwan et al., 2015) and ED "batch admitting", which occurs when several boarding patients are queued and admitted to the IW altogether at a time convenient to the staff (Luo et al., 2013). On the one hand, Armony et al. (2015) group plausible causes for ED-to-IW delays into four main categories: inadequate synchronization between the ED and the IW, bad work methods, shortage in staff availability and lack of equipment availability. Shi et al. (2016), on the other hand, attribute boarding times to a mismatch between the daily number of arrivals and discharges, and

a mismatch between the discharge timing and hourly arrival pattern. Depending on which process or resource that is defined to be the main cause of inpatient boarding, the proposed improvement strategies differ.

Articles that study the consequences of inpatient boarding tend to model the system from an ED perspective. Performance measures considered in these studies are the rate of LWB (Bair, Song, Chen & Morris, 2010), ambulance bypass (Au et al., 2008), LOS in the ED (Khare, Powell, Reinhardt & Lucenti, 2009), waiting times, service levels, and maximum throughput (Carmen, Van Nieuwenhuysse & Van Houdt, 2018). Even if boarding is not the main focus, ED studies should take this phenomenon into account to obtain realistic results (Carmen et al., 2015; Saghaian, Hopp, Van Oyen, Desmond & Kronick, 2012).

Depending on the focus of the article, studies model either the ED or the IW in detail, but to the best of our knowledge, none model both the ED and the IW. With boarding being located at the intersection of the ED and IW, modelling it is complicated and it is often neglected. Furthermore, even when it is included in the study, it is usually modelled in a very basic way. On the one hand, when focusing on the IW, boarding patients are typically modelled as IW admissions which arrive according to a certain arrival process (Bagust, Place & Posnett, 1999; Luo et al., 2013; Rashwan et al., 2015). On the other hand, when using an ED focus, it is often assumed that boarding patients only occupy ED beds and do not need any further treatment in the ED (Carmen et al., 2015; Carmen et al., 2018; Crawford et al., 2014). However, data analysis (e.g. Armony et al., 2015) shows that boarding patients take up to 11% of physician time in the ED. Neglecting this can considerably underestimate the staff's workload. This paper aims to contribute to the existing literature by adding these additional treatments required by boarding patients to an ED-focused analysis.

Another way to categorize the literature that concentrates on inpatient boarding is to classify articles based on the OR/OM methodologies used for investigating this topic (Table 1). Simulation can be used for the main analysis (Saghaian et al., 2015; Zeltyn et al., 2011), or as an additional methodology to verify an analytical model by comparing the simulation output with empirical estimates or to conduct a sensitivity analysis on the outcome of the study (e.g. Shi et al., 2016). Simulation has been the preferred OR/OM tool to study ED operations for years (e.g. Bagust et al., 1999) since it is able to simulate a complex and stochastic environment and therefore can capture the real system's behaviour (Rashwan et al., 2015; Saghaian et al., 2015). Also, it is cheaper, safer, and faster than performing real-time experimentation (Mutafee et al., 2012). However, the main drawback is the lack of generality over the wide range of different simulation studies (Günal & Pidd, 2010). Most of the time, simulation is used for modelling specific units and facilities, and these models are rarely reused in other similar studies, which may make the generalization of the results questionable.

Alternative OR/OM methodologies used in the reviewed articles are queueing theory (Au et al., 2008; Carmen et al., 2018; Saghaian et al., 2012; Shi et al., 2016), statistical methods (Khanna et al., 2012; Luo et al., 2013) and data analysis (Armony et al., 2015). In contrast to the numerous articles that

apply simulation, studies that use other OR/OM methodologies than simulation are remarkably rare. This is not surprising since adding inpatient boarding to the analysis makes it more realistic and accurate, but also more complex. With this rise in complexity, the need for an OR/OM methodology that can capture all these complex aspects of the ED also grows. As discussed before, simulation is particularly suitable for modelling such environments. Moreover, simulation models are often easy to understand and use for hospital staff. Consequently, we will also rely on simulation in this paper when investigating the impact of needy boarding patients on ED performance and comparing different priority policies for ED physicians.

3. Methodology and case study

Given that boarding patients and the pressure they put on not only ED beds but also ED personnel is often neglected, we explore the severity of this caveat using a stylised discrete-event simulation. As data on the time a boarding patient spends in the ED is scarce and data on the time they require check-ups by a physician is almost non-existent, we rely on the little data we were able to find and keep the simulation model simple so as not to distract the focus from the main issue.

3.1. Case study: ED simulation model

Patients arriving at EDs differ with respect to their complaints and the severity of their condition. One approach often applied to deal with the heterogeneity in ED patients is the use of triage (e.g. by using an emergency severity index (ESI)) (Ferrand, Magazine, Rao & Glass, 2018). Another solution consists of dividing the ED in separated zones with dedicated beds and physicians, also referred to as patient streaming (Saghafian et al., 2012; Saghafian, Hopp, Van Oyen, Desmond & Kronick, 2014). Moreover, several combinations of these two approaches exist (e.g. using an ESI prioritization in conjunction with a dedicated fast-track zone (Ferrand et al., 2018)). In this paper, we focus on EDs where patients are treated in separated zones, implying that within one zone, patients have similar treatment characteristics. Especially in larger hospitals, this is a realistic setting and it enables us to use real-life data of an Israeli hospital (see Section 3.3).

Based on Carmeli et al. (2018), the ED is analysed from a patient flow perspective, where each zone in the ED is treated as a separate flow network. In each flow network (zone), patients can return to the physician for additional treatment several times (re-entrant patients) and boarding patients generate additional check-ups and thus extra workload (needy boarding patients). A schematic overview of such a zone is provided in Figure 1.

Figure 1: Overview of *one* ED zone.

In each zone, a homogenous flow of patients arrives according to a nonstationary Poisson arrival process with time-varying arrival rate λ_t (where t represents time). Before starting the actual treatment process, they wait in the external queue until they obtain a bed, of which there are N . After obtaining a bed, patients enter the physician queue in order to see one of the physicians. This initial treatment process is exponentially distributed with mean $\frac{1}{\mu_1}$ and it is assumed that any of the physicians can treat any patient. The assumption of exponential service times is based on Zeltyn et al. (2011) who validated this fit through a statistical analysis on the same data that we use in our paper (see Section 3.3). The number of physicians present in a zone s_t depends on the time of the day and varies hourly. In order to determine s_t , we apply the staffing algorithm proposed in Yom-Tov and Mandelbaum (2014).

After the initial treatment step, there are three ways in which patients can continue their treatment process. First, with probability p_1 , patients need to return to the physician for additional treatment after a delay which is exponentially distributed with mean $\frac{1}{\delta}$. This delay represents any treatment step that does not require the physician's presence like nurse treatments, lab tests, or scans. We assume that this treatment step has infinite capacity. Second, with probability $(1 - p_1)p_b$, patients need admission into the hospital, but there is no IW bed available for them. Therefore, they have to wait in the ED until they can be hospitalized in the IW. This boarding time is exponentially distributed with mean $\frac{1}{\beta}$. While waiting, boarding patients generate additional check-ups according to a Poisson distribution with rate μ_b . The durations of these check-ups by the physician are exponentially distributed with mean $\frac{1}{\mu_2}$. When a patient's boarding time is over, he can go to the IW if, at that time, he is not seeing the physician or waiting for a check-up. Otherwise, the patient will be transferred to the IW immediately after returning from the physician. Last, with probability $(1 - p_1)(1 - p_b)$, patients leave the ED after treatment.

With needing boarding patients, there are three kinds of patients waiting in the physician queue: patients that have just obtained a bed and need to start their treatment process, patients that have returned from tests, and boarding patients that require a check-up. In this paper, the former two will be referred to as test patients. Consequently, two types of patients are analysed in each ED zone: test and boarding

patients. This distinction is made clear in the flow model displayed in Figure 1 by using different colours: red for test patients and blue for boarding patients. Remark that each of the parameters in the flow model can vary over the different ED zones, reflecting the heterogeneity in patients' characteristics between the zones.

It should be noted that several modelling assumptions (i.e., exponentially distributed service times, memoryless routing probabilities, test patient homogeneity within each zone) deviate from observations as generally reported in practice. However, these assumptions are based on a specific case study and on literature (see also Section 3.3). So, given the available information, we conduct a case study which, although presenting a simplified version of reality in terms of the limitations mentioned above, provides a good starting point for a first exploration on the impact of needy boarding patients on ED performance.

3.2. Analysed scenarios

In order to analyse the effect of adding needy boarding patients to the modelled system and to determine a priority policy for the physicians, we compare several scenarios by varying four parameters: the type of boarding patients, the priority policy, the staffing levels, and the bed utilization (Table 2).

Table 2: Overview of analysed parameters in the scenarios.

Parameter	Considered values
Boarding patients	BP_{b+p} ; BP_b ; BP_{none}
Priority policy	U; FFS; B; T; $AP(\frac{z_b}{z_t})$
Staffing levels	Staffing_LQoS; Staffing_HQoS
Bed utilization	Bed_HU; Bed_LU

Type of boarding patients. To investigate whether it is necessary to take into account needy boarding patients, scenarios where boarding patients occupy both beds and physicians (BP_{b+p}), occupy only beds (BP_b), and do neither occupy beds, nor physicians (BP_{none}) are modelled.

Priority policy. When both test and boarding patients need physician treatment (BP_{b+p}), physicians are faced with a prioritization decision and hence the need for a priority policy arises. Given the different characteristics of test and boarding patients, the choice whether to give priority to one or another can have an impact on system performance. On the one hand, prioritizing test patients increases the probability that boarding patients are blocked in the physician queue (i.e. a patient's boarding time is over and he is allowed to enter the IW, but is still waiting for a physician check-up). Consequently, boarding patients occupy beds in the ED longer than needed and thus prevent newly arriving patients waiting in the external queue from entering the treatment process (bed blocking effect of boarding patients). On the other hand, giving priority to boarding patients can also prevent newly arriving patients from entering the process. That is, every test patient has a probability $(1 - p_1)(1 - p_b)$ of leaving the system, and thus releasing a bed that can be used for patients waiting in the external queue, after seeing

the physician. But due to the prioritization of boarding patients, they have to wait longer in the physician queue and therefore may occupy beds longer than needed (bed blocking effect of test patients).

To investigate the effect of the physicians' priority policy on system performance in our case study, three static priority policies and one dynamic priority policy are analysed. In static priority policies, physicians apply a prioritization rule independent from the system state. In a dynamic priority policy however, the system state dictates which type of patients gets priority in the physician queue. The following priority policies are analysed and compared:

1. The first static priority policy applies a first-come, first-served priority rule (FCFS).
2. In the second static priority policy, physicians always give priority to boarding patients (B). This policy is equivalent to a priority policy that prioritizes patients with the shortest average processing time. Indeed, the average duration of the check-ups required by boarding patients is much shorter than the average duration of the treatment times of test patients (see also Section 3.3)
3. Alternatively, in the third static priority policy, physicians always give priority to test patients (T). Likewise, this policy is equivalent to a priority policy that prioritizes patients with the longest average processing time.
4. The fourth priority policy is a dynamic priority policy that implements an accumulating priority queue where a patient's priority is a linear function of his waiting time in the physician queue (Fajardo & Dreier, 2017; Li & Stanford, 2016; Sharif, Stanford, Taylor & Ziedins, 2014; Stanford, Taylor & Ziedins, 2014). As indicated by Ferrand et al. (2018), such a policy is suited for ED applications. The rate z_i at which the priority increases depends on the patient type i ($i \in \{t, b\}$ where t stands for test patients and b for boarding patients). The physician prioritizes the patient with the highest accumulated priority in the physician queue.

The effect of the accumulating priority policy ($AP(\frac{z_b}{z_t})$) is analysed by varying $\frac{z_b}{z_t}$. This is done by altering z_b , while keeping z_t constant and equal to 1. An overview and explanation of the analysed values of $\frac{z_b}{z_t}$ are given in Table 3. As can be seen from Table 3, the static priority policies are included in the analysis of the dynamic $AP(\frac{z_b}{z_t})$ priority policy; $AP(1)$, $AP(9999)$, and $AP(0)$ equal the FCFS, B, and T priority policies respectively.

Table 3: Overview and explanation of analysed values of $\frac{z_b}{z_t}$.

$\frac{z_b}{z_t}$	Explanation ²
0	T priority policy
0.2	$z_b = 0.2 * z_t$
0.4	$z_b = 0.4 * z_t$
0.6	$z_b = 0.6 * z_t$
0.8	$z_b = 0.8 * z_t$
1	FCFS priority policy
1.25	$z_t = 0.8 * z_b$
1.67	$z_t = 0.6 * z_b$
2.5	$z_t = 0.4 * z_b$
5	$z_t = 0.2 * z_b$
9999	B priority policy

Staffing levels. The effect of needy boarding patients on system performance and the impact of the applied priority policy are analysed in scenarios with low and high staffing levels. As mentioned in Section 3.1, the number of physicians present in a zone varies over time and is determined by applying the staffing algorithm proposed in Yom-Tov and Mandelbaum (2014). This algorithm relies on a parameter β_{QoS} that reflects the quality of service (the higher β_{QoS} , the higher the quality of service and hence the staffing levels). In this paper, we consider two scenarios with lower quality of service staffing (Staffing_LQoS) and scenarios with high quality of service staffing (Staffing_HQoS) by setting β_{QoS} equal to 0.1 and 0.5 respectively. Hence, Staffing_LQoS and Staffing_HQoS equal scenarios with low and high staffing levels respectively.

Bed utilization. Likewise, scenarios with high bed utilization (Bed_HU) and low bed utilization (Bed_LU) are modelled. This is done by targeting an average bed utilization rate of 85% and 70% respectively, in a scenario with BP_{b+p} , FCFS, and Staffing_LQoS, by varying the number of beds.

Henceforth, scenarios are referred to as {Type of boarding patients, Priority policy, Staffing levels, Bed utilization} by using the abbreviations in Table 2. Obviously, the choice of priority policy is only relevant in a scenario with BP_{b+p} . Moreover, when all combinations of a certain parameter are analysed, this is indicated by “(·)”. Given a full factorial design of the parameters introduced in Table 2, 52 scenarios are modelled and compared.

² The values of $\frac{z_b}{z_t}$ displayed on the x-axis in the graphs in this paper are chosen such that on the left of $\frac{z_b}{z_t} = 1$ the formula $z_b = \alpha * z_t$ applies and on the right of $\frac{z_b}{z_t} = 1$ the formula $z_t = \alpha * z_b$. Furthermore, α goes from 0 to 1 in steps of 0.2. Consequently, values of $\frac{z_b}{z_t}$ on the left and right of $\frac{z_b}{z_t} = 1$ which are based on the same α (for example $\frac{z_b}{z_t} = 0.2$ and $\frac{z_b}{z_t} = 5$) are located equally far from $\frac{z_b}{z_t} = 1$ on the x-axis in the graphs.

3.3. Case study: input data

The input data used in this paper is based on Yom-Tov and Mandelbaum (2014) and Armony et al. (2015), who both use real-life data of an Israeli hospital. Data on this hospital's ED is available via the Technion SEE center. ED patients are divided in four separate patient types: Internal (patient type 1), Orthopaedic (patient type 2), Surgical (patient type 3), and Trauma (patient type 4). The hourly arrival rates of these patient types are shown in Figure 2, and the corresponding hourly staffing levels are depicted in Figure 3 and Figure 4. An overview of the remaining input parameters according to patient type is given in Table 4. Since very little empirical data exist that describe the parameters related to boarding patients (i.e. p_b , β , μ_b and μ_2), we were not able to collect these data for each patient type separately. Consequently, we use the same average values for all patient types as reported by Armony et al. (2015).

Figure 2: Hourly arrival rates.

Figure 3: Hourly staffing levels (Staffing_LQoS).

Figure 4: Hourly staffing levels (Staffing_HQoS).

Table 4: Input data used in the simulation model.

Patient type	μ_1	δ	p_1	p_b	β	μ_2	μ_3	$N(Bed_HU)$	$N(Bed_LU)$
1	8.91	0.953	0.7743	0.35	$\frac{1}{3}$	4	40	30	36
2	8.86	0.969	0.6094	0.35	$\frac{1}{3}$	4	40	7	9
3	10.33	0.572	0.6441	0.35	$\frac{1}{2}$	4	40	18	22
4	12.37	1.31	0.7268	0.35	$\frac{1}{3}$	4	40	18	22

3.4. Performance measures

To evaluate and compare scenarios in our case study, we mainly rely on the performance measures given in Table 5. However, we also track numerous other performance measures in our simulation study (e.g. total physician queue waiting time, number of patients in the physician and external queue, $P(W_{phys} \leq 15 \text{ min})$, $P(W_e \leq 15 \text{ min})$, overtime boarding) and report them as an average over all patients as well as for test and boarding patients separately when relevant. These data can be found in the online Appendix.

Table 5: Analysed performance measures.

Notation	Explanation
$E[LOS_{exl.b}]$	Average length of stay excluding boarding time
$E[LOS_{incl.b}]$	Average length of stay including boarding time
$E[W_{phys}]$	Average waiting time in the physician queue
$E[W_e]$	Average waiting time in the external queue
$P(W_{phys} = 0)$	Probability that a patient does not have to wait in the physician queue
$P(W_{phys} \leq 30 \text{ min})$	Probability that a patient does not have to wait longer than 30 minutes in the physician queue
$P(W_e = 0)$	Probability that a patient does not have to wait in the external queue
$P(W_e \leq 30 \text{ min})$	Probability that a patient does not have to wait longer than 30 minutes in the external queue

3.5. Modelling approach

Arena Simulation Software® is used to simulate the ED. Since our aim is to study the ED in a steady-state, the graphical procedure of Welch (Mahajan & Ingalls, 2004) is used to determine the length of the warm-up period. After making 100 replications for $\{BP_{b+p}, FCFS, Staffing_LQS, Bed_HU\}$, each with a length of 150 hours, we calculate the cross-replication averages of the utilization rates of physicians ρ_{phys} and beds ρ_{bed} for every 10 hours. Next, we calculate the moving averages over these periods using a window of 3 periods. This means that we calculate the moving average $\overline{M}_t$ at period t as follows:

$$\overline{M}_t = \begin{cases} \frac{\sum_{i=1}^{2t-1} \overline{Y}_i}{2t-1} & \text{for } t \leq 3 \\ \frac{\sum_{i=t-3}^{t+3} \overline{Y}_i}{7} & \text{for } 3 < t \leq 12 \end{cases} \quad \text{with } \overline{Y}_i \text{ the cross-replication average of period } i$$

As apparent from Figure 5, the moving averages of both utilization rates become relatively stable after 5 periods or 50 hours. Following the graphical procedure of Welch, the simulation model enters the steady state at that point. To build in extra safety, a warm-up period of 100 hours is used.

Figure 5: Moving averages of ρ_{phys} and ρ_{bed} .

We use a replication length of 8860 hours of which the first 100 hours count as a warm-up period. This means that we collect statistics over the period of one year (8760 hours). In each scenario, the number of replications equals 100. Using these settings, the half-widths of ρ_{phys} and ρ_{bed} never exceed 0.28% in the analysed scenarios.

In order to compare the different scenarios, we rely on the 95% confidence intervals; scenarios significantly differ from each other when their 95% confidence intervals do not overlap.

4. Numerical results of case study

This section presents the obtained numerical results of our case study. The effect of needy boarding patients on ED system performance is analysed in Section 4.1. Next, Section 4.2 investigates whether the choice of priority policy impacts ED system performance. Based on these two sections, we analyse

whether an optimal priority policy exists in Section 4.3. Finally, Section 4.4 presents a short discussion of the numerical results with respect to the generalizability of our conclusions beyond the case study. Since it is not possible to present all combinations of performance measures, scenarios and patient types due to space limitations, we focus on those who are most outspoken or illustrate the main findings and conclusions of our case study. However, a full overview of the performance measures for all scenarios and patient types can be found in the online appendix.

4.1. Effect of needy boarding patients on ED system performance

In order to investigate the effect of needy boarding patients on ED performance, we compare scenarios with BP_{b+p} , BP_b and BP_{none} while fixing the other parameters (priority policy, staffing levels and bed utilization). We find that the differences in ED system performance between BP_{b+p} , BP_b and BP_{none} depend on the considered performance measure and patient type.

To illustrate and explain the differing effect of needy boarding patients between patient types, we look at $E[LOS_{incl_b}]$ in $\{(\cdot), AP(0), \text{Staffing_LQoS}, \text{Bed_LU}\}$ for patient type 1 and patient type 2. As is clear from Figure 6, the impact of boarding patients on $E[LOS_{incl_b}]$ differs substantially between the two patient types when going from BP_b to BP_{b+p} . This observed divergence in impact of needy boarding patients is due to a difference in average physician utilization $E[\rho_{phys}]$. Whereas $E[\rho_{phys}]$ is only 61.37% for patient type 2 in $\{BP_{b+p}, AP(0), \text{Staffing_LQoS}, \text{Bed_LU}\}$, it reaches 86.83% for patient type 1. The low value of $E[\rho_{phys}]$ for patient type 2 is explained by the fact that the ED zone in which patient type 2 is treated is a small system. Since the staffing rule applied in this paper rounds up the calculated values when determining the number of staffed physicians, it tends to overstaff small systems, resulting in low physician utilizations. Defraeye and Van Nieuwenhuyse (2013) argue that this sort of staffing rules is not appropriate for small systems and recommend using an extended version of the iterative staffing algorithm (ISA(1)). As overstaffing in small systems is indicated as a disadvantage of the staffing rule applied in this paper, other staffing rules might result in higher $E[\rho_{phys}]$. This implies that we cannot conclude that the effect of needy boarding patients can be neglected in small systems, as other staffing rules might result in higher $E[\rho_{phys}]$ and therefore introduce an important impact of needy boarding patients.

Figure 6: $E[LOS_{incl_b}]$ and $P(W_{phys} = 0)$ for patient type 1 and patient type 2 in $\{BP_b, AP(0), Staffing_LQoS, Bed_LU\}$.

Moreover, the effect of needy boarding patients also depends to a great extent on the performance measure we consider. Although there is no important difference in $E[LOS_{incl_b}]$ for patient type 2 between $\{BP_{b+p}, AP(0), Staffing_LQoS, Bed_LU\}$ and $\{BP_b, AP(0), Staffing_LQoS, Bed_LU\}$, the impact on $P(W_{phys} = 0)$ is ample (Figure 6). This might greatly impact staffing rules as these are often based on a target delay probability (e.g. Feldman, Mandelbaum, Massey and Whitt (2008), Green, Kolesar and Whitt (2007), Yom-Tov and Mandelbaum (2014)). Indeed, when staffing is set based on a target delay probability in $\{BP_b, AP(0), Staffing_LQoS, Bed_LU\}$, the realised delay probability (i.e. the delay probability in $\{BP_{b+p}, AP(0), Staffing_LQoS, Bed_LU\}$) will be higher than the targeted one. Or, in other words, when taking into account needy boarding patients (i.e. $\{BP_{b+p}, AP(0), Staffing_LQoS, Bed_LU\}$), staffing levels need to be higher in order to reach the same delay probability as in $\{BP_b, AP(0), Staffing_LQoS, Bed_LU\}$. However, currently, no research exists that accounts for needy boarding patients when setting staffing levels and hence, this provides an opportunity for future research.

It is important to note that the differing impact of needy boarding patients (high impact versus low impact) on ED system performance only holds when comparing different patient types or performance measures. Indeed, if we consider the impact of needy boarding patients to be substantial when varying scenarios between BP_{b+p} , BP_b and BP_{none} results in a substantial impact for at least one patient type and considered performance measure, we find that needy boarding patients impact ED system performance for all investigated scenarios, regardless of the priority policy, staffing levels or bed utilization.

To conclude, we observe that ED system performance is significantly impacted by needy boarding patients in our case study. Even though exceptionally, isolated cases are present where it can be justified to approximate BP_{b+p} by BP_b or even BP_{none} , no single scenario exists in which this holds for all patient types and all performance measures considered. This quantitatively demonstrates the need to explicitly incorporate needy boarding patients into the analysed ED system.

4.2. Effect of priority policy on ED system performance

The effect of the priority policy on ED system performance is analysed by varying $\frac{z_b}{z_t}$ for the dynamic $AP(\frac{z_b}{z_t})$ priority policy. As explained in Section 3.2, the static priority policies are included in the analysis of the $AP(\frac{z_b}{z_t})$ priority policy; $AP(1)$, $AP(9999)$, and $AP(0)$ equal the FCFS, B, and T priority policies respectively.

Similar to what we observed in Section 4.1, the effect of the priority policy on ED system performance depends on the considered performance measure and patient type. As can be seen in Figure 7(a), although the priority policy has an enormous impact on $E[LOS_{incl_b}]$ in $\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_LU}\}$ for patient type 1 and patient type 4, the effect on $E[LOS_{incl_b}]$ for patient type 2 and patient type 3 is much smaller. This is again explained by the observed $E[\rho_{phys}]$. Also, the effect of the priority policy depends on the performance measure that is taken into account. Even though the effect on $E[LOS_{incl_b}]$ in $\{BP_{b+p}, (\cdot), \text{Staffing_HQoS}, \text{Bed_LU}\}$ is limited (Figure 7(b)), the effect on $P(W_{phys} = 0)$ is substantial for patient types 1, 3 and 4 (Figure 8(b)).

Figure 7: $E[LOS_{incl_b}]$ in (a) $\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_LU}\}$ and (b) $\{BP_{b+p}, (\cdot), \text{Staffing_HQoS}, \text{Bed_LU}\}$.

Figure 8: $P(W_{phys} = 0)$ in (a) $\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_LU}\}$ and (b) $\{BP_{b+p}, (\cdot), \text{Staffing_HQoS}, \text{Bed_LU}\}$.

Similar to Section 4.1, we observe that although isolated instances exist where the priority policy has only a limited impact, no single scenario exists in which this holds for all patient types and all performance measures considered. We therefore conclude that the choice of priority policy has an effect on ED system performance in our case study and it is thus worthwhile to investigate the existence of an optimal priority policy.

4.3. Optimal priority policy

The optimal priority policy in our case study is determined by comparing scenarios with varying $AP(\frac{z_b}{z_t})$ priority policies, while fixing staffing levels and bed utilization. Table 6 and Table 7 give an overview of the (range of) optimal priority policies for each patient type and each of the considered performance measures in our case study. When multiple priority policies are indicated, this means that these priority policies are not significantly different from each other with respect to the performance measure under study.

Table 6: Optimal priority policy for different scenarios, patient types and performance measures $E[LOS_{exl,b}]$, $E[LOS_{incl,b}]$, $E[W_{phys}]$, and $E[W_e]$.

Scenario	Patient type	Considered performance measure			
		$E[LOS_{exl,b}]$	$E[LOS_{incl,b}]$	$E[W_{phys}]$	$E[W_e]$
$\{BP_{b+pp}, (-), \text{Staffing_LQoS, Bed_HU}\}$	Patient type 1	AP(0)	AP(0), AP(0.2)	AP(0.6) - AP(5)	AP(0), AP(0.2)
	Patient type 2	AP(0)	AP(0)	AP(9999)	AP(0), AP(0.2)
	Patient type 3	AP(0)	AP(0), AP(0.2)	AP(1) - AP(9999)	AP(0) - AP(0.6)
	Patient type 4	AP(0)	AP(0)	AP(0.2) - AP(1.67)	AP(0)
$\{BP_{b+pp}, (-), \text{Staffing_LQoS, Bed_LU}\}$	Patient type 1	AP(0)	AP(0.2) - AP(0.6)	AP(0.6) - AP(2.5)	AP(0.2) - AP(0.6)
	Patient type 2	AP(0)	AP(0), AP(0.2)	AP(9999)	AP(0) - AP(0.4)
	Patient type 3	AP(0)	AP(0) - AP(0.4)	AP(0.8) - AP(5)	AP(0.2) - AP(0.8)
	Patient type 4	AP(0)	AP(0), AP(0.2)	AP(0.4) - AP(1.25)	AP(0.2), AP(0.4)
$\{BP_{b+pp}, (-), \text{Staffing_HQoS, Bed_HU}\}$	Patient type 1	AP(0)	AP(0.2), AP(0.4)	AP(1.67) - AP(9999)	AP(0) - AP(0.6)
	Patient type 2	AP(0), AP(0.2), AP(0.6), AP(0.8), AP(1.25)	AP(0), AP(0.2), AP(0.6), AP(0.8), AP(1.25)	AP(9999)	AP(0), AP(0.2), AP(0.6), AP(0.8), AP(1.25)
	Patient type 3	AP(0), AP(0.2)	AP(0) - AP(0.4)	AP(1.67) - AP(9999)	AP(0) - AP(0.4)
	Patient type 4	AP(0)	AP(0), AP(0.2)	AP(1) - AP(9999)	AP(0), AP(0.2)
$\{BP_{b+pp}, (-), \text{Staffing_HQoS, Bed_LU}\}$	Patient type 1	AP(0)	AP(0.2) - AP(0.6)	AP(1.67) - AP(9999)	AP(0.2) - AP(1.25)
	Patient type 2	AP(0) - AP(1)	AP(0) - AP(1), AP(1.67)	AP(9999)	AP(0) - AP(1), AP(1.67), AP(2.5)
	Patient type 3	AP(0)	AP(0) - AP(0.4)	AP(2.5) - AP(9999)	AP(0) - AP(0.8)
	Patient type 4	AP(0)	AP(0), AP(0.2)	AP(0.8) - AP(9999)	AP(0) - AP(0.6)

Table 7: Optimal priority policy for different scenarios, patient type and performance measures $P(W_{phys} = 0)$, $P(W_{phys} \leq 30 \text{ min})$, $P(W_e = 0)$, and $P(W_e \leq 30 \text{ min})$.

Scenario	Patient type	Considered performance measure			
		$P(W_{phys} = 0)$	$P(W_{phys} \leq 30 \text{ min})$	$P(W_e = 0)$	$P(W_e \leq 30 \text{ min})$
$\{BP_{b+p}, (\cdot), \text{Staffing_LQoS, Bed_HU}\}$	Patient type 1	AP(0)	AP(0)	AP(0), AP(0.2)	AP(0), AP(0.2)
	Patient type 2	AP(0)	AP(2.5), AP(5)	AP(0)	AP(0), AP(0.2)
	Patient type 3	AP(0)	AP(0)	AP(0), AP(0.2)	AP(0), AP(0.2)
	Patient type 4	AP(0)	AP(0)	AP(0)	AP(0)
$\{BP_{b+p}, (\cdot), \text{Staffing_LQoS, Bed_LU}\}$	Patient type 1	AP(0)	AP(0)	AP(0.2) - AP(0.6)	AP(0.2) - AP(0.6)
	Patient type 2	AP(0)	AP(2.5), AP(5)	AP(0) - AP(0.8)	AP(0) - AP(0.8)
	Patient type 3	AP(0)	AP(0)	AP(0) - AP(0.8)	AP(0) - AP(0.8)
	Patient type 4	AP(0)	AP(0)	AP(0), AP(0.2)	AP(0), AP(0.2)
$\{BP_{b+p}, (\cdot), \text{Staffing_HQoS, Bed_HU}\}$	Patient type 1	AP(0)	AP(0)	AP(0) - AP(0.4)	AP(0) - AP(0.4)
	Patient type 2	AP(0)	AP(2.5), AP(5)	AP(0), AP(0.2), AP(0.6), AP(0.8), AP(1.25), AP(2.5)	AP(0), AP(0.2), AP(0.6), AP(0.8), AP(1.25), AP(2.5)
	Patient type 3	AP(0)	AP(2.5)	AP(0) - AP(0.4)	AP(0) - AP(0.4)
	Patient type 4	AP(0)	AP(1.67), AP(2.5)	AP(0)	AP(0)
$\{BP_{b+p}, (\cdot), \text{Staffing_HQoS, Bed_LU}\}$	Patient type 1	AP(0)	AP(2.5), AP(5)	AP(0.2) - AP(0.8)	AP(0.2) - AP(1)
	Patient type 2	AP(0)	AP(2.5)	AP(0) - AP(2.5)	AP(0) - AP(2.5)
	Patient type 3	AP(0)	AP(2.5)	AP(0) - AP(0.8)	AP(0) - AP(0.8)
	Patient type 4	AP(0)	AP(1) - AP(2.5)	AP(0) - AP(0.4)	AP(0) - AP(0.4)

A first observation is that for almost all performance measures, the optimal priority policies in our case study tend to favour the same kind of patients (either boarding patients by applying high values of $\frac{z_b}{z_t}$ or test patients by setting low values of $\frac{z_b}{z_t}$). Moreover, they do so regardless the scenario or patient type. There is only one exception; when maximizing $P(W_{phys} \leq 30 \text{ min})$, there is a big difference in optimal priority policies between scenarios and patient types. This difference is due to a difference in $E[\rho_{phys}]$; when $E[\rho_{phys}]$ is high, giving more priority to test patients results in higher values of $P(W_{phys} \leq 30 \text{ min})$, whereas for lower values of $E[\rho_{phys}]$, an opposite effect is observed. However, as Figure 9 indicates, the instances where it is optimal to favour boarding patients already reach $P(W_{phys} \leq 30 \text{ min})$ of more than 90% for any priority policy. One might therefore argue that focusing on $P(W_{phys} \leq 30 \text{ min})$ to determine the priority policy is not appropriate in these instances, as the effect of the prioritization policy is not relevant, even though it is significant (according to the 95% confidence intervals).

Figure 9: $P(W_{phys} \leq 30 \text{ minutes})$ in (a) $\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_HU}\}$ and (b) $\{BP_{b+p}, (\cdot), \text{Staffing_HQoS}, \text{Bed_LU}\}$.

A second major finding is that the optimal priority policy in our case study strongly depends on the performance measure under study. Whereas minimizing $E[W_e]$ results in giving more priority to test patients (low values for $\frac{z_b}{z_t}$), minimizing $E[W_{phys}]$ favours prioritizing more boarding patients (high values for $\frac{z_b}{z_t}$). There is also a slight difference between $E[LOS_{exl_b}]$ and $E[LOS_{incl_b}]$. Although both performance measures agree to give more priority to test patients, they disagree on the value of $\frac{z_b}{z_t}$ for patient type 1 in $\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_LU}\}$, $\{BP_{b+p}, (\cdot), \text{Staffing_HQoS}, \text{Bed_HU}\}$, and $\{BP_{b+p}, (\cdot), \text{Staffing_HQoS}, \text{Bed_LU}\}$.

The essential difference between $E[LOS_{exl_b}]$ and $E[LOS_{incl_b}]$ lies in the average overtime boarding (i.e. the average time that boarding patients are blocked in the ED because they are still waiting for a physician check-up while there is a bed available for them in the IW). As shown in Figure 10, the

average overtime boarding increases sharply for patient type 1 when going from an $AP(0.2)$ to an $AP(0)$ priority policy. Reflecting this sharp increase, $E[LOS_{incl_b}]$ also increases and therefore eliminates $AP(0)$ as an optimal priority policy. In contrast, $E[LOS_{exl_b}]$ is not influenced by the overtime boarding as it does not account for the boarding time. Hence, we argue that $E[LOS_{incl_b}]$ is more appropriate as a performance measure when deciding on the applied priority policy.

Figure 10: (a) $E[LOS_{exl_b}]$, (b) $E[LOS_{incl_b}]$ and (c) $E[overtime boarding]$ in $\{BP_{b+p}, (\cdot), Staffing_LQoS, Bed_LU\}$.

In order to better understand the evolution of $E[LOS_{incl_b}]$ for different priority policies, we investigate its underlying factors: $E[W_e]$ and $E[W_{phys}]$. It is clear from Figure 11 and Figure 12 that $E[W_e]$ and $E[W_{phys}]$ follow a different pattern regarding $AP(\frac{z_b}{z_t})$ priority policies. Indeed, as could also be concluded from Table 6, minimizing $E[W_e]$ results in giving more priority to test patients (low values for $\frac{z_b}{z_t}$), whereas minimizing $E[W_{phys}]$ favours prioritizing more boarding patients (high values for $\frac{z_b}{z_t}$).

Figure 11: (a) $E[W_{phys}]$ and (b) $E[W_e]$ in $\{BP_{b+p}, (\cdot), Staffing_LQoS, Bed_LU\}$.

Figure 12: (a) $E[W_{phys}]$ and (b) $E[W_e]$ in $\{BP_{b+p}, (\cdot), Staffing_HQoS, Bed_LU\}$.

On the one hand, $E[W_e]$ shows a non-linear increasing trend from AP(0.2) onwards as $\frac{z_b}{z_t}$ increases for patient types 1, 3 and 4 in $\{BP_{b+p}, (\cdot), Staffing_LQoS, Bed_LU\}$ (Figure 11(b)). Contrarily, going from AP(0) to AP(0.2) decreases $E[W_e]$. This is due to a trade-off between the bed blocking effect of boarding patients and that of test patients. As is clearly shown in Figure 11(b), the bed blocking effect of test patients dominates the bed blocking effect of boarding patients, resulting in an increase in $E[W_e]$ in $\{BP_{b+p}, (\cdot), Staffing_LQoS, Bed_LU\}$ when giving more priority to boarding patients (higher values of $\frac{z_b}{z_t}$). Only when going from AP(0.2) to AP(0), the bed blocking effect of boarding patients becomes more outspoken, implying the initial decrease in $E[W_e]$ observed in Figure 11(b). This is only the case

for patient types 1, 3 and 4; for patient type 2, the bed blocking effect of test patients is dominant, regardless the value of $\frac{z_b}{z_t}$.

On the other hand, $E[W_{phys}]$ follows a U-shape for patient types 1, 3 and 4 in $\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_LU}\}$, which becomes more outspoken as $E[\rho_{phys}]$ increases (Figure 11(a)). Contrarily, a decreasing trend in $E[W_{phys}]$ is observed for all patient type in $\{BP_{b+p}, (\cdot), \text{Staffing_HQoS}, \text{Bed_LU}\}$ when increasing $\frac{z_b}{z_t}$ (Figure 12(a)). The explanation is again found in the observed values for $E[\rho_{phys}]$. When $E[\rho_{phys}]$ stays below a certain threshold value, $E[W_{phys}]$ is minimized by always giving priority to boarding patients (i.e. applying the AP(9999), or equivalently the B, priority policy). This is due to the fact that the average processing time of boarding patients is much smaller than the one of test patients. Interestingly, however, is that when a certain threshold value of $E[\rho_{phys}]$ is reached, $E[W_{phys}]$ starts to increase again when giving too much priority to boarding patients. This effect becomes more outspoken as $E[\rho_{phys}]$ further increases, ultimately eliminating AP(9999) as an optimal priority policy for minimizing $E[W_{phys}]$. This is explained by the fact that prioritizing boarding patients increases their average number of physician visits (e.g. Figure 13(a)). Indeed, every time boarding patients return from the physician and their boarding time is not over yet, they generate a new check-up at rate μ_b . In contrast, since test patients return to the physician with probability p_1 , independent from the applied priority policy, prioritizing them does not impact their physician visits at all (e.g. Figure 13(b)). When reaching a certain threshold of $E[\rho_{phys}]$, the increase in boarding patients' physician visits results in a congestion of the physician queue and hence in an increase in $E[W_{phys}]$. This threshold value of $E[\rho_{phys}]$ differs between patient types and is thus dependent on the patient characteristics.

Figure 13: (a) $E[\text{number of boarding patients' physicians visits}]$ and (b) $E[\text{number of test patients' physicians visits}]$ in $\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_LU}\}$.

So, the impact of increasing $\frac{z_b}{z_t}$ on $E[LOS_{incl_b}]$ is the result of a trade-off between an increasing trend in $E[W_e]$ and a decreasing trend in $E[W_{phys}]$. However, the effect of $E[W_e]$ is more outspoken and is therefore the key driver of the evolution in $E[LOS_{incl_b}]$. Since focusing solely on $E[W_e]$ or $E[W_{phys}]$ results in contradicting recommended priority policies and disregards patients in the physician queue and external queue respectively, we argue that $E[LOS_{incl_b}]$ is better suited as a performance measure for determining the priority policy. That is, $E[LOS_{incl_b}]$ takes into account both $E[W_e]$ and $E[W_{phys}]$, and provides a way of making a trade-off between the two.

Hence, based on the observations in this section, we propose the following strategy to decide on an optimal priority policy for our case study. We solely consider $E[LOS_{incl_b}]$ and service levels $P(W_{phys} = 0)$, $P(W_{phys} \leq 30 \text{ min})$, $P(W_e = 0)$, and $P(W_e \leq 30 \text{ min})$. Moreover, we only take into account a service level in the analysis if it is below 90% for at least one value of $\frac{z_b}{z_t}$ in the considered scenario. Applying this strategy allows us to reduce the number of optimal priority policies for our case study to one or two per scenario (Table 8).

Table 8: Optimal priority policy for different scenarios, patient types and performance measures $E[LOS_{incl_b}]$, $P(W_{phys} = 0)$, $P(W_{phys} \leq 30 \text{ min})$, $P(W_e = 0)$, and $P(W_e \leq 30 \text{ min})$.

Scenario	Patient type	Optimal priority policy
$\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_HU}\}$	Patient type 1	AP(0)
	Patient type 2	AP(0)
	Patient type 3	AP(0)
	Patient type 4	AP(0)
$\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_LU}\}$	Patient type 1	AP(0) or AP(0.2)
	Patient type 2	AP(0)
	Patient type 3	AP(0)
	Patient type 4	AP(0)
$\{BP_{b+p}, (\cdot), \text{Staffing_HQoS}, \text{Bed_HU}\}$	Patient type 1	AP(0) or AP(0.2)
	Patient type 2	AP(0)
	Patient type 3	AP(0)
	Patient type 4	AP(0)
$\{BP_{b+p}, (\cdot), \text{Staffing_HQoS}, \text{Bed_LU}\}$	Patient type 1	AP(0) or AP(0.2)
	Patient type 2	AP(0)
	Patient type 3	AP(0)
	Patient type 4	AP(0)

As Table 8 indicates, it is always optimal in our case study to give more priority to test patients (low values of $\frac{z_b}{z_t}$). For patient types 2, 3, and 4, all considered performance measures are optimized by

always prioritizing test patients (applying an AP(0), or equivalently a T, priority policy). For patient type 1, the optimal value of $\frac{z_b}{z_t}$ mostly depends on the considered performance measure; $E[LOS_{incl_b}]$, $P(W_e = 0)$, and $P(W_e \leq 30 \text{ min})$ are optimized by applying an AP(0.2) priority policy, whereas focusing on $P(W_{phys} = 0)$ and $P(W_{phys} \leq 30 \text{ min})$ results in an AP(0) priority policy. Only in $\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_HU}\}$, AP(0) is the optimal priority policy for patient type 1, regardless of the considered performance measure.

If we look at individual service levels of test and boarding patients, it can be seen that boarding patients are not strongly disadvantaged by the application of an AP(0) or AP(0.2) priority policy (e.g. Figure 14). On the one hand, $P(W_{phys} = 0)$ for boarding patients is maximized by prioritizing test patients (e.g. Figure 14(a)). On the other hand, for $P(W_{phys} \leq 30 \text{ min})$, a trade-off is observed between test and boarding patients; giving more priority to one type penalizes the other. Moreover, test patients suffer more from priority policies that favour boarding patients than the other way around. Hence, we conclude that it is optimal in our case study, and acceptable for both test and boarding patients, to give more priority to test patients in order to optimize ED system performance.

Figure 14: (a) $P(W_{phys} = 0)$ and (b) $P(W_{phys} \leq 30 \text{ min})$ in $\{BP_{b+p}, (\cdot), \text{Staffing_LQoS}, \text{Bed_HU}\}$.

4.4. Discussion and generalizability of numerical results

The numerical results presented in the previous sections rely on the assumptions and input data of our case study. As noted in Section 3.1, several modelling assumptions (i.e., exponentially distributed service times, memoryless routing probabilities, test patient homogeneity within each zone) deviate from observations as generally reported in practice. This implicates that while some conclusions are generalizable, others might be case specific.

A first conclusion states that omitting needy boarding patients from any ED modelling is very likely to misrepresent the ED system's performance. Moreover, since the choice of priority policy might have an effect on ED system performance, it is worthwhile to investigate the existence of an optimal priority policy. This conclusion is generalizable beyond our case study since a more complex ED system is likely to be even less forgiving if needy boarding patients were to be ignored.

A second conclusion is about recommending a priority policy. These recommendations are based on the numerical modelling results that, in turn, are based on the case-specific modelling assumptions. Consequently, the recommended priority policies in our case study might not necessarily be generalizable to other ED settings.

Altogether, the presented numerical results clearly demonstrate that needy boarding patients can significantly impact system performance and hence should be accounted for in the analysis and planning of an emergency department. Also, the analysis indicates that simulation is an appropriate method for analysing priority policies in EDs.

5. Conclusion and future research

This paper investigates the impact of needy boarding patients and the prioritization decision of physicians between test and boarding patients in an ED. It does so by analysing and comparing scenarios with varying types of boarding patients, priority policies, staffing levels and bed utilizations for various performance measures using discrete-event simulation.

The main contribution of this paper is that it quantitatively demonstrates the relevance of incorporating needy boarding patients into the analysis and the importance of defining a priority policy for the ED physicians. These are two areas often neglected in current literature. When incorporating needy boarding patients, the prioritization decision has significant consequences for ED system performance, especially if the physician utilization is already high without the returning boarding patients. In our case study, it is optimal to prioritize test patients. Dependent on patients' characteristics, this is done by *always* giving priority to test patients or by only giving priority to boarding patients if their waiting time in the physician queue becomes very long. In this way, ED system performance is optimized.

Following these results, several options for future research emerge. First of all, we focus on an ED where patients are treated in separated zones with dedicated physicians and beds. This might be a realistic setting for large hospitals, but is mostly not feasible for small ones. Indeed, in many cases (including small hospitals), patients with varying characteristics need to be treated in one zone. Therefore, analyzing the priority policy in a system with several categories of test and boarding patients provides an opportunity for future research. Also, case studies that relax the assumptions of exponential service times and memoryless routing probabilities might provide more generalizable insights on recommended priority policies. This will, however, require better data on the bed utilization and

physician workload boarding patients require. Future research might address this gap by conducting case studies and reporting statistical analysis of boarding patients' characteristics in EDs.

Secondly, the impact of needy boarding patients on service levels, and hence delay probabilities, is significant. Since staffing rules often rely on target delay probabilities, taking into account needy boarding patients when setting staffing levels might be an interesting research topic.

Acknowledgements

This research was supported by the GlaxoSmithKline Research Chair on Re-Design of Healthcare Supply Chains in Developing Countries to increase Access to Medicines. We are grateful for the feedback of the editor and the anonymous reviewers that greatly improved the quality of this paper.

References

- Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseurtlin, Y., & Yom-Tov, G. B. (2015). On patient flow in hospitals: a data-based queueing-science perspective. *Stochastic systems*, 5(1), 146-194.
- Au, L., Byrnes, G.B., Bain, C.A., Fackrell, M., Brand, C., Campbell, D.A., & Taylor, P. G. (2008). Predicting overflow in an emergency department. *IMA Journal of management mathematics*, 20(1), 39-49.
- Bair, A.E., Song, W.T., Chen, Y.-C., & Hsieh, B.A. (2010). The impact of inpatient boarding on ED efficiency: A discrete-event simulation study. *Journal of Medical Systems*, 34(5), 919-929.
- Bagust, A., Place, M., & Posnett, J. W. (1999). Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. *British Medical Journal*, 319(7203), 155-158.
- Carmen, R., Van Nieuwenhuyse, I., & Van Houdt, B. (2018). Inpatient boarding in emergency departments: Impact on patient delays and system capacity. *European Journal of Operational Research*, 271(3), 953-967.
- Carmen, R., Defraeye, M., & Van Nieuwenhuyse, I. (2015). A decision support system for capacity planning in emergency departments. *International Journal of Simulation Modelling*, 14(2), 299-312.
- Crawford, E.A., Berlin, P.J., Kong, N., & Thakar, C.V. (2014). Analyzing discharge strategies during acute care: A discrete-event simulation study. *Medical decision making*, 34(2), 231-241.
- Defraeye, M., & Van Nieuwenhuyse, I. (2013). Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. *Decision Support Systems*, 54(4), 1558-1567.
- Fajardo, V. A., & Drekić, S. (2017). Waiting time distributions in the preemptive accumulating priority queue. *Methodology and Computing in Applied Probability*, 19, 255-284.
- Feldman, Z., Mandelbaum, A., Massey, W.A., & Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2), 324-338.

Ferrand, Y.B., Magazine, M.J., Rao, U.S., & Glass, T.F. (2018). Managing responsiveness in the emergency department: Comparing dynamic priority queue with fast track. *Journal of Operations Management*, 58-59, 15-26.

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

Günel, M.M., & Pidd, M. (2010). Discrete event simulation for performance modelling in health care: a review of the literature. *Journal of Simulation*, 4(1), 42-51.

Hoot, N. R., & Aronsky, D. (2008). Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions. *Annals of Emergency Medicine*, 52(2), 126-135.

Khanna, S., Boyle, J., Good, N., & Lind, J. (2012). Unravelling relationships: Hospital occupancy levels, discharge timing and emergency department access block. *Emergency Medicine Australasia*, 24(5), 510-517.

Khare, R.K., Powell, E.S., Reinhardt, G., & Lucenti, M. (2009). Adding more beds to the emergency department or reducing admitted patient boarding times: Which has a more significant influence on emergency department congestion? *Annals of Emergency Medicine*, 53(5), 575-585.

Li, N., & Stanford, D.A. (2016). Multi-server accumulating priority queues with heterogeneous servers. *European Journal of Operational Research*, 252(3), 866-878.

Luo, W., Cao, J., Gallagher, M., & Wiles, J. (2013). Estimating the intensity of ward admission and its effect on emergency department access block. *Statistics in Medicine*, 32(15), 2681-2694.

Mahajan, P.S., & Ingalls, R.G. (2004). Evaluation of methods used to detect warm-up period in steady state simulation. *Proceedings of the 2004 Winter Simulation Conference*, 663-671.

Moskop, J.C., Sklar, D.P., Geiderman, J.M., Schears, R.M., & Bookman, K.J. (2009). Emergency Department Crowding, Part 1—Concept, Causes, and Moral Consequences. *Annals of Emergency Medicine*, 53(5), 605-611.

Mustafee, N., Lyons, T., Rees, R., Davies, L., Ramsey, M., & Williams, M. D. (2012). Planning of bed capacities in specialized and integrated care units: Incorporating bed blockers in a simulation of surgical throughput. *Proceedings of the 2012 Simulation Conference*, 1-12.

Rashwan, W., Al-Hama, W. & Arisha, A. (2015). A system dynamics view of the acute bed blockage problem in the Irish healthcare system. *European Journal of Operational Research*, 247(1), 276-293.

Saghafian, S., Austin, G., & Traub, S.J. (2015). Operations research/management contributions to emergency department patient flow optimization: review and research prospects. *IIE Transactions on Healthcare Systems Engineering*, 5(2), 101-123.

Saghafian, S., Hopp, W.J., Van Oyen, M.P., Desmond, J.S., & Kronick, S.L. (2012). Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60(5), 1080-1097.

Saghafian, S., Hopp, W.J., Van Oyen, M.P., Desmond, J.S., & Kronick, S.L. (2014). Complexity-augmented triage: a tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management*, 16(3), 329-345.

SEE center, Technion – Israel Institute of Technology. <https://seeserver.ima.technion.ac.il/see-terminal/contacts.aspx>

Sharif, A.B., Stanford, D.A., Taylor, P., & Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care*, 3(2), 73-79.

Shi, P., Chou, M.C., Dai, J.G., Ding, D., & Sim, J. (2016). Models and insights for hospital inpatient operations: time-dependent ED boarding time. *Management Science*, 62(1), 1-28.

Stanford, D.A., Taylor, P., & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 77(3), 297-330.

Yom-tov, G.B., & Mandelbaum, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2), 283-299.

Zeltny, S., Marmor, Y.N., Mandelbaum, A., Carmeli, E., Greenspan, O., Mesika, Y., et al. (2011). Simulation-based models of emergency departments. Operational, tactical, and strategic staffing. *ACM Transactions on Modelling and computer simulation (TOMACS)*, 21(4), 1-25.