6-2021

Analytics for hospital resource planning: Two case studies

Jingui XIE
Brunel University

Weifen ZHUANG
Xiamen University

Marcus ANG
Singapore Management University, marcusang@smu.edu.sg

Mabel C. CHOU
National University of Singapore

Li LUO
Sichuan University

See next page for additional authors

Citation

XIE, Jingui; ZHUANG, Weifen; ANG, Marcus; CHOU, Mabel C.; LUO, Li; and YAO, David D.. Analytics for hospital resource planning: Two case studies. (2021). *Production and Operations Management*. 30, (6), 1863-1885.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/6525

Analytics for Hospital Resource Planning—Two Case Studies

Jingui Xie
Brunel Business School, Brunel University London, Uxbridge, UB8 3PH, UK, jingui.xie@brunel.ac.uk
School of Management, University of Science and Technology of China, Hefei, 230062, China, xiej@ustc.edu.cn

Weifen Zhuang*
School of Management, Xiamen University, Xiamen 361005, China, wfzhuang@xmu.edu.cn

Marcus Ang
Lee Kong Chian School of Business, Singapore Management University, Singapore City 178899, Singapore, marcusang@smu.edu.sg

Mabel C. Chou
NUS Business School and Institute of Operations Research and Analytics, National University of Singapore, Singapore City 119245, Singapore, bizchoum@nus.edu.sg

Li Luo
Business School, Sichuan University, Chengdu 610064, China, luolicc@scu.edu.cn

David D. Yao
IEOR Department, Columbia University, New York, New York 10027, USA, yao@columbia.edu

Using real data and process flows from two large hospitals (in Singapore and in Chengdu, China) as cases, we illustrate how to apply certain modeling and optimization techniques, along with simulation as a validation tool, to hospital resource planning problems. We demonstrate how these simple analytical tools can help achieve significant improvements in both patient service and resource utilization, and without the need to increase the overall level of existing capacities. Two resource planning problems are studied in detail, one concerns the rebalancing of bed capacities among various wards, and the other addresses the allocation of medical diagnostic resource among different types of patients.

Key words: resource allocation; optimization; asymptotic optimality
History: Received: August 2019; Accepted: December 2019 by J. George Shanthikumar, after 1 revision.

1. Introduction

Critical health care delivery mostly takes place in a large and complex system, the hospital, involving expensive resources and highly skilled workers such as doctors, nurses and other medical professionals. Yet, many of today's hospitals have yet to utilize and benefit from the kind of tools and technologies that run modern manufacturing plants such as a semiconductor wafer fab, let alone all the research and innovations that have gone into enterprise supply chain management over the last two decades embodied broadly in the *analytics* paradigm. Healthcare delivery, according to Atul Gawande ("Big Med." *New Yorker*, August 16, 2012)—a distinguished surgeon, author, and public-health researcher, has a host of principles and practices to learn from a restaurant chain, Cheesecake Factory. These include the basic tenet of any successful business: offering a quality product/service or experience to customer at an affordable price; the standardization of routine treatments and procedures so as to improve efficiency and reduce costs; and the adoption of "systems" concepts and technologies so as to provide effective solutions to the entire supply chain—in this case, the health care value chain.

The goal of this study is to demonstrate, through case studies at two large hospitals, how applying some simple modeling and optimization techniques, along with simulation as a validation tool, to hospital resource planning problems can achieve significant improvements in both patient service and resource utilization, without the need to increase the existing capacity.

Please Cite this article in press as: Xie, J., et al. Analytics for Hospital Resource Planning—Two Case Studies. *Production and Operations Management* (2020), https://doi.org/10.1111/poms.13158

In the first case, we study the rebalancing of bed capacity among various wards at a large hospital in Singapore, referred to as Hospital-1 (or the first case) below. Data reveals that while the average bed occupancy rate (BOR) is around 86%, there is a big variation among the wards, due to demand changes over the years. This supply–demand mismatching leads to a high overflow rate: on average 18.91% of emergency patients are wrong-sited (the highest overflow rate being 68.67%), that is, these patients have to be sent to some ward that is not the best choice for their treatment. In addition, it causes long waiting time (weeks or months) for elective patients to be admitted.

With a simple optimization model, we perform a load balancing among wards, that is, optimize the number of beds allocated to each ward, while maintaining the total bed capacity at its existing level. This reallocation reduces the overflow rate from 18.91% to 4.5%. Furthermore, even if we allow an increase in elective patient admissions by 15%, the optimal reallocation can still maintain the overflow rate at around 6%–7%, and keep the waiting time of emergency patients at about 0.5 hours, well within the required 6-hour limit.

The second case concerns medical diagnostic facilities (such as CT and MRI) at another hospital, located in Chengdu, China, referred to as Hospital-2 (or the second case) below. Such facilities are shared among emergency patients, inpatients, and outpatients, with emergency patients given priority service, immediately upon arrival, while inpatients and outpatients are served by appointment only. The hospital's practice is to set daily quota for outpatient and inpatient appointments (denoted n_1 and n_2)—once the quota is used up, appointments will be deferred to the following day(s); while the remaining capacity is reserved for emergency patients ($n_3 = N - n_1 - n_2$, with N being the total number of available slots per day).

Our approach is to first optimize the reserved capacity for emergency patients, n_3^*. We then apply a "nested" policy among inpatients and outpatients to share the remaining $N - n_3^*$ slots, that is, with an optimized quota n_1^* limiting the outpatient appointments, but allowing inpatients to have access up to all $N - n_3^*$ slots. This optimization leads to a 14% improvement in same-day service for inpatients and outpatients; and for the delayed appointments, a 33% reduction in patient-day counts. It also improves the utilization of the facility by 10% and reduces overtime by 11%.

The key to the first case is a *square-root allocation rule*, by which the number of beds allocated to any given ward i is $\rho_i + \beta_i\sqrt{\rho_i}$, where ρ_i is the traffic intensity (patient arrival rate times the average length of stay at the ward), and β_i is a decision variable to be optimized. When the objective is to minimize the maximal delay among all wards and with all wards carrying equal weight, we show the optimal allocation is an "equal β rule," that is, to set $\beta_i = \beta$ for all i. Variations of this square-root allocation rule have been widely used in call-center staffing. It comes from what is known as the Erlang-C formula associated with the $M/M/c$ queueing model; refer to the Appendix A.1.

In the second case, the nested policy plays the key role. To construct the policy, we first derive upper- and lower-bounds on the optimal value of the dynamic programming (DP) solution to the allocation problem. (The DP solution has a switching curve structure, with the curves depending on both time and states; as such its practical implementation is questionable.) In particular, n_3^*, the capacity reserved for emergency patients comes from the upper-bound solution, and n_1^*, the quota on outpatients, comes from the lower-bound solution. Interestingly, similar to the first case, there is also a square-root phenomenon here. Let D_i, for $i = 1, 2, 3$, denote the demand (daily total) of the three types of patients; and suppose $\mathbf{E}(D_i) = \lambda_i T$, with T being a scaling factor, and $\lambda_i > 0$ a constant parameter. Then, we show that the upper- and lower-bounds on the optimal DP solution coincide on a term that is of order $O(T)$, while differing on a second term of order $O(\sqrt{T})$. As the nested policy falls in between the two bounds (and so does the optimal DP solution), it is asymptotically optimal when $T \to \infty$, that is, the gap from optimality vanishes when T is large.

The square-root factor in both cases is not coincidental. Fundamentally, it is rooted in the central limit theorem: the sum of i.i.d. random variables $X_1 + \cdots + X_T$, with $\mathbf{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2$, can be approximated, when T is large, by a normal random variable, $\mu T + \sigma\sqrt{T}Z$ with Z following the standard normal distribution. The demands (patient number counts) in both cases can be approximated by normal distributions when their means are reasonably large (signifying high-volume demand on heavily utilized resources). To the extent that each demand brings along a revenue (if served) or a cost/penalty (if denied), the objective value will scale accordingly. Specifically, there will be a first-order (mean) term that is of order $O(T)$, and a second-order (variability or standard deviation) term of order $O(\sqrt{T})$. The lesson learned from the first case is that the allocation of the beds among the various wards is to first assign ρ_i beds to ward i, strictly according to their traffic intensity (first-order), and then assign any remaining beds to each ward i proportional to $\sqrt{\rho_i}$, which measures variability. In the second case, the insight is that both the upper- and lower-bound solutions achieve the same first-order $O(T)$ value as the DP optimal solution, and so does the nested policy. This provides a

performance guarantee for the proposed nested policy, which has the additional advantage of easy to implement (whereas the DP solution is significantly more complex as mentioned above).

Aside from these similarities, there are significant differences in the two cases. The main challenge in the first case comes from the cross-referenced wards and clusters and the intertwined patient flows and overflow priorities (refer to details below), and these complications are clearly reflected in the complexity of the data sets. Accordingly, much of our effort is spent on studying the data and identifying a suitable model that aggregates the 19 wards into 8 "super wards." Once this aggregation is done, the reallocation of the beds is readily solved as convex optimization problems. The second case, in contrast, has a clear process flow along with a relatively clean-cut dataset, which readily calls for a DP formulation as overviewed above. Yet, the challenge is to adapt the DP optimal policy into a solution that can be practically implemented at the hospital, comparable to its existing practice in simplicity but with improved performance. Our exposition below is clearly influenced by the differences between the two cases, and follows closely the actual steps we went through in the two studies: from data to models and solutions in the first case, whereas from models and solutions to heuristics (the nested policy) so as to adapt to the process reality in the second case.

In what remains of this introductory section, we present a brief overview of the related literature. This is followed by two sections, Sections 2 and 3, detailing the two cases, respectively. Concluding remarks are summarized in section 4. To ease the flow of exposition, as well as to facilitate reading, materials that are more specialized in nature are collected in the Appendix A.

1.1. Literature Review

The coordination of emergency, elective, and other patients so as to better utilize hospital resources has been extensively studied in the operations management literature. For example, Armony et al. (2015) and Hall et al. (2006) conduct detailed studies of patient flows in various departments at an Israeli and a US hospital, respectively. Discrete-event simulation and queuing theory are two commonly used approaches for modeling and improving patient flows (see, e.g., Allon et al. 2013, Green 2006, Green et al. 2006b, Jacobson et al. 2006, Zeltyn et al. 2011).

Compared to the rich literature on patient flow models of emergency department (e.g., Huang et al. 2015), and appointment-scheduling in outpatient clinics and other hospital facilities (e.g., Green and Savin 2008, Green et al. 2006a, Kong et al. 2013), inpatient flow management and the interface between emergency department and inpatient wards have received less attention; see the same discussion in section 4 of Armony et al. (2015). Related works on inpatient operations include capacity allocation and flow improvement in specialized hospitals or wards (Bavafa et al. 2018, Cochran and Bharti 2006, Deo et al. 2013, Green 2004, Green and Nguyen 2001, Griffin et al. 2012), ward nurse staffing (de Véricourt and Jennings 2011, Yankovic and Green 2011), bed assignment (Mandelbaum et al. 2012, Thompson et al. 2009), admission control and design (Helm and Van Oyen 2014, Helm et al. 2011, Kim et al. 2014), and discharge policy (Chan et al. 2012, Shi et al. 2015).

Shi et al. (2015) focus on understanding the effect of inpatient discharge policies and other operational policies on the time-of-day waiting time performance, such as the fraction of patients waiting for more than six hours at the emergency department before being admitted, and studying the impact of discharge and other operational policies, whereas our study aims at analyzing the impact of supply–demand imbalance among the hospital wards, and the benefit of reallocating the bed capacities in improving overflow rates and waiting times. Also note that most existing works focus on acute beds and intensive care units (ICUs) (e.g., Chan et al. 2012, Costa et al. 2003, Kim et al. 2000, Kim et al. 2014), and only a few were on general wards where most inpatient beds are located in.

The study of medical diagnostic facility management is related to two streams of literature, appointment scheduling in health care and demand management (revenue management). For comprehensive reviews on appointment scheduling and resource allocation in health care, see Magerlein and Martin (1978), Smith-Daniels et al. (1988), Chapman and Carmel (1992), (Cayirli and Veral 2003), Mondschein and Weintraub (2003), Gupta (2007), Gupta and Denton (2008), Jack and Powers (2009), Cardoen et al. (2010), Guerriero and Guido (2011), and May et al. (2011).

The first stream of literature can be classified into *single-day scheduling* and *multi-day scheduling*. Our study of medical diagnostic resource allocation is related to the single-day framework of Green et al. (2006a) and Gupta and Wang (2008), while it differs in many significant ways. The heuristic scheduling policies they proposed do not explicitly concern the multi-day applications with demand volumes dependent on the day of the week. For a single-day application, the "newsvendor policy" of Green et al. (2006a) and the heuristic policy from the newsvendor solution of Gupta and Wang (2008) are similar to our lower-bound solution, which we have shown to be dominated by our nested-partition policy. Two other heuristics of Green et al. (2006a) can be viewed as special cases of our nested-partition policy. (For instance, "fill all slots" is just fully pooling, which specialize to

letting $n_1 = \bar{N}$ in our nested-partition policy. In contrast, our choice of n_1 and $\bar{N}$ is optimized via the lower-bound problem.) Furthermore, our proposed nested-partition policy can be applied on a rolling-horizon manner to a multi-day setting.

Our work is also in the same spirit as the multi-day scheduling literature (e.g., Erdelyi and Topaloglu 2011, Gerchak et al. 1996, Huh et al. 2013, Patrick et al. 2008, Sauré et al. 2015, Truong 2015), as we address a detailed appointment scheduling problem from a resource allocation point of view, and in doing so, bring out the key tradeoff between reserving capacity for emergency patients and risking unused (wasted) resources during regular hours. We contribute to the literature a different approach (a nested-partition via the lower- and upper-bound solution to the DP model), with proven asymptotic optimality, and with additional new insights into the tradeoff between reserving capacity for emergency patients and serving outpatients and inpatients the best one can.

Another related stream of literature is demand management or revenue management (RM), which is essentially dynamic resource allocation among multiple classes of stochastic demands. Service capacity in health care is constrained and perishable while demand is fluctuating, hence it meets the necessary conditions for an effective demand management (revenue management) (Kimes 1989). For the single-resource multi-class RM problem, the two most widely used and well-performing heuristics EMSR-a (Belobaba 1987) which aggregates protection levels pairwise and EMSR-b (Belobaba 1989) which aggregates weighted-average demand by taking into account the pooling effect, are merely feasible solutions to our lower-bound problem which, in turn, is dominated by our proposed nested-partition policy.

2. Bed-Capacity Allocation at Hospital-1

Hospital-1 is a major hospital in Singapore. Over the last 15 years or so, it has witnessed a substantial expansion of its facilities and capacities along with a steady increase in their demand. As bed capacity increases over time, its allocation among different wards has not always been mapped to the demand dynamics, certainly not in any systematic manner. As in any organization, there are turf barriers and administrative hurdles. Once a (new) set of beds is allocated to a certain ward, it will be difficult if not impossible to reallocate them to other wards, even when shifts in demand clearly warrant such reallocation. The mismatch in supply and demand results in patients being wrong-sited, meaning they have to be sent to wards that are different from their primary wards where they would be best treated. This stream of wrong-sited patients is referred to as "overflow," and the substantial rate of overflows has over the years become a serious performance issue. Further exacerbating the problem is the hospital's obligation to meet the standard of the (Singapore) Ministry of Health on the emergency department boarding time. Under this time pressure, the hospital management is more willing to overflow emergency patients as opposed to waiting for beds to become available in their primary wards.

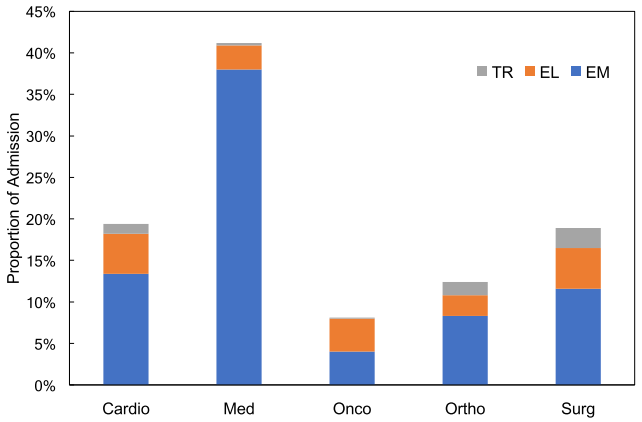
To illustrate, consider the cardiology department. In the dataset that we are given access to, its annual total of admitted patients is 8585, while the average number of beds over the same period is 61, and the average length of stay is 4.02 days, which translates into a required capacity of 95 beds. Thus, it is no surprise that the cardiology department has a large overflow rate, 37.55%, which is the highest among all departments in our study. Oncology is better, but it has a different kind of problem. The annual total of admitted oncology inpatients is 3572, and the average length of stay is 6.09 days, which means that 60 beds are required. The department has exactly this number of beds; however, there are still 690 (or, 19.32%) oncology patients assigned to other wards, which is the second highest overflow rate in our study. The reason for this is many of its beds are occupied by patients overflowed from other departments, so that its own patients also need to be overflowed.

2.1. Data, Facts and Problems

In our study, we focus on five major clusters at Hospital-1: medicine, surgery, cardiology, orthopedic, and oncology. (A cluster refers to a broad medical specialty, under which there are several wards, often with further division of subspecialties.) These five clusters have a total of 629 beds, across 19 wards, and account for a total of 44,075 admissions (over the year that we focus on). The clusters we left out include obstetrics and gynecology, pediatrics, and ophthalmology, since they are quite independent, in the sense that no beds are shared with other clusters.

There are three sources of inpatient admissions. Emergency (EM) patients start from visiting the emergency department, and then admitted by emergency department physicians. Elective (EL) patients are admitted from the clinics by outpatient specialists. Internal transfer (TR) patients are from other wards (i.e., outside of the 19 wards that we focus on), such as ICU wards, isolation wards, high-dependence wards, etc. The breakdown of the inpatients is 75.28% EM, 19.12% EL, and 5.60% TR. Their distribution over the five clusters is displayed in Figure 1. Observe that among the five clusters, Medicine has the largest proportion of EM admissions, followed by Cardiology and Surgery.

Figure 1 Admission Proportion of Five Major Clusters and Three Patient Types [Color figure can be viewed at wileyonlinelibrary.com]



Once admitted, the number of days a patient spends in a ward is called length of stay (LOS). It differs by wards and by cluster; the relevant data are summarized in Table 1. The last two columns in the table report the number of beds (or, bed capacity) in each ward and their utilization (bed occupancy rate, or BOR). The BOR reported in the table varies from 66.48% to 97.33%. When averaged over all wards, the BOR is 85.87%. While this utilization level is not excessively high, the variation (or imbalance) among the wards is the main reason for a high level of overflows, as detailed below.

The detailed accounting of overflow is presented in Table 2. In the left (first) column, each ward carries in parentheses its association with one (or in some cases, two) clusters. For example, ward 41 is associated with both surgical (S) and cardiology (C) clusters; and the first row records the head count of patients admitted into ward 41, breaking down by the clusters. Out of the total of 2465 patients, those from outside of the surgical and cardiology clusters are counted as overflow, a total of 323, or 13.1%. To understand the columns, consider the third one, the medicine cluster. Listed in the column are patient head counts from all the wards, and the total is 18,144 (see the "Total" figure at the bottom). Since only eight wards (42, 44, 53, 55, 57, 64, 66, 76) belong to this cluster, patients in other wards are counted as overflow, and that is the 2644 figure, mapping to an overflow rate of 14.57%.

Data and statistics in Table 2 are further illustrated in Figure 2 and in Figure 3, with the latter also presenting a breakdown among EM, EL, and TR patients. From Table 2, we can see that Cardiology and Medicine have the largest number of overflowed inpatients, while Cardiology and Oncology have the highest overflow rates. The total number of overflowed patients is 8335, and the overall overflow rate is 18.91%. Among those overflows, there are 6178 (14.02%) EM patients, 1760 (3.99%) EL patients, and 397 (0.90%) TR patients. As shown in Figure 2, NW52, 54, 76, and 78 have overflow rates more than 30%. In particular, overflow in NW76 is 68.67%.

In summary, the biggest problem as revealed from the data is the high overflow rate: about 18.91% of patients are wrong-sited. Overflow not only leads to poor quality of patient care and service, it also increases the workload of physicians and nurses (just the extra time to walk to a remotely located ward is non-trivial) and leads to high cost for management.

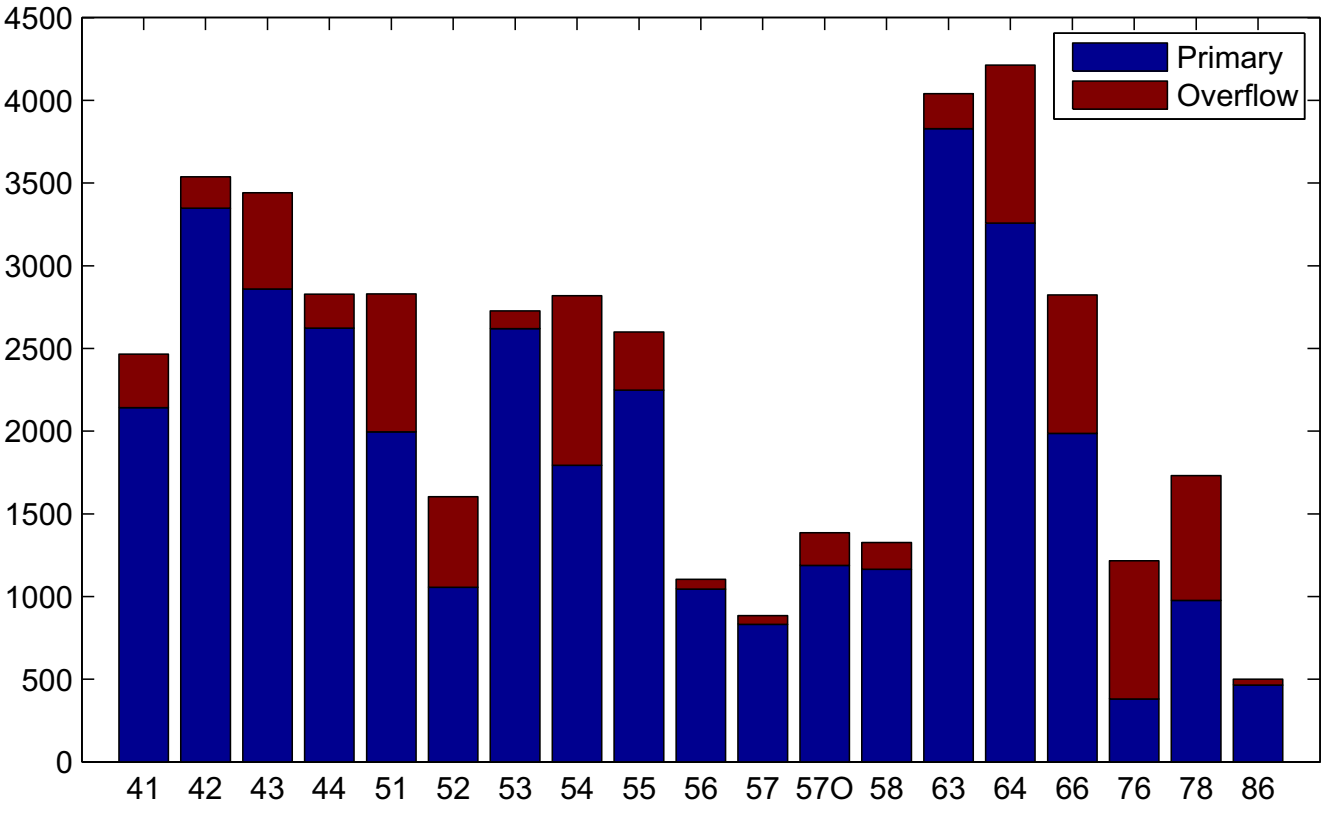
Table 1 Length of Stay (LOS), Bed Capacity, and Bed Occupancy Rate (BOR) in General Wards

	LOS							
Wards	Cardio	Med	Onco	Ortho	Surg	All clusters	Ave. bed capacity	BOR
NW41	4.98	2.93	3.48	4.10	5.83	5.34	43.58	82.68%
NW42	2.38	4.26	3.11	5.54	2.46	4.18	44.00	92.07%
NW43	4.34	2.55	2.38	4.17	3.89	3.82	41.24	87.39%
NW44	3.76	4.34	2.63	3.16	3.18	3.96	38.68	79.31%
NW51	2.19	1.90	4.00	3.85	2.30	3.34	39.00	66.48%
NW52	2.47	2.67	1.96	5.21	2.90	4.34	28.66	66.55%
NW53	3.90	6.03	3.11	13.00	2.94	5.98	45.92	97.33%
NW54	3.49	2.53	7.45	6.15	4.98	4.83	48.50	76.83%
NW55	3.43	5.16	3.16	1.75	3.53	4.92	40.64	86.14%
NW56	5.45	3.18	2.91	N.A.	2.75	5.33	17.00	94.79%
NW57	3.77	5.63	2.83	16.00	N.A.	5.49	14.00	95.05%
NW57O	3.92	2.59	6.43	1.00	2.86	5.91	23.92	93.78%
NW58	2.29	2.15	6.52	2.33	8.67	6.03	24.00	91.34%
NW63	3.85	2.65	3.15	6.75	1.56	3.79	43.59	96.25%
NW64	3.52	3.92	3.24	7.71	3.25	3.82	47.01	93.89%
NW66	3.65	3.53	3.64	4.49	3.88	3.72	34.00	84.65%
NW76	4.86	4.43	6.52	5.50	5.03	4.93	18.00	91.28%
NW78	3.56	3.72	4.38	4.81	3.93	4.00	25.00	75.88%
NW86	2.78	2.38	8.17	2.50	2.46	7.75	12.02	88.54%
Total	4.02	4.37	6.09	4.90	4.18	4.47	628.76	85.87%

Table 2 Overflow Statistics

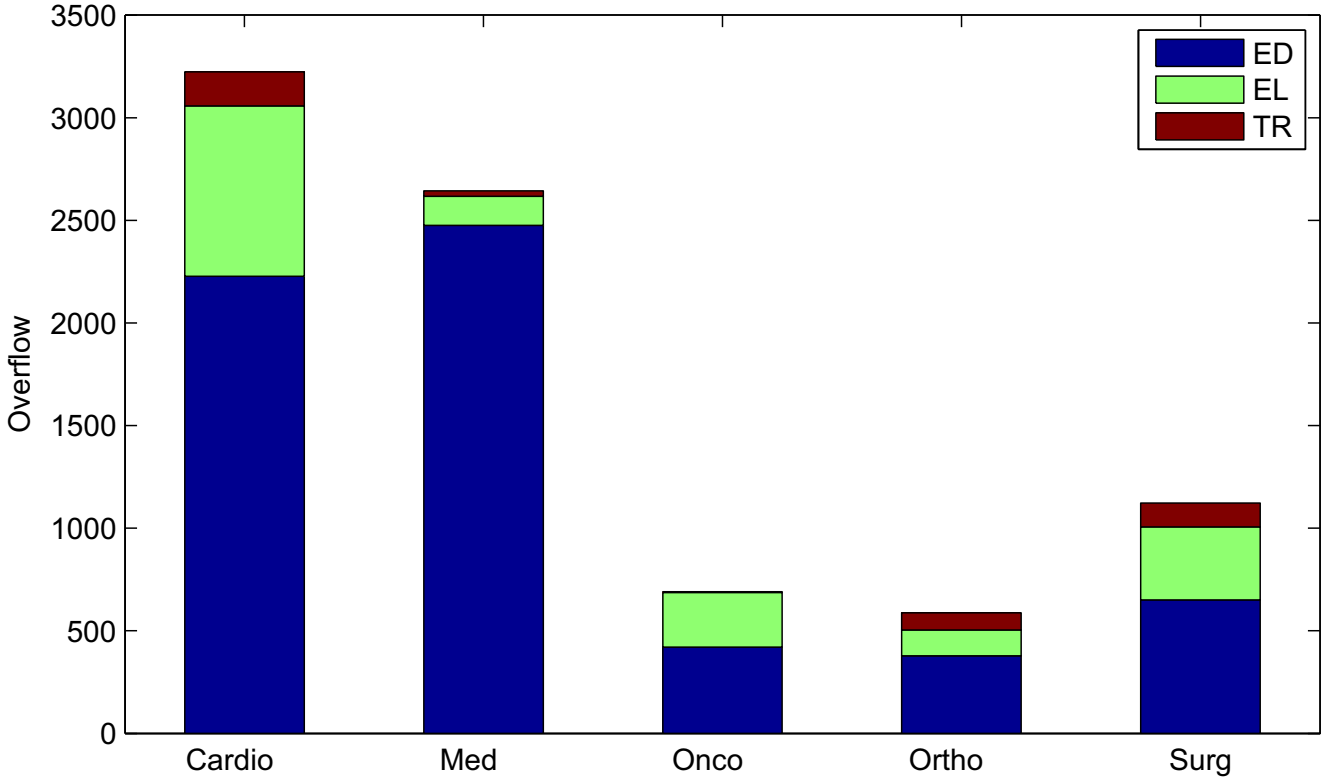
	Cardio(C)	Med(M)	Onco(On)	Ortho(Or)	Surg(S)	Total	Overflow	(%)
NW41(SC)	488	197	23	103	1654	2465	323	13.10%
NW42(M)	84	3347	46	13	48	3538	191	5.40%
NW43(S)	196	234	26	126	2859	3441	582	16.91%
NW44(SM)	168	1824	19	19	798	2828	206	7.28%
NW51(Or)	135	315	2	1996	381	2829	833	29.45%
NW52(Or)	106	246	23	1056	172	1603	547	34.12%
NW53(M)	49	2620	27	15	16	2727	107	3.92%
NW54(SOr)	256	747	22	1474	320	2819	1025	36.36%
NW55(M)	204	2248	69	8	70	2599	351	13.51%
NW56(C)	1045	44	11		4	1104	59	5.34%
NW57(M)	13	831	40	1		885	54	6.10%
NW57O(On)	25	165	1188	1	7	1386	198	14.29%
NW58(On)	21	126	1164	6	9	1326	162	12.22%
NW63(C)	3828	182	13	8	9	4040	212	5.25%
NW64(M)	644	3258	232	14	65	4213	955	22.67%
NW66(SM)	647	991	83	108	995	2824	838	29.67%
NW76(M)	288	381	54	164	329	1216	835	68.67%
NW78(SOr)	379	375	66	350	561	1731	820	47.37%
NW86(On)	9	13	464	2	13	501	37	7.39%
Total	8585	18,144	3572	5464	8310	44,075	8335	18.91%
Overflow	3224	2644	756	588	1123	8335		
(%)	37.55%	14.57%	21.16%	10.76%	13.51%	18.91%		

Figure 2 Overflow Breakdown by Wards [Color figure can be viewed at wileyonlinelibrary.com]



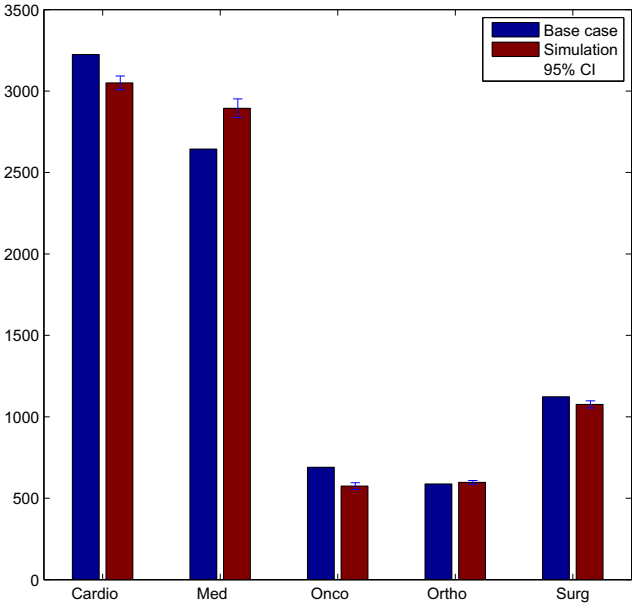
Another problem, not directly shown in the data but universally acknowledged and widely reported in the media, is the excessive delay experienced by EL patients, who have to wait up to weeks and months to get a bed. This excessive delay puts those patients in severe risks and increases the probability of deteriorating conditions, which, in turn, leads to more emergency admissions, creating a vicious cycle.

Figure 3 Overflow Breakdown by Clusters and by Patient Types [Color figure can be viewed at wileyonlinelibrary.com]



A high overflow rate along with what is merely a moderately high bed utilization suggests a serious issue of load imbalance. While this has not caused a serious delay for EM patients, that is only because the hospital management appears to be very serious in enforcing the 6-hour rule for emergency admissions, and thus quite willing to send EM patients to wrong-sited wards. The more direct negative impact is the reluctance of the management to admit more EL patients, lest this might result in even more EM overflows.

Figure 4 Overflow of Base Case and Simulation [Color figure can be viewed at wileyonlinelibrary.com]



2.2. Aggregation of Wards and Simulation Modeling

From the above discussions, and from Table 2 in particular, it is quite clear that to reduce overflow, the key is to reduce the imbalance among the bed capacities, not at the ward level but at the cluster level. Thus, we want to first aggregate the wards into “super wards," according to the clusters they are associated with as indicated in the first column of Table 2. This leads to the eight super wards in Table 3. Note that the first five super wards are mapped to the five clusters, whereas each of the last three covers two clusters: “SOr" (Surg and Ortho), “SC" (Surg and Card), and “SM” (Surg and Med).

With this aggregation, the LOS and BOR statistics are recalculated and summarized in Tables 4–6. As observed from Table 5, the lowest BOR is at SW4, 66.51%; and the highest BOR at SW1, 95.84%. (Note there is a small gap (about 2 beds) in total bed capacity due to rounding.)

We then use simulation to evaluate the performance of the aggregated model, to make sure that it matches the hospital's base case in all major performance measures. In the simulation model, we try to follow all the rules and work flows of the hospital's practice. In this regard, three details are worth noting. The first is the bed allocation priority (which is followed in the simulation), which is in this order: 1st priority: EM having already more than 6 hours; 2nd priority: TR; 3rd priority: other EM (waiting time less than 6 hours); 4th priority: EL.

The second rule from the hospital practice that we follow in the simulation is its overflow table: where to send a patient if no bed in the primary ward is available. Refer to Table 7. For instance, if the patient's primary ward is in SW1, and no bed is available there, then the first choice is to overflow the patient to an available bed in one of the wards in SW2, SW4, SW6, or SW8; if still no bed is available there, then look for a bed in SW3 or SW5.

The third practice has to do with how the EM waiting time is calculated. There are six steps involved in admitting an EM patient: (1) bed request from emergency department, (2) request acknowledged by bed management unit (BMU), (3) allocation by BMU, (4) confirmation from the emergency department, (5) discharge from emergency department, (6) admission into a general ward. The EM waiting time is defined as the duration between a bed request is initiated (step 1) and the bed confirmation is received (step 4). In the literature (e.g., Shi et al. 2015), the emergency department waiting time (or "boarding time") is defined as the duration between bed request time and the time to exit from the emergency department. This boarding time includes the discharge time, component 5) in the above list. As our focus here is on bed allocation, we do not include any delay involved in Steps 5 and 6. (These two times average to 45 minutes and 18 minutes from the data, and their inclusion will correspond to what is usually called "boarding time.") According to the above definition, the average EM waiting time is 1.71 hours. A major part of this time is spent on the search for an appropriate bed and related negotiations, which is typically longer in an overflow case. The average pre-allocation delay is 1.16 hours for right-sited allocations; hence, the delay due to bed shortage is 0.55 hours, the difference between the two delay times.

The simulated overall bed utilization is $85.93 \pm 0.22\%$, while the hospital BOR is 85.87%. Notably, the adjustment on LOS due to discharge windows in the simulation model has no statistical significance. For the overall overflow rate, simulation reports $18.70 \pm 0.23\%$ versus 18.91% for the base case. The breakdown among the five clusters is reported in Figure #4.

Table 3 Aggregate General Wards Into Super Wards

General wards	Super wards	Clusters
NW56,63	SW1	Card
NW42,53,55,57,64,76	SW2	Med
NW57O,58,86	SW3	Onco
NW51,52	SW4	Ortho
NW43	SW5	Surg
NW54,78	SW6	Surg & Ortho (SOr)
NW41	SW7	Surg & Card (SC)
NW44,66	SW8	Surg & Med (SM)

Table 4 Length of Stay of Base Case

Super wards	SW1	SW2	SW3	SW4	SW5	SW6	SW7	SW8	All
SW1	4.19	2.75	3.04	6.75	1.92				4.12
SW2	3.75	4.79	3.55	6.08	4.31				4.67
SW3	3.11	2.40	6.75	2.22	4.48				6.25
SW4	2.32	2.24	2.12	4.32	2.49				3.71
SW5	4.34	2.55	2.38	4.17	3.89				3.82
SW6	3.53	2.93	5.15			5.38			4.51
SW7		2.93	3.48	4.10			5.64		5.34
SW8	3.67		3.45	4.29				3.87	3.84
All	3.96	4.43	6.09	4.41	3.75	5.38	5.64	3.87	4.47

2.3. Technical Preliminaries and Motivation

To prepare for the analytical approach (in the next subsection) to the reallocation of beds to correct the supply–demand imbalance among the wards, we need some technical preliminaries, which we overview and motivate here.

The celebrated Erlang-C formula originated from the $M/M/c$ queueing model, where there are c parallel servers, the arrival process is Poisson with rate λ and the service times are i.i.d. exponential with rate μ. Refer to Appendix A.1 for more details.

Let $\rho := \lambda/\mu$ denote the traffic intensity, the rate of work (or "demand" rate) that comes to the system requiring service. Without loss of generality, think of each server works at unit rate, thus c is the *capacity* of the system—the maximum rate to deplete work from

Table 5 Bed Capacity and Occupancy Rate of Super Wards

Super wards	# Admissions	Ave bed capacity (and round-off)		BOR
SW1	5144	60.59	61	95.84%
SW2	15,178	209.56	210	92.61%
SW3	3213	59.94	60	91.75%
SW4	4432	67.66	68	66.51%
SW5	3441	41.24	41	87.39%
SW6	4550	73.5	74	76.51%
SW7	2465	43.58	44	82.68%
SW8	5652	72.68	73	81.81%
Total	44,075	628.76	631	85.87%

Table 6 Overflow Table of Base Case

Super wards	SW1	SW2	SW3	SW4	SW5	SW6	SW7	SW8	Total	Overflow	(%)
SW1	4873	226	24	8	13				5144	271	5.27%
SW2	1282	12,685	468	215	528				15,178	2493	16.43%
SW3	55	304	2816	9	29				3213	397	12.36%
SW4	241	561	25	3052	553				4432	1380	31.14%
SW5	196	234	26	126	2859				3441	582	16.91%
SW6	635	1122	88			2705			4550	1845	40.55%
SW7		197	23	103			2142		2465	323	13.10%
SW8	815		102	127				4608	5652	1044	18.47%
Total	8097	15,329	3572	3640	3982	2705	2142	4608	44,075	8335	18.91%
Overflow	3224	2644	756	588	1123				8335		
(%)	39.82%	17.25%	21.16%	16.15%	28.20%	0.00%	0.00%	0.00%	18.91%		

Table 7 Overflow Priority Table

From	Cardio SW1	Med SW2	Onco SW3	Ortho SW4	Surg SW5	SOr SW6	SC SW7	SM SW8
to (1st priority)	SW2,4,6,8	SW4	SW2,4,6,8	SW2,5,6,8	SW4	SW4,7,8	SW4,6,8	SW4,6,7
to (2nd priority)	SW3,5	SW1,3,5,6,7	SW1,5,7	SW1,3,7	SW1,2,3			

the system. Clearly, we require $\rho < c$ in order to have a stable system, that is, a system that has enough capacity to handle all the work that is pumped into it. (If not, backlogged work will cumulate and grow to infinity as time goes by.)

The ratio, ρ/c (which is <1 due to $\rho < c$) is the server utilization, denoted by u; it also measures the proportion of time the server is occupied (as opposed to idling). Thus, when the utilization is high, all c servers are occupied most of the time. Consequently, any new arrivals will have to wait in the queue before receiving service from the next available server. Let α denote the probability for this to happen, that is, how frequent an arrival will need to wait (as opposed to entering service right away).

A great insight from the Erlang-C formula is that such a system can be loaded to a very high utilization, but still very few arriving job need to wait, that is, while the utilization u can be very close 1, the delay probability α can be close to 0. For this to happen, the system needs to have a (relatively) large number of servers c. High utilization then means, ρ is also large, close to c (but still less than c). Suppose we express the difference

$$c - \rho = \beta\sqrt{\rho}, \tag{1}$$

with $\beta > 0$ being a constant. This indeed captures all the essential points mentioned above: c and ρ are both large, and their difference is order-of-magnitude smaller ($\sqrt{\rho}$); so, ρ is close to c, and hence the utilization is close 1.

Then, the Erlang-C formula gives a simple expression that relates the wait probability α to β via Equation (A.8) in the Appendix:

$$\alpha \approx \frac{1}{1 + u\beta\Phi(\beta)/\phi(\beta)} \approx \bar{\Phi}(\beta), \tag{2}$$

where Φ and ϕ denote the standard normal distribution and its density function, $\bar{\Phi} := 1 - \Phi$; and $u := \rho/c$. Furthermore, even when an arrival must wait, the average waiting time (in queue) is a fraction, $1/(\beta\sqrt{\rho})$ of the mean service time $(1/\mu)$; refer to Equation (A.12) in the Appendix A.1. Note, the second approximation in Equation (2) is a rather crude one (relative to the first approximation). It is based on Equation (A.7): $\phi(\beta)/\beta \approx \bar{\Phi}(\beta)$ when $\beta > 0$ is moderately large (see result #4 in the Appendix A.2).

Now, let us apply these results to Hospital-1. It has about 800 beds, of which our study focuses on a large portion, 629 beds, which served a total of 44,075 patients over a certain year. So, we have $c = 629$, and the daily patient arrival rate is $\lambda = 44{,}075/365 = 120.75$. The average length of stay (LOS) per patient is 4.47 days, which is $1/\mu$. Hence, the traffic intensity is $\rho = \lambda/\mu = 539.77$, and the utilization is $u = 85.81\%$. This also leads us to $\beta = (c - \rho)/\sqrt{\rho} = 3.84$, and hence $\alpha = 0.000076$ following Equation (2). That is, out of the 44,075 patients admitted over the year, only about 3 $(= 44{,}075 \times 0.000076)$ need to wait before a bed becomes available. This is obviously far from the observed performance at the hospital.

Of course, in any real hospital, and Hospital-1 in particular, the total number of beds must be organized into separate functional wards. In other words, they cannot be all lumped together as the c parallel servers in the $M/M/c$ model so as to achieve the maximal resource pooling effect. Yet, given a set of

functional wards, one can still optimize the allocation of the total number of beds among these wards according to their individual utilizations (patient arrival rate times LOS) so as to minimize the overflow and delay.

From the study on the hospital's data detailed above, the main cause of a high overflow rate is clearly a mismatch between demand and supply (bed capacity) among the wards, or more precisely, the super wards. So, our proposed solution is to do a reallocation of the total number of beds, across the eight super wards, indexed by $i = 1, \ldots, I(=8)$.

2.4. Reallocation and Performance Improvement

Suppose, after leaving out beds allocated to all elective patients, there is a total of C' bed remaining to be allocated to emergency patients, among the I super wards, termed "classes" below and indexed by $i = 1, \ldots, I$. Let $\rho_i := \lambda_i/\mu_i$ denote the traffic intensity of class i (emergency) patients. Let $C := C' - \sum_{i=1}^{I} \rho_i$, and assume $C > 0$. Following the square-root allocation rule detailed in the Appendix A.1, our problem is to allocate these C beds among the I classes, with $\beta_i\sqrt{\rho_i}$ for class i, and β_i's being the decision variables.

Given this bed allocation, the delay/wait probability for class i is $\alpha_i \approx \bar{\Phi}(\beta_i)$, following the approximation in Equation (2). Our objective is to minimize the largest wait probability among all classes:

$$\min_{(\beta_i)} \max_i \bar{\Phi}(\beta_i) \quad \text{s.t.} \sum_{i=1}^{I} \sqrt{\rho_i}\beta_i \le C; \quad \beta_i \ge 0, \quad \forall i. \tag{3}$$

Since the objective function is (strictly) decreasing, the (capacity) constraint must be binding. A moment's reflection also tells us that all β_i must be equal to achieve optimality, that is, the optimal solution is

$$\beta_i = C/\sum_{i=1}^{I} \sqrt{\rho_i}, \quad \forall i = 1, \ldots, I. \tag{4}$$

To reason formally, rewrite the above optimization problem as follows:

$$\min_{z,(\beta_i)} z \quad \text{s.t. } \bar{\Phi}(\beta_i) \le z, \quad \forall i; \quad \sum_{i=1}^{I} \sqrt{\rho_i}\beta_i \le C; \quad \beta_i \ge 0, \forall i. \tag{5}$$

Let η_i be the Lagrangian multiplier corresponding to the constraint $\bar{\Phi}(\beta_i) \le z$; and θ be the Lagrangian multiplier corresponding to the capacity constraint. Then, the optimality equations from taking (partial) derivatives on β_i and on z are:

$$\eta_i \phi(\beta_i) = \theta\sqrt{\rho_i}, \quad i = 1, \ldots, I; \quad \sum_{i=1}^{I} \eta_i = 1. \tag{6}$$

Summing up the first set of equations over i and taking into account the last one, we have

$$\phi(\beta_i) = \theta \sum_{i=1}^{I} \sqrt{\rho_i}, \quad i = 1, \ldots, I.$$

That is, the optimal β_i is equal for all i. This leads to:

$$\begin{aligned} \beta_i &= \frac{C}{\sum_{j=1}^{I} \sqrt{\rho_j}} := \beta, \quad \eta_i = \frac{\sqrt{\rho_i}}{\sum_{j=1}^{I} \sqrt{\rho_j}}, \quad i = 1, \ldots, I; \\ \theta &= \frac{\beta\phi(\beta)}{\sum_{j=1}^{I} \sqrt{\rho_j}}, \quad z = \bar{\Phi}(\beta). \end{aligned} \tag{7}$$

The above satisfy all the optimality equations, all the constraints, and the complementarity condition (each positive Lagrangian corresponds to a binding constraint). Since the optimal solution sets all $\beta_i = \beta$ as in Equation (7), we shall refer to this as the "equal β" allocation.

Alternatively, we can minimize the total overflow. Let $w > 0$ be a given upper-limit on the delay beyond which overflow must be triggered (e.g., $w = 6$ hours). Then, following Equation (A.12) in the Appendix A.1, with $c_i := \rho_i + \beta_i\sqrt{\rho_i}$, we have

$$\mathrm{P}(W_i(\beta_i) \ge w) = \alpha_i e^{-(c_i\mu_i - \lambda_i)w} = \alpha_i e^{-\beta_i\sqrt{\rho_i}\mu_i w}; \tag{8}$$

and the optimization problem is

$$\min_{(\beta_i)} \sum_{i=1}^{I} \lambda_i \mathrm{P}(W_i(\beta_i) \ge w) \quad \text{s.t.} \quad \sum_{i=1}^{I} \sqrt{\rho_i}\beta_i \le C; \quad \beta_i \ge 0, \quad \forall i. \tag{9}$$

Again, the objective function is decreasing, so the constraint must be binding. And, with θ as the Lagrangian multiplier, and $\alpha_i \approx \bar{\Phi}(\beta_i)$, the optimality equations are:

$$\lambda_i e^{-\beta_i\sqrt{\rho_i}\mu_i w}\left[\phi(\beta_i) + \bar{\Phi}(\beta_i)\sqrt{\rho_i}\mu_i w\right] = \theta\sqrt{\rho_i}, \quad i = 1, \ldots, I. \tag{10}$$

The resulting solution will be referred to below as the "MinOF" (minimal overflow) allocation.

Table 8 shows the bed allocations in the base case, and the reallocated bed capacities under the two new rules above. Notably, the two new allocation rules yield very similar solutions.

We then use simulation to estimate the performance of the re-allocation, and find the equal β allocation reduces the overall overflow rate drastically, from 18.91% to about 4.5%. The average EM waiting time also decreases from 0.55 hour to about 0.20 hour, as shown in Table 9 where the detailed ward-by-ward performance under the reallocation is also reported. The performance of the MinOF allocation is very similar. (Hence, in what remains of this section we shall focus on the equal β allocation only.)

The overall utilization of about 86% means there is a 14% bed capacity "reserved" for EM patients. So the natural question to ask is, can we reduce this reserved capacity and increase EL admissions accordingly? How much will this affect the overflow rates and waiting times for EM patients?

We use the simulation model for the experiments of increasing the EL admissions by 15%, 30%, and 50%. In the hospital's base case, EL patients constitute about 20% of the overall admissions; hence the proposed increase amounts to increasing the overall admissions by 3%, 6% and 10%, respectively. The utilization then increases from 86% to 88%, 92%, 96%, respectively.

As shown in Table 10 below, when we increase the EL admissions by 15%, and reallocate the bed capacity following the equal β allocation rule, the overall overflow rate is around 6%–7% and EM waiting is about 0.5 hour. Even when we increase the EL admissions by 50%, the overflow rate is around 9%–10%, and the average EM wait is around 4 hours, which is still significantly below the 6 hours requirement.

3. Diagnostic Resource Allocation at Hospital-2

Medical diagnostic facilities, such as computed tomography (CT) or magnetic resonance imaging (MRI), are a critical part of a comprehensive health care system, and play an important role in the proper diagnosis and timely treatment of diseases. Due to their high fixed and operational costs, these facilities often appear to be bottlenecks in many patient care processes, and thus negatively impact patient service, adding pressure on hospital managers to come up with more efficient and effective ways to manage these scarce resources.

Consider a typical facility. On any particular day, there are three types of patients requesting its service: outpatients, inpatients and emergency patients, indexed below by $i = 1, 2, 3$, respectively. The total time available during the day for each facility to serve patients (say, 12 or 13 hours) is divided into a fixed number of slots, denoted N. The usual practice is to further divide the N slots into three components, n_i for type i patients, with $n_1 + n_2 + n_3 = N$. The choice of n_i's is based on the daily averages of the three types, taking into account that emergency patients will have priority over the other two types.

Clearly, this practice is suboptimal. The optimal policy can be derived from a (stochastic) dynamic programming (DP) formulation of the problem, assuming serving each type i customer earns a revenue of r_i, a rejection costs c_i and any unused slot at

Table 8 Bed Allocation

Super wards	SW1	SW2	SW3	SW4	SW5	SW6	SW7	SW8	All
Base case									
No. admitted	8097	15,329	3572	3640	3982	2705	2142	4608	44,075
LOS	3.96	4.43	6.09	4.41	3.75	5.38	5.64	3.87	4.47
Bed capacity	61	210	60	68	41	74	44	73	631
Bed re-allocation									
Equal β	101	209	69	53	50	49	41	59	631
MinOF	102	209	70	53	50	48	40	59	631

Table 9 System Performance under Different Bed Re-Allocation Policies

	SW1	SW2	SW3	SW4	SW5	SW6	SW7	SW8	All
Equal β									
Bed re-allocation	101	209	69	53	50	49	41	59	631
BOR	86.74%	89.00%	85.49%	87.76%	83.57%	85.35%	83.62%	86.06%	86.81%
Overflow rate (in)	1.74%	2.34%	3.87%	12.29%	3.99%	10.03%	9.76%	5.47%	4.57%
Overflow rate (out)	4.03%	3.77%	3.65%	9.06%	4.90%	4.76%	4.53%	4.93%	4.57%
Ave EM wait	0.16	0.13	0.20	0.32	0.24	0.37	0.42	0.31	0.20
MinOF									
Bed re-allocation	102	209	70	53	50	48	40	59	631
BOR	86.01%	88.76%	84.39%	87.57%	82.90%	85.37%	84.42%	85.75%	86.45%
Overflow rate (in)	1.66%	2.11%	3.80%	11.88%	3.69%	8.93%	8.43%	5.17%	4.23%
Overflow rate (out)	3.45%	3.44%	2.97%	8.46%	4.44%	4.65%	5.38%	4.85%	4.23%
Ave EM wait	0.14	0.12	0.17	0.30	0.22	0.38	0.45	0.32	0.20

Table 10 EL Admissions Increase by 15%, 30% and 50%

	SW1	SW2	SW3	SW4	SW5	SW6	SW7	SW8	All
115% EL									
Equal β									
Bed re-allocation	102	206	71	53	50	49	41	59	631
BOR	90.35%	91.14%	90.80%	87.81%	87.45%	86.70%	86.51%	87.52%	89.42%
Overflow rate (in)	5.18%	4.14%	9.82%	8.55%	6.68%	10.29%	13.49%	6.95%	6.62%
Overflow rate (out)	7.16%	5.83%	6.74%	7.48%	7.70%	6.07%	6.25%	6.93%	6.62%
Ave EM wait	0.53	0.45	0.62	0.55	0.56	0.46	0.49	0.38	0.48
130% EL									
Equal β									
Bed re-allocation	103	202	76	53	50	49	40	58	631
BOR	92.19%	92.93%	92.25%	90.69%	90.48%	90.35%	90.61%	90.62%	91.78%
Overflow rate (in)	6.14%	4.81%	12.12%	10.77%	8.27%	12.32%	14.68%	8.98%	8.00%
Overflow rate (out)	8.27%	6.82%	6.50%	9.51%	9.92%	7.97%	9.13%	9.17%	8.00%
Ave EM wait	1.12	1.07	1.14	1.28	1.27	0.99	1.12	0.88	1.09
150% EL									
Equal β									
Bed re-allocation	104	200	78	53	50	49	40	57	631
BOR	95.70%	96.03%	96.21%	95.22%	95.03%	94.99%	94.84%	95.16%	95.62%
Overflow rate (in)	8.71%	6.53%	13.61%	13.09%	10.61%	14.96%	17.09%	10.83%	10.11%
Overflow rate (out)	11.04%	8.03%	9.22%	12.14%	13.05%	10.20%	10.67%	11.23%	10.11%
Ave EM wait	3.52	3.45	3.38	3.88	3.62	3.34	3.35	3.14	3.46

the end of working day is penalized at a cost rate of π, and the objective is to maximize the expected total net profit—revenues from serving the three types of patients minus cost penalties) over the day. Denote $\bar{r}_i := r_i + c_i$ $(i = 1, 2, 3)$, and assume

$$\bar{r}_1 \le \bar{r}_2 \le \bar{r}_3. \tag{11}$$

That is, serving an emergency patient is more valuable, in terms of the sum of revenue received and penalty avoided, than serving an inpatient, which, in turn, is more valuable than serving an outpatient.

We can show that the solution to the DP, that is, the optimal policy, has a threshold structure: accept a type 1 or type 2 patient if and only if the number of remaining (i.e., available) slots is above a certain threshold; and there's another threshold that applies to accepting inpatients (type 2) only. (Refer to the Appendix A.3.) While this structure is appealing, it is still too complex for practical implementation, as the thresholds not only depend on the current state (how many slots have been used up) but also on the time of the day.

Our goal is to design a simple and easily implementable policy that is close to the performance of the optimal policy. To this end, we pursue two alternative formulations that yield, respectively, an upper bound and a lower bound to the DP objective value. The upper- and lower-bound problems are *static* optimization problems; as such, their solutions are state- and time-independent, and hence easy to implement. Furthermore, solutions from both bounds are asymptotically optimal (more on this below). Based on these solutions, we construct a simple *nested policy*, detailed below, as the solution to the resource allocation problem.

3.1. A Nested Policy

The nested policy is directly built upon the lower-bound solution, which, in turn, is obtained from splitting the N appointment slots among the three types of patients, that is, accept each type as long as there are slots (allocated to that type) available, on a first-come-first-served basis. When all slots allocated to any type are exhausted, further arrivals from that type will be rejected. This yields a lower-bound to the DP since this static partition is clearly a feasible policy.

For $i = 1, 2, 3$, let D_i denote type-i demand, the total number of patients over the day. Consider normal demand distributions for all three types, with mean μ_i and variance σ_i^2 for $i = 1, 2, 3$. First, for emergency patients, use the upper-bound solution derived in the Appendix A.4, $\mu_3 + \sigma_3 z_3^\ell$, and reproduced below:

$$z_3^u = \frac{N - \mu_3}{\sigma_3}, \quad z_3^\ell = z_3^u \wedge \Phi^{-1}\left(\frac{\bar{r}_3 - \bar{r}_2}{\bar{r}_3 + \pi}\right), \tag{12}$$

where $\Phi(\cdot)$ denotes the distribution function of the standard normal $Z \sim N(0, 1)$.

Next, for the other two types, let $\mu_i + \sigma_i y_i$ denote the allocation to type-i requests, $i = 1, 2$, with y_i as decision variables. Then, similar to the upper-bound objective function in (A.32), the objective function of the lower-bound problem can be derived as follows

$$V^\ell := \sum_{i=1}^{3} r_i \mu_i - \pi\sigma_3 z_3^\ell - (\bar{r}_3 + \pi)\sigma_3 G(z_3^\ell) - \sum_{i=1}^{2} \bar{r}_i \sigma_i G(y_i), \tag{13}$$

where $G(x) := \mathbf{E}(Z - x)^+$ is the shortfall function; refer to Appendix A.2.

Hence, (y_1, y_2) should be the solution to:

$$\begin{aligned} &\min_{y_1,y_2} \sum_{i=1}^{2} \bar{r}_i \sigma_i G(y_i), \\ &\text{s.t.} \sum_{i=1}^{2} \sigma_i y_i \le N' := N - \mu_1 - \mu_2 - \mu_3 - \sigma_3 z_3^{\ell}. \end{aligned} \tag{14}$$

Since $G(y)$ is decreasing in y, the above constraint must be binding. Hence, the problem reduces to a single-variable convex minimization problem, which, upon a transformation of variables, can be expressed as follows:

$$\min_{x \ge -\mu_1} \bar{r}_1 \sigma_1 G\left(\frac{x}{\sigma_1}\right) + \bar{r}_2 \sigma_2 G\left(\frac{N' - x}{\sigma_2}\right). \tag{15}$$

The above problem can be solved via a line search. Furthermore, it is readily verified that the solution must satisfy

$$\frac{x}{\sigma_1} \le \frac{N' - x}{\sigma_2}, \text{ or } x \le \frac{N' \sigma_1}{\sigma_1 + \sigma_2}, \tag{16}$$

since $\bar{r}_1 \le \bar{r}_2$.

Comparing the lower bound V^{ℓ} in Equation (13) with the upper bound V^u in Equation (A.36), we observe the leading term in both is equal to $\sum_{i=1}^{3} r_i \mu_i$, which is of order T; whereas all other terms involve a factor of σ_i, and hence will be of order $\sqrt{T}$, With more work we can further bound the gap between the upper- and lower-bound, $V^u - V^{\ell}$, as follows,

$$\delta \le V^u - V^{\ell} \le \delta + (\bar{r}_2 - \bar{r}_1)\sigma_2 G(\gamma) \quad \text{where } \delta \sim O(\sqrt{T}), \tag{17}$$

and γ is a constant independent of T.

The above clearly implies, with V^* denoting the optimal DP value,

$$\lim_{T\to\infty} \left| \frac{V^* - V^u}{V^*} \right| = 0, \quad \lim_{T\to\infty} \left| \frac{V^* - V^{\ell}}{V^*} \right| = 0.$$

Hence, both the upper- and lower-bound solutions are asymptotically optimal.

The lower-bound solution V^{ℓ} obtained above is a fully partitioned policy: it partitions the N slots into $n_3 := \mu_3 + \sigma_3 z_3^{\ell}$ slots for emergency patients, where z_3^{ℓ} follows Equation (12), $n_1 := \mu_1 + x^*$ slots for outpatients, where x^* is the solution to the minimization problem in (15), and the remaining $n_2 = N - n_1 - n_3$ slots for inpatients. Let $\bar{N} := n_1 + n_2 = N - n_3$. It is natural to modify the partition among outpatients and inpatients by allowing the latter to have access to all $\bar{N}$ slots, since $\bar{r}_2 \ge \bar{r}_1$. Specifically, outpatients will be accepted if and only if the number of accepted outpatients is below n_1; whereas inpatients will be accepted if and only if the *sum* of accepted outpatients and inpatients is below $\bar{n}$. Call this the *nested* (partition) policy, and denote its value function as V^{n}. It is readily shown that $V^u \ge V^{\mathrm{n}} \ge V^{\ell}$. Thus, the nested policy is also asymptotically optimal.

3.2. Hospital-2 Data and Practice

Here we use process flows and data from Hospital-2 in numerical studies along with simulation, to conduct performance comparisons between the nested policy developed above and the baseline practice at the hospital. The dataset covers a 1-year period with 90,790 patient records of CT regular scans, retrieved from the hospital's information system, following standard compliance with the hospital's patient privacy guidelines. We choose to focus on the regular CT scans (as opposed to enhanced scans), as these involve all three types of patients, and have a high daily volume, with short and less variable service times.

For regular CT scans at the hospital, both outpatients and inpatients have to make appointments in advance while emergency patients randomly walk in and receive service with priority. The normal CT working hours are 08:00–21:00, a total of 13 hours per day. The average service time per patient is 2.4 minutes, with very little variation among the three types. This translates into $N = 325$ available slots as the daily capacity for each facility.

Requests for service, or "demand rates," from all three types of patients, are reported in Table 11, with an hourly breakdown. Statistical tests applied to data have confirmed Poisson distribution as the best fit. Demand over the weekdays usually exceeds the daily capacity, inevitably causing a substantial number of service requests from outpatients and inpatients to be deferred for 1 or even 2 days. For the two weekend days (Saturday and Sunday), the hospital's practice is to close (new) appointment requests from outpatients and inpatients, and only serve emergency patients and clear out the outpatients and inpatients deferred from the weekdays.

The hospital's practice is to accept up to 120 outpatients and up to 50 inpatients each day and to reserve the remaining slots for emergency patients. This will be referred to the hospital's baseline policy, and summarized in Table 12, with a period-by-period breakdown of the two thresholds, 120 and 50. In reality, however, these reference numbers are often adjusted *ad hoc*, resulting in the fluctuations and demand–supply mismatch.

3.3. Sensitivity Analysis

There are cost parameters involved in the analytical models, including both the DP optimal solution and the nested policy. They can be chosen or adjusted to

reflect the relative importance among the on-time service of the three types of patients, and also the relative value of the CT resource.

Therefore, we first conduct a sensitivity analysis to compare the performance of the nested policy and the hospital's baseline policy under different combinations of these cost parameters using the average demand rates in Table 11. The results are reported in Table 13, where α^n and α^b are the two policies' relative gap in percentages below the optimal DP solution (which achieves the maximal expected net profit). For all cases, $c_1 = 500$ is fixed as a reference point.

Several observations can be drawn:

- The performance of the nested policy is significantly better, under all combinations of parameters: its average gap from the optimal is 2.22%, as opposed to the baseline's 17.81%. Refer to part (b) of the table.
- The performance of the nested policy is relatively insensitive to the cost parameters. In comparison, the performance of the baseline policy significantly deteriorates when either c_2 or π increases. Its performance generally improves when c_3 increases.
- Thus, the baseline policy seems to single-mindedly focus on emergency patients, making sure there are enough slots reserved for these patients.
- The (best) partition thresholds under the nested policy are detailed in part (c) of the table, under various parametric combinations. They appear to be quite insensitive as well. The small adjustments are also easy to understand. For instance, the slots reserved for emergency patients, $n_3 = N - \bar{N}$ increases in c_3, and decreases in π (reserve less if any wasted slot is priced higher).

Table 12 The Hospital's Baseline Policy

Time period	Time length (mins)	Patient type	Number of scheduled patients
08:00–08:30	30	Outpatient	25
08:30–11:00	150	Inpatient	20
11:00–11:30	30	Inpatient	10
11:30–12:00	30	Outpatient	15
12:00–13:00	60	N.A.	0
13:00–14:30	90	Outpatient	10
14:30–17:00	150	Inpatient	20
17:00–18:00	60	Outpatient	10
18:00–19:00	60	N.A.	0
19:00–21:00	120	Outpatient	60
Total no. of appointments		Inpatient	50
Total no. of appointments		Outpatient	120

3.4. Performance Improvement Using the Nested Policy

Next, we do a detailed comparison between the nested policy and the hospital's baseline policy, when applied to the hospital data over any week, the five-day cycle. From the above sensitivity analysis, we have observed that the nested policy of (120, 194, 131) connects well with the baseline policy: they have the same threshold (120) for outpatients, the largest source of demand. This nested policy, in turn, corresponds to the following cost parameters (per Table 13),

$$r_1 = r_2 = r_3 = 800,\ c_1 = 500,\ c_2 = 750,\ c_3 = 2000,\ \pi = 800.$$

Using these parameters, we will compute the best nested partition for each weekday using the demand data from Table 11.

We then run simulation to compare the performance of the two policies over a typical one-week cycle including all five week days. (The weekends

Table 11 Demand Rates from Three Types of Patients

		Time Interval													
DOW	Patient type	08–09	09–10	10–11	11–12	12–13	13–14	14–15	15–16	16–17	17–18	18–19	19–20	20–21	Sum
Mon	Outp.	20.2	33.6	34.7	30.7	13.1	15.6	24.5	23.6	11.5	2.2	0	0	0	210
	Inp.	14.8	12.6	14.7	1.9	2.8	5.8	17.1	6.3	3.5	0	0	0	0	80
	Emg.	4.5	9.5	15.3	9.7	12.3	12.7	11.9	12.2	12.6	9.8	8.4	11.4	11.3	142
Tue	Outp.	18.3	32.6	32.0	27.3	11.4	15.8	23.4	21.0	9.7	2.0	0	0	0	194
	Inp.	23.3	17.0	14.2	2.4	3.4	7.3	12.9	4.8	1.6	0	0	0	0	87
	Emg.	5.3	11.4	14.9	9.0	10.3	12.3	10.1	9.0	13.0	9.3	8.7	10.6	10.9	135
Wed	Outp.	14.1	25.9	27.8	22.8	8.5	14.3	21.8	18.4	10.0	2.1	0	0	0	166
	Inp.	22.0	12.9	11.0	3.8	4.3	5.0	14.6	4.4	1.5	0	0	0	0	80
	Emg.	5.4	9.9	13.9	9.5	10.4	10.8	9.9	10.4	11.9	7.3	8.1	12.2	11.8	132
Thu	Outp.	11.7	26.2	26.3	19.5	7.2	10.3	15.9	13.7	8.9	2.0	0	0	0	142
	Inp.	19.4	14.1	10.0	3.4	8.3	7.4	14.7	4.0	1.7	0	0	0	0	83
	Emg.	5.8	9.8	14.7	8.4	10.0	11.0	10.1	10.3	12.0	8.4	8.7	12.4	11.7	133
Fri	Outp.	14.3	23.8	23.7	18.7	7.0	9.8	13.6	12.0	5.6	1.4	0	0	0	130
	Inp.	22.9	12.5	11.3	1.4	10.5	7.9	14.2	5.8	2.6	0	0	0	0	89
	Emg.	5.6	11.4	16.2	9.1	9.9	10.6	9.9	10.1	11.1	8.3	9.3	12.5	9.9	134
Average	Outp.	15.7	28.4	28.9	23.8	9.4	13.2	19.8	17.7	9.1	1.9	0	0	0	168
demand	Inp.	20.5	13.8	12.2	2.6	5.9	6.7	14.7	5.1	2.2	0	0	0	0	84
rate	Emg.	5.3	10.4	15.0	9.1	10.6	11.5	10.4	10.4	12.1	8.6	8.6	11.8	11.1	135

Table 13 Sensitivity Analysis for Performance Comparisons between the Nested Policy and the Hospital's Baseline Policy

(a) Relative gap (%) below the optimal DP solution

		$\pi = 400$		$\pi = 800$		$\pi = 1200$	
c_2	c_3	α^n	α^b	α^n	α^b	α^n	α^b
750	2000	1.59	13.21	1.63	16.39	1.62	19.71
	2500	1.62	12.32	1.60	15.37	1.60	18.58
	3000	1.61	11.60	1.62	14.52	1.66	17.64
1000	2000	2.93	17.12	2.87	20.32	2.86	23.67
	2500	2.84	16.27	2.79	19.35	2.78	22.60
	3000	2.83	15.58	2.79	18.55	2.79	21.71

(b) Overall gap (%)

	α^n	α^b
Minimum	1.59	11.60
Maximum	2.93	23.67
Average	2.22	17.81

(c) The best partition thresholds under the nested policy

		$\pi = 400$			$\pi = 800$			$\pi = 1200$		
c_2	c_3	n_1	$\bar{N}$	n_3	n_1	$\bar{N}$	n_3	n_1	$\bar{N}$	n_3
750	2000	118	192	133	120	194	131	121	195	130
	2500	116	190	135	117	191	134	118	192	133
	3000	114	188	137	115	189	136	117	191	134
1000	2000	117	195	130	118	196	129	119	197	128
	2500	114	192	133	115	193	132	116	194	131
	3000	112	190	135	113	191	134	114	192	133

are left out since no service requests are accepted from inpatients and outpatients; hence, no decisions to be made.) The simulation uses random patient arrivals and random service times according to the hospital data. It also captures two idiosyncratic and inevitable phenomena: (i) the occasional forced idling of the facility, and (ii) the need to run overtime in order to clear those inpatients and outpatients who are already scheduled but delayed by emergency patients. The first one is explicitly accounted for in the analytical model by the penalty π for each wasted slot; the second one is only tangentially reflected in the c_3 cost.

The simulation for the nested policy is detailed as follows.

1. Get the "static" partition (n_1, n_2, n_3) from the LB solution (14) with $n_i = \mu_i + \sigma_i y_i^* (i = 1, 2)$ and $n_3 = \mu_3 + \sigma_3 z_3^*$ for each of the 5 days.
2. Implement the above partition (n_1, n_2, n_3) as a nested policy $(n_1, \bar{N}, n_3)$, that is, reserve n_3 slots for type 3, and use the remaining $\bar{N} = N - n_3$ slots for types 1 and 2, with a further limit n_1 applied to type 1.
3. Run simulation to get the numbers (served, deferred, utilization and overtime) for Monday.
4. Starting from Tuesday, first assign the deferred inpatients to the $\bar{N}$ slots (for Tuesday, obtained from Step 1), and deduct these slots from n_1 (also for Tuesday, obtained from Step 1). Use the reduced n_1 to control the number of outpatients (assigning those deferred from Monday first).
5. Repeat the above for Wednesday, Thursday and Friday.

We run the simulation until the relative error is within 0.5%. The performance comparison with respect to served and deferred is summarized in Table 14(a), where the column "served" counts the number of patients served on the same day (emergency patients are all served on the same day); and "deferred" is measured in patient-days, as some patients may be deferred by more than one day. (Hence, one patient deferred for 2 days counts as 2 patient-days.) The "total" in the "served" column counts the total number of three types of patients served; the "total" in the "deferred" column counts the total patient-days deferred for inpatients and outpatients. The performance comparison with respect to utilization and overtime is summarized in Table 14(b). From the table, the advantages of the nested policy include the following:

- Its on-time service (i.e., on the same day) is a total of 975 inpatients and outpatients over a week as opposed to the baseline's 850, a 14% improvement.
- Its deferred patient-days (over a week) is 1068, as opposed to the baseline's 1431, a 33% improvement. In particular, the deferred patient-day for inpatients is 38, in sharp contrast to the baseline's 493, with only a slight increase of 92 deferred patient-days for outpatients.
- It improves the utilization significantly (by about 10%), while also reducing the overtime (by about 11%).

4. Concluding Remarks

Using two resource allocation problems from two large hospitals as cases, along with real data and process flows, we have demonstrated in this study how analytical modeling and optimization can significantly improve hospital performance, in terms of both patient service and resource utilization. Furthermore, this improvement is achieved without any increase in the overall level of resource capacity, entirely through revamping existing operations. In the case of Hospital-1, the key is to use the square-root allocation rule to optimize the number of beds allocated to the various wards, so as to reduce the overflow rates and waiting times for emergency patients, and to admit more elective patients. In the case of Hospital-2, the key is a nested partition of the diagnostic resource capacity

Table 14 Performance Comparisons of the Nested Policy and the Hospital's Baseline Policy through Simulation

(a) Performance comparison w.r.t. served (no. of patients) and deferred (patient-days)

	Nested Policy						Baseline Policy					
	Served			Deferred			Served			Deferred		
DOW	Outpatient	Inpatient	Total	Outpatient	Inpatient	Total	Outpatient	Inpatient	Total	Outpatient	Inpatient	Total
Mon	117	71	330	93	9	102	120	50	312	90	30	120
Tue	108	86	329	178	9	187	120	50	305	164	67	231
Wed	117	81	330	227	8	235	120	50	302	210	97	307
Thu	114	85	332	255	6	261	120	50	303	232	130	362
Fri	108	89	331	277	6	283	120	50	304	242	169	411
Sum	564	411	1652	1030	38	1068	600	250	1526	938	493	1431

(b) Performance comparison w.r.t. utilization and overtime

	Nested Policy		Baseline Policy	
DOW	Utilization (%)	Overtime (mins)	Utilization (%)	Overtime (mins)
Mon	91.8	75	85.8	79
Tue	93.5	61	84.1	76
Wed	91.3	78	82.3	82
Thu	92.5	75	82.8	82
Fri	93.6	63	83.5	78
Average	92.5	70	83.7	79

among emergency patients, inpatients, and outpatients, along with optimized threshold values derived from the upper- and lower-bound solutions to a dynamic programming formulation. In both cases, the solutions are either in the closed form or easy to compute; as such, they facilitate implementation, and can be readily updated whenever there is a shift in demand or resource availability.

In both cases, the challenges and difficulties involved in applying analytics and related tools are also clearly demonstrated. In particular, it appears that no tool can really be a "plug and play"; some level of modification and adaptation, applied to either the model or the solution or both, always seems needed. In this regard, analytics is almost as much of an art as it is a science.

Acknowledgments

This study is supported in part by the National Natural Science Foundation of China, Grants No. 71571176, 71672160, 71201140, 71432004, 71532007, 71131006, 71172197; by the National University of Singapore Academic Research Fund, Grant No. R-314-000-082-112; the National University of Singapore Medicine-Business Seed Grant, Grant No. C-311-004-003-001; the Ministry of Health Singapore, Grant No. HSRG/0002/2010; and by Hong Kong Research Grants Council, Grant No. T32-102/14N. We are deeply grateful to the administration and staff at the two hospitals where our studies were conducted for their support and assistance.

Appendix A

A.1. Erlang-C Formula and the Square-Root Allocation Rule

Consider the $M/M/c$ queueing model, with arrivals following a Poisson process and service times following an i.i.d. exponential distribution, and there are c parallel servers. Let λ and μ denote, respectively, the arrival and the service rates; let $\rho := \lambda/\mu$. Assume $c > \rho$ to ensure stability of the system. Let π_i, $i = 0, 1, \ldots$, denote the steady-state probability that there are i jobs in the system. We know

$$\pi_i = \frac{\rho^i}{i!}\pi_0,\ i = 0,1,,c;\ \pi_{c+k} = \left(\frac{\rho}{c}\right)^k \pi_c,\ k = 0,1,2,\ldots; \tag{A.1}$$

and, in particular, $\pi_0 = 1/S$, where S is the normalizing constant:

$$S = \sum_{i=0}^{c-1} \frac{\rho^i}{i!} + R\frac{\rho^c}{c!},\ \text{and}\ R := \frac{1}{1-\rho/c} = \frac{c}{c-\rho}. \tag{A.2}$$

Other probabilities of interest:

$$\pi_0 = \frac{1}{S},\quad \pi_c = \frac{\rho^c}{c!}\pi_0 = \frac{\rho^c}{Sc!}; \tag{A.3}$$

and the probability that an arriving job will have to wait in queue,

$$\alpha := \sum_{k=0}^{\infty} \pi_{c+k} = R\pi_c = \frac{R}{S}\frac{\rho^c}{c!}, \tag{A.4}$$

which is known as the Erlang-C formula.

Approximations to the above formulas, the last one in particular, can be worked out via approximating the Poisson distribution by normal. Let N follow a Poisson distribution with mean (and variance) ρ, and we approximate it by a normal variate X with the same mean and variance. Then,

$$\begin{aligned}\sum_{i=0}^{c-1}\frac{\rho^i}{i!} &= e^{\rho}[1-\mathrm{P}(N\geq c)] \approx e^{\rho}[1-\mathrm{P}(X\geq c)]\\ &= e^{\rho}\Phi\left(\frac{c-\rho}{\sqrt{\rho}}\right),\end{aligned} \tag{A.5}$$

where $\Phi(x)$ denotes the distribution function of the standard normal. Similarly, approximating $\mathrm{P}(N = c)$ by a normal density, and with $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ denoting the density function of the standard normal, we have

$$\frac{\rho^c}{c!} = e^{\rho}\mathrm{P}(N=c) \approx e^{\rho}\frac{1}{\sqrt{\rho}}\phi\left(\frac{c-\rho}{\sqrt{\rho}}\right). \tag{A.6}$$

Write

$$\beta := \frac{c-\rho}{\sqrt{\rho}}. \tag{A.7}$$

Then, we have, taking into account $\frac{R}{\sqrt{\rho}} = \frac{c}{\rho\beta}$,

$$\alpha = \frac{R\frac{\rho^c}{c!}}{\sum_{i=0}^{c-1}\frac{\rho^i}{i!}+R\frac{\rho^c}{c!}} \approx \frac{1}{1+(\rho/c)\beta\Phi(\beta)/\phi(\beta)}. \tag{A.8}$$

It is known that $\bar{\Phi}(x) := 1-\Phi(x) \approx (\leq)\phi(x)/x$ when $x > 0$. This leads to, with $u := \rho/c$ being the server utilization,

$$\frac{\bar{\Phi}(\beta)}{\bar{\Phi}(\beta)+u\Phi(\beta)} \leq \alpha \leq \frac{\phi(\beta)}{\phi(\beta)+u[\beta-\phi(\beta)]}. \tag{A.9}$$

For the normal approximation outlined above to work well, ρ needs to be large; and hence, so must be c (recall $c > \rho$), such that β in Equation (A.7) is a positive constant. (Numerically, a reasonably large ρ, say, in the 10–15 range, will already make the approximation work quite well.) This can be made more precise as follows:

$$\begin{aligned}\rho\sim O(n),\ c\sim O(n),\ u := \frac{\rho}{c}<1;\ \sqrt{\rho}(1-u)\to\beta\\ \in(0,\infty).\end{aligned} \tag{A.10}$$

The last limit says, the server utilization approaches 1 in the order of $1-\frac{\beta}{\sqrt{\rho}}$ (as $\rho\to\infty$). This is the so-called "heavy traffic condition."

The waiting time in *queue* (i.e., excluding service time), denoted W, of any arriving job follows an Erlang distribution of $k+1$ iid exponential phases, each with mean $1/(c\mu)$, denoted $E(k+1, c\mu)$, if there are $c+k$ jobs in the system, for $k = 0, 1, \ldots$; o.w., the waiting time in queue is zero. That is,

$$\begin{aligned}W = 0 \text{ w.p. } 1-\alpha;\ W = E(k+1,c\mu) \text{ w.p. } \pi_{c+k},\ k\\ = 0,1,\ldots.\end{aligned} \tag{A.11}$$

Therefore, the density function of W is, for $x > 0$:

$$\begin{aligned}\mathrm{P}(W=x) &= \mathrm{P}[E(k+1,c\mu)=x]\\ &= \sum_{k=0}^{\infty}\frac{c\mu(c\mu x)^k}{k!}e^{-c\mu x}\left(\frac{\rho}{c}\right)^k\pi_c = c\mu\pi_c e^{-(c\mu-\lambda)x}.\end{aligned}$$

Hence,

$$\mathrm{P}(W\geq x) = \frac{c\mu}{c\mu-\lambda}\pi_c e^{-(c\mu-\lambda)x}.$$

Recognizing

$$\frac{c\mu}{c\mu-\lambda}\pi_c = R\pi_c = \alpha,$$

we have

$$\mathrm{P}(W\geq x) = \alpha e^{-(c\mu-\lambda)x},\ x>0. \tag{A.12}$$

That is, given a job has to wait in queue (which happens w.p. α), the waiting time is exponentially distributed, with rate $c\mu-\lambda = (c-\rho)\mu = \beta\sqrt{\rho}\mu$.

In summary, $1-\alpha$ is the key service-level measure: the probability of zero queuing delay ($W = 0$); and α follows the approximation in Equation (A.8). To achieve this service level, the required capacity c follows a square-root rule, via Equation (A.7), with β being the "safety factor": $c = \rho+\beta\sqrt{\rho}$. With this rule, the probability for any job to wait in queue is α, in which case the waiting time is exponentially distributed with a mean $1/(\beta\sqrt{\rho}\mu)$, that is, only a fraction $1/(\beta\sqrt{\rho})$ of the mean service time $1/\mu$.

A.2. Useful Results for the Shortfall Function of Standard Normal

Here we collect some useful results that relate to the standard normal variable, Z, and the shortfall function involving Z.

The density and distribution functions of Z will be denoted by $\phi(x)$ and $\Phi(x)$, respectively. Let $\bar{\Phi}(x) := 1-\Phi(x)$. Define the "shortfall function" as

$$\begin{aligned}G(x) &:= \mathrm{E}(Z-x)^+ = \int_x^{+\infty}(z-x)\phi(z)dx\\ &= \phi(x)-x\bar{\Phi}(x)\end{aligned} \tag{A.13}$$

where the last equality follows directly from $\phi'(z) = -z\phi(z)$.

The following is a collection of useful properties of $G(x)$.

1. $G(x)$ is decreasing and convex in x, since $(Z - x)^+$ is decreasing and convex in x. Also, $G'(x) = -\bar{\Phi}(x)$.
2. The part of $G(x)$ that has significant curvature is limited to $[-1, 1]$. For $x > 1$, we have $G(x) \approx 0$; and for $x < -1$, we have $G(x) \approx -x$.
3. Direct calculation implies, for any $a \le b$,

$$2\int_a^b G(x)dx = \Phi(b) - \Phi(a) + bG(b) - aG(a). \tag{A.14}$$

4. From $(1 - \frac{1}{x^2})\frac{\phi(x)}{x} \le \bar{\Phi}(x) \le \frac{\phi(x)}{x}$ for $x > 0$, we have

$$0 \le G(x) = \phi(x) - x\bar{\Phi}(x) \le \frac{\phi(x)}{x^2}, \ \forall x > 0. \tag{A.15}$$

5. Applying L'Hôspital twice, we have

$$\frac{x^2 G(x)}{\phi(x)} \to 1 \text{ as } x \to +\infty. \tag{A.16}$$

That is, $G(x) \sim \frac{\phi(x)}{x^2}$ as $x \to +\infty$, in contrast with $\bar{\Phi}(x) \sim \frac{\phi(x)}{x}$.

A.3. Optimal Threshold Policy for the Diagnosis Resource Allocation Problem

Here we formulate the resource allocation problem of a medical diagnostic facility among the three types of patients (indexed by $i = 1, 2, 3$ as in section 3.1) as a finite-horizon stochastic dynamic program (DP). The working hours of the facility during a day are divided into N equal "slots," each being the time needed to complete one patient's diagnosis. The allocation problem concerns a single working day of the facility: how to split the N slots among the three types of patients, with the provision that outpatients and inpatients ($i = 1, 2$) need to make a request for an appointment (which may not be accepted), whereas emergency patients ($i = 3$) can just walk in and will always be accepted.

As in section 3.1, for $i = 1, 2, 3$, let D_i denote type-i demand, the total number of patients over the day. Suppose requests (for appointments) from type 1 and type 2 patients arrive during the planning horizon following two independent Poisson processes. The planning horizon is divided into T equal intervals, referred to as "periods" below and indexed backward by $t = T, \ldots, 1$. The choice of T should be large enough, such that we can model a discrete-time version of the two Poisson arrival processes. Specifically, over each period t, the probability of having a type-i request is $\lambda_t^i \in (0, 1)$ (for either $i = 1$ or $i = 2$, but not both); and $1 - \lambda_t^1 - \lambda_t^2$ is the probability that there is no request (from either type) in period t. Upon the arrival of each request, the decision is to either accept or reject the request. An accepted request will be given a slot to perform the diagnosis. Suppose for $i = 1, 2$, each accepted type-i request earns a revenue of r_i, and each rejected request incurs a penalty c_i. The slots that remain at the end of the planning horizon ($t = 0$) are those reserved for emergency patients (type 3). Let r_3 be the revenue from serving each type-3 patient, and c_3 be the penalty cost for rejecting a type-3 patient. Let π be the penalty cost for any un-used slot at the end of the working day.

Let $V_t(n)$ denote the value-to-go function, starting from period t onward and with n slots *remaining* (out of a total of N). The DP recursion is given as follows, for $t = T, \ldots, 1$; and $n = 1, \ldots, N$, we have

$$V_t(n) = \sum_{i=1}^{2} \lambda_t^i \max\{V_{t-1}(n-1) + r_i, V_{t-1}(n) - c_i\} + \left(1 - \sum_{i=1}^{2} \lambda_t^i\right) V_{t-1}(n), \tag{A.17}$$

and at the boundary, $t = 0$,

$$V_0(n) = r_3 \mathbf{E}(D_3 \wedge n) - c_3 \mathbf{E}(D_3 - n)^+ - \pi \mathbf{E}(n - D_3)^+. \tag{A.18}$$

The total value generated over the planning horizon is $V_T(N) = V^*$, which we want to maximize.

To simplify notation, denote

$$\Delta V_t(n) := V_t(n) - V_t(n-1), \quad t = T, \ldots, 0.$$

Define

$$n_t^i := \min\{n | \Delta V_{t-1}(n) \le \bar{r}_i\}, \ i = 1, 2; \ t = T, \ldots, 1, \tag{A.19}$$

which is the minimum n such that the marginal value of an appointment is no greater than $\bar{r}_i$, and in particular, define

$$\bar{n}_i := n_1^i = F_3^{-1}\left(\frac{\bar{r}_3 - \bar{r}_i}{\bar{r}_3 + \pi}\right), \quad i = 1, 2. \tag{A.20}$$

where $F_3(\cdot)$ denotes the distribution function of D_3, the number of emergency requests (during the day). Note that when D_3 follows a normal distribution with mean μ_3 and variance σ_3^2. Denote $a \wedge b := \min(a, b)$, then,

$$N \wedge \bar{n}_2 \equiv n_3^* = \mu_3 + \sigma_3 z_3^* \text{ where } z_3^* = \Phi^{-1}\left(\frac{\bar{r}_3 - \bar{r}_2}{\bar{r}_3 + \pi}\right) \wedge \left(\frac{N - \mu_3}{\sigma_3}\right), \quad \text{(A.21)}$$

with $\Phi(\cdot)$ being the distribution function of the standard normal random variable Z_3. Substituting n_3^*, the number of slots reserved for emergency patients, into V_0, along with $D_3 = \mu_3 + \sigma_3 Z_3$, we have

$$V_0(n_3^*) = r_3\mu_3 - \pi\sigma_3 z_3^* - (\bar{r}_3 + \pi)\sigma_3 G(z_3^*), \quad \text{(A.22)}$$

where $G(x) := \mathbf{E}(Z - x)^+$ is the well-known shortfall (or, loss) function. Refer to section A.2 for properties of the shortfall function used in the analysis below.

From the DP model in Equations (A.17)–(A.18), we can derive the optimal policy in Proposition 1. Throughout, we use the words "increasing" and "decreasing" in the non-strict sense, meaning non-decreasing and non-increasing, respectively.

PROPOSITION 1. *The optimal policy is characterized as follows. If $\bar{n}_2 \geq N$, no appointment will be accepted; otherwise, in each period $t = T, \ldots, 1$ and for $i = 1, 2$,*

(a) $n_t^1 \geq \bar{n}_1 \geq \bar{n}_2 = n_t^2$, and n_t^i is increasing in t;
(b) if $n > n_t^1$, accept both types of requests;
(c) if $\bar{n}_2 < n \leq n_t^1$, accept inpatient (type-2) requests only;
(d) if $n \leq \bar{n}_2$, reject both types of requests.

In particular, $n_3^* = N \wedge \bar{n}_2$ is the number of slots reserved for emergency patients, and $\bar{N} = (N - \bar{n}_2)^+$ is the total number of appointment slots can be accepted.

Proposition 1 characterizes the optimal reservation policies as the switching-curve or threshold type, which are illustrated in Figure 5.

The key to prove Proposition 1 lies in the structural properties of the value function, which are summarized in the following Lemma 2.

LEMMA 2. *$V_t(n)$ satisfies the following first- and second-order properties:*

(a) it is concave in n, that is, for each given t, $\Delta V_t(n)$ is decreasing in n;
(b) it is concave in t, i.e., for each given n: $V_t(n) - V_{t-1}(n) \geq V_{t-1}(n) - V_t(n)$;
(c) it is submodular in (t, n), i.e., $\Delta V_t(n)$ is increasing in t:

$$V_t(n) - V_t(n-1) \geq V_{t-1}(n) - V_{t-1}(n-1).$$

To prove Lemma 2, rewrite Equations (A.17)–(A.18) as

$$V_t(n) = V_{t-1}(n) + \sum_{i=1}^2 \lambda_t^i \left((r_i + c_i - \Delta V_{t-1}(n))^+ - c_i\right), \quad t = T, \ldots, 1, \quad \text{(A.23)}$$

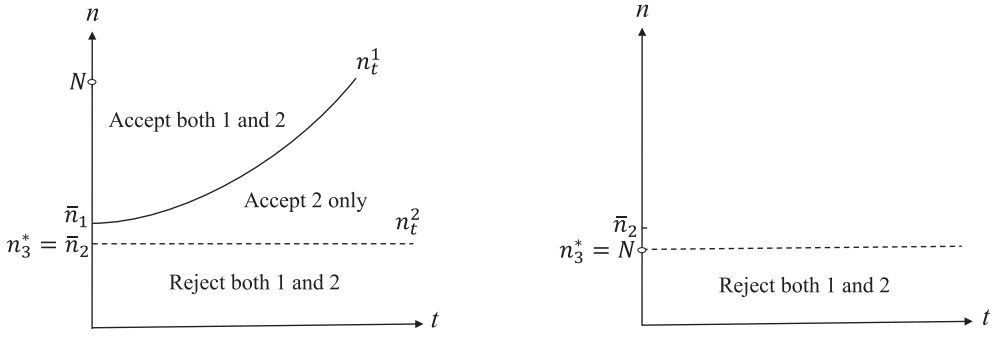
$$V_0(n) = r_3\mathbf{E}(D_3 \wedge n) - c_3\mathbf{E}(D_3 - n)^+ - \pi\mathbf{E}(n - D_3)^+. \quad \text{(A.24)}$$

(a) From (A.24), $V_0(n)$ is clearly concave in n, as $\wedge$ is concave and $(x)^+$ is convex. Inductively, suppose $V_{t-1}(n)$ is concave in n, that is, $\Delta V_{t-1}(n)$ is decreasing in n. From Equation (A.23), we have

$$\Delta V_t(n) = \Delta V_{t-1}(n) + \sum_{i=1}^2 \lambda_t^i [r_i + c_i - \Delta V_{t-1}(n)]^+ - \sum_{i=1}^2 \lambda_t^i [r_i + c_i - \Delta V_{t-1}(n-1)]^+. \quad \text{(A.25)}$$

It suffices to argue that the first two terms on the right hand side are decreasing in n. (The third term,

Figure 5 Optimal Control Policy



(a) $\bar{n}_2 < N$ **(b)** $\bar{n}_2 \geq N$

with a minus sign, is already decreasing in n, given the induction hypothesis.) Note, these two terms reduce to one of the following three possibilities:

$$\Delta V_{t-1}(n);\ (1-\lambda_t^i)\Delta V_{t-1}(n)+\lambda_t^i(r_i+c_i),$$
$$i=1,2;\ (1-\lambda_t^1-\lambda_t^2)\Delta V_{t-1}(n)+\sum_{i=1}^{2}\lambda_t^i(r_i+c_i);$$

and all three are decreasing in n, given the induction hypothesis.

(b) Concavity in t is equivalent to $V_t(n)-V_{t-1}(n)\geq V_{t-1}(n)-V_t(n)$. However, this follows immediately from Equation (A.23)—the second equality, along with $\Delta V_{t-1}(n)\leq \Delta V_t(n)$, the submodularity in (c)—to be proven below.

(c) It is equivalent to show $\Delta V_t(n)\geq \Delta V_{t-1}(n)$. However, this follows immediately from Equation (A.25), since the second term on the right- hand side dominates the third term, taking into account $\Delta V_{t-1}(n)\leq \Delta V_{t-1}(n-1)$.

Lemma 2 says that the value function is concave in n and t, respectively, and submodular in (t,n). The marginal value of an appointment is increasing in time, that is, it is higher at the start and lower when approaching to the end of the booking horizon.

We are now ready to prove Propostion 1. We will show (a) only, as the other claims are readily verified. To this end, we need to show

$$\Delta V_t(n)\leq \bar{r}_2,\ \forall n\geq \bar{n}_2;\ \Delta V_t(n)>\bar{r}_2,\ \forall n<\bar{n}_2. \quad \text{(A.26)}$$

by induction on t. The above claim clearly all holds at $t=0$. Suppose it holds for $t-1$ and we will show it holds for t.

By the definition of n_t^i in Equation (A.19) and Lemma 2(c) that $\Delta V_t(n)$ is increasing in t, it follows that n_t^i is increasing in t and $n_t^1\geq n_t^2$ given $\bar{r}_1\leq \bar{r}_2$. Consider n_t^1 along with $\bar{n}_2$ and Equation (A.20), we have

$$n_t^1\geq n_1^1\geq n_1^2:=\bar{n}_2.$$

Make use of Equation (A.25), and consider the following five cases.

(1) If $n\geq n_t^1+1$, we have $\Delta V_{t-1}(n)\leq \bar{r}_1\leq \bar{r}_2$ and $\Delta V_{t-1}(n-1)\leq \bar{r}_1\leq \bar{r}_2$, thus

$$\begin{aligned}\Delta V_t(n)&=\Delta V_{t-1}(n)+\lambda_t^1(\Delta V_{t-1}(n-1)-\Delta V_{t-1}(n))\\&\quad+\lambda_t^2(\Delta V_{t-1}(n-1)-\Delta V_{t-1}(n))\\&=(1-\sum_{i=1}^{2}\lambda_t^i)\Delta V_{t-1}(n)+\sum_{i=1}^{2}\lambda_t^i\Delta V_{t-1}\\&\quad(n-1)\leq \bar{r}_2.\end{aligned}$$

(2) If $n=n_t^1$, we have $\Delta V_{t-1}(n)\leq \bar{r}_1\leq \bar{r}_2$ and $\bar{r}_1<\Delta V_{t-1}(n-1)$; and taking into account $(\bar{r}_2-\Delta V_{t-1}(n-1))^+\geq \bar{r}_2-\Delta V_{t-1}(n-1))$, we have

$$\begin{aligned}\Delta V_t(n)&\leq \Delta V_{t-1}(n)+\lambda_t^1(\bar{r}_1-\Delta V_{t-1}(n))\\&\quad+\lambda_t^2(\Delta V_{t-1}(n-1)-\Delta V_{t-1}(n))\\&=(1-\sum_{i=1}^{2}\lambda_t^i)\Delta V_{t-1}(n)+\lambda_t^1\bar{r}_1+\lambda_t^2\Delta V_{t-1}\\&\quad(n-1)\leq \bar{r}_2.\end{aligned}$$

(3) If $\bar{n}_2+1\leq n<n_t^1$, we have $\bar{r}_1<\Delta V_{t-1}(n)\leq \bar{r}_2$ and $\bar{r}_1<\Delta V_{t-1}(n-1)\leq \bar{r}_2$, thus

$$\begin{aligned}\Delta V_t(n)&=\Delta V_{t-1}(n)+\lambda_t^2(\Delta V_{t-1}(n-1)-\Delta V_{t-1}(n))\\&=(1-\lambda_t^2)\Delta V_{t-1}(n)+\lambda_t^2\Delta V_{t-1}(n-1)\leq \bar{r}_2.\end{aligned}$$

(4) If $\bar{n}_2=n\ (<n_t^1)$, we have $\bar{r}_1<\Delta V_{t-1}(n)\leq \bar{r}_2$ and $\Delta V_{t-1}(n-1)>\bar{r}_2\geq \bar{r}_1$, thus

$$\begin{aligned}\Delta V_t(n)&=\Delta V_{t-1}(n)+\lambda_t^2(\bar{r}_2-\Delta V_{t-1}(n))\\&=(1-\lambda_t^2)\Delta V_{t-1}(n)+\lambda_t^2\bar{r}_2\leq \bar{r}_2.\end{aligned}$$

(5) If $n<\bar{n}_2$, we have $\Delta V_{t-1}(n)>\bar{r}_2\geq \bar{r}_1$ and $\Delta V_{t-1}(n-1)>\bar{r}_2\geq \bar{r}_1$, thus

$$\Delta V_t(n)=\Delta V_{t-1}(n)>\bar{r}_2.$$

This completes the induction, and hence, the proof.

A.4. Upper-Bound Solution to the Diagnosis Resource Allocation Problem

As before, for $i=1, 2, 3$, let D_i denote type-i demand, the total number of patients over the day. An upper-bound to the DP value function can be obtained from a standard "hind-sight" optimal argument. (This amounts to taking the expectation outside of max in the DP, which leads to an upper bound due to Jensen's inequality.)

Specifically, suppose we know D_i, the (realized) total number of type-i requests, for $i=1, 2$ (but not D_3). Then, we will allocate x_i slots to type-i, for $i=1, 2, 3$, so as to maximize the following objective:

$$\sum_{i=1}^{2}[r_iD_i-c_i(D_i-x_i)]+r_3\mathrm{E}(D_3\wedge x_3)-c_3\mathrm{E}(D_3-x_3)^+ -\pi\mathrm{E}(x_3-D_3)^+,$$

where for $i=1, 2, 3$, r_i is the revenue for each accepted patient and c_i the penalty cost for each rejected patient, and π is the penalty cost for any slot left unused. Note, for $i=1, 2$, we must have $x_i\leq D_i$; hence, there will not be any wasted slots.

Write $\bar{r}_i:=r_i+c_i$, for $i=1, 2, 3$. Re-organize terms, taking into account the identify $a\wedge b=a-(a-b)^+$, the upper-bound problem can be expressed as follows:

$$\max_{x_1,x_2,x_3} \bar{r}_1 x_1 + \bar{r}_2 x_2 - c_1 D_1 - c_2 D_2 + \bar{r}_3 x_3 - c_3 \mathbf{E}(D_3) - (\bar{r}_3 + \pi_3)\mathbf{E}(x_3 - D_3)^+ \quad \text{s.t. } x_i \le D_i,\ i = 1,2;\ x_1 + x_2 + x_3 = N;\ x_i \ge 0,\ i = 1,2,3. \tag{A.27}$$

To solve the above maximization problem, it is useful to envision an algorithm that increases the three variables x_i, $i = 1, 2, 3$, from zero until all N slots are exhausted or one or both of the other two constraints, $x_i \le D_i$, $i = 1, 2$, become binding. This will lead to the following:

$$\sum_{i=1}^{3} r_i \mathbf{E}(D_i) + \pi[\mathbf{E}(D_3) - x_3^*] - (\bar{r}_3 + \pi)\mathbf{E}(D_3 - x_3^*)^+ - (\bar{r}_2 - \bar{r}_1)\mathbf{E}[D_2 - (N - x_3^*)]^+ - \bar{r}_1 \mathbf{E}[D_1 + D_2 - (N - x_3^*)]^+, \tag{A.28}$$

where $x_3^* = F_3^{-1}\left(\frac{\bar{r}_3 - \eta}{\bar{r}_3 + \pi}\right)$, with F_3^{-1} being the inverse distribution function of D_3; and η is the Lagrangian multiplier w.r.t. the constraint $x_1 + x_2 + x_3 = N$ of the maximization problem in (A.27), satisfying

$$\eta \in [\eta^\ell, \eta^u], \text{ with } \eta^\ell := \bar{r}_3 - (\bar{r}_3 + \pi)F_3(N),\ \eta^u := \bar{r}_2 \vee \eta^\ell. \tag{A.29}$$

In the DP, the patient arrivals follow Poisson processes. Hence, D_i follows a Poisson distribution with mean λ_i, for $i = 1, 2, 3$. Now, for asymptotic analysis, let T be a scaling parameter, such that the means now become $\lambda_i T$. We can then approximate D_i by the normal distribution $N(\mu_i, \sigma_i^2)$, $i = 1, 2, 3$, with

$$\mu_i = \lambda_i T = \sigma_i^2, \quad i = 1, 2, 3. \tag{A.30}$$

For the analysis to be meaningful, we must assume N is scaled accordingly:

$$N = \lambda T + \kappa\sqrt{T}, \quad \exists \lambda > 0,\ \exists \kappa. \tag{A.31}$$

Because if N does not grow with T, then for sufficiently large T, it is trivially optimal to reserve all N slots for emergency patients and reject all other requests.

This way, with straightforward algebra, the upper bound in Equation (A.28) can be expressed as follows:

$$\sum_{i=1}^{3} r_i \mu_i - \bar{r}_3 \sigma_3 G(z_3^*) - \pi\sigma_3[z_3^* + G(z_3^*)] - (\bar{r}_2 - \bar{r}_1)\sigma_2 G\left(\frac{N - \mu_2 - \mu_3 - \sigma_3 z_3^*}{\sigma_2}\right) - \bar{r}_1\sqrt{\sigma_1^2 + \sigma_2^2}\, G\left(\frac{N - \mu_1 - \mu_2 - \mu_3 - \sigma_3 z_3^*}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \tag{A.32}$$

where

$$z_3^* := \Phi^{-1}\left(\frac{\bar{r}_3 - \eta}{\bar{r}_3 + \pi}\right), \quad \eta \in [\eta^\ell, \eta^u]; \tag{A.33}$$

with η^u and η^ℓ defined in Equation (A.29), replacing $F_3(N)$ by $\Phi(\frac{N - \mu_3}{\sigma_3})$.

In the upper bound in Equation (A.32), z_3^* still depends on η, which, in turn, depends on D_1 and D_2 (via x_1^* and x_2^*). To overcome this handicap, we replace η by either η^ℓ or η^u, that is, replace z_3^* by its upper- and lower-bounds, z_3^u and z_3^ℓ, as follows:

$$z_3^u := \Phi^{-1}\left(\frac{\bar{r}_3 - \eta^\ell}{\bar{r}_3 + \pi}\right) \ge z_3^* \ge z_3^\ell := \Phi^{-1}\left(\frac{\bar{r}_3 - \eta^u}{\bar{r}_3 + \pi}\right). \tag{A.34}$$

Making use of the expressions for η^u and η^ℓ in Equation (A.29), we have

$$z_3^u = \frac{N - \mu_3}{\sigma_3}, \quad z_3^\ell = z_3^u \wedge \Phi^{-1}\left(\frac{\bar{r}_3 - \bar{r}_2}{\bar{r}_3 + \pi}\right). \tag{A.35}$$

Note, on the RHS of Equation (A.36), $-\bar{r}_3\sigma_3 G(z_3^*)$ is the only increasing term (in particular, note that $-\pi\sigma_3(z_3^* + G(z_3^*))$ is decreasing, since $x + G(x)$ is increasing). Hence, we replace z_3^* by z_3^u in $-\bar{r}_3\sigma_3 G(z_3^*)$ and by z_3^ℓ in all other terms. Thus, the upper-bound objective value can be expressed as follows:

$$V^u := \sum_{i=1}^{3} r_i\mu_i - \bar{r}_3\sigma_3 G(z_3^u) - \pi\sigma_3(z_3^\ell + G(z_3^\ell)) - (\bar{r}_2 - \bar{r}_1)\sigma_2 G\left(\frac{N - \mu_2 - \mu_3 - \sigma_3 z_3^\ell}{\sigma_2}\right) - \bar{r}_1\sqrt{\sigma_1^2 + \sigma_2^2}\, G\left(\frac{N - \mu_1 - \mu_2 - \mu_3 - \sigma_3 z_3^\ell}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \tag{A.36}$$

References

Allon, G., S. Deo, W. Lin. 2013. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Oper. Res.* **61**(3): 544–562.

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stoch. Syst.* **5**(1): 146–194.

Bavafa, H., C. Leys, L. Örmeci, S. Savin. 2018. Managing portfolio of elective surgical procedures: A multidimensional inverse newsvendor problem. Working paper, University of Wisconsin, Madison, WI.

Belobaba, P. P. 1987. Airline yield management: An overview of seat inventory control. *Transport. Sci.* **21**(2): 63–73.

Belobaba, P. P. 1989. Application of a probabilistic decision model to airline seat inventory control. *Oper. Res.* **37**(2): 183–197.

Cardoen, B., E. Demeulemeester, J. Beliën. 2010. Operating room planning and scheduling: A literature review. *Eur. J. Oper. Res.* **201**(3): 921–932.

Cayirli, T., E. Veral. 2003. Outpatient scheduling in health care: A review of literature. *Prod. Oper. Manag.* **12**(4): 519–549.

Chan, C. W., V. F. Farias, N. Bambos, G. J. Escobar. 2012. Optimizing intensive care unit discharge decisions with patient readmissions. *Oper. Res.* **60**(6): 1323–1341.

Chapman, S. N., J. I. Carmel. 1992. Demand/capacity management in health care: An application of yield management. *Health Care Manage. Rev.* **17**(4): 45–54.

Cochran, J. K., A. Bharti. 2006. Stochastic bed balancing of an obstetrics hospital. *Health Care Manage. Sci.* **9**(1): 31–45.

Costa, A. X., S. A. Ridley, A. K. Shahani, P. R. Harper, V. De Senna, M. S. Nielsen. 2003. Mathematical modelling and simulation for planning critical care capacity. *Anaesthesia* **58**(4): 320–327.

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

Deo, S., S. Iravani, T. Jiang, K. Smilowitz, S. Samuelson. 2013. Improving health outcomes through better capacity allocation in a community-based chronic care model. *Oper. Res.* **61**(6): 1277–1294.

Erdelyi, A., H. Topaloglu. 2011. Approximate dynamic programming for dynamic capacity allocation with multiple priority levels. *IIE Trans.* **43**, 129–142.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Sci.* **42**(3): 321–334.

Green, L. V. 2005. Capacity planning and management in hospitals. M. L. Brandeau, F. Sainfort, W. P. Pierskalla eds. *Operations Research and Health Care*. International Series in Operations Research & Management Science, vol. 70. Springer, Boston, MA, 15–41.

Green, L. V. 2006. Queueing analysis in healthcare. R. Hall, ed. *Patient Flow: Reducing Delay in Healthcare Delivery*. International Series in Operations Research & Management Science, vol. 91. Springer, Boston, MA, 281–307.

Green, L. V., V. Nguyen. 2001. Strategies for cutting hospital beds: The impact on patient service. *Health Serv. Res.* **36**(2): 421.

Green, L. V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Oper. Res.* **56**(6): 1526–1538.

Green, L. V., S. Savin, B. Wang. 2006a. Managing patient service in a diagnostic medical facility. *Oper. Res.* **54**(1): 11–25.

Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006b. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Acad. Emer. Med.* **13**, 61–68.

Griffin, J., S. Xia, S. Peng, P. Keskinocak. 2012. Improving patient flow in an obstetric unit. *Health Care Manage. Sci.* **15**(1): 1–14.

Guerriero, F., R. Guido. 2011. Operational research in the management of the operating theatre: A survey. *Health Care Manage. Sci.* **14**(1): 89–114.

Gupta, D. 2007. Surgical suite's operations management. *Prod. Oper. Manag* **16**(6): 689–700.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* **40**, 800–819.

Gupta, D., L. Wang. 2008. Revenue management for a primary-care clinic in the presence of patient choice. *Oper. Res.* **56**(3): 576–592.

Hall, R., D. Belson, P. Murali, M. Dessouky. 2006. Modeling patient flows through the healthcare system. R. Hall, ed. *Patient Flow: Reducing Delay in Healthcare Delivery*. International Series in Operations Research & Management Science, vol. 91. Springer, Boston, MA, 1–44.

Helm, J. E., M. P. Van Oyen. 2014. Design and optimization methods for elective hospital admissions. *Oper. Res.* **62**(6): 1265–1282.

Helm, J. E., S. AhmadBeygi, M. P. Van Oyen. 2011. Design and analysis of hospital admission control for operational effectivenss. *Prod. Oper. Manag.* **20**(3): 359–374.

Huang, J., B. Carmeli, A. Mandelbaum. 2015. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Oper. Res.* **63**(4): 892–908.

Huh, W. T., N. Liu, V. Truong. 2013. Multiresource allocation scheduling in dynamic environments. *Manuf. Serv. Oper. Manag.* **15**(2): 280–291.

Jack, E. P., T. L. Powers. 2009. A review and synthesis of demand management, capacity management and performance in health-care service. *Int. J. Manage. Rev.* **11**(2): 149–174.

Jacobson, S. H., S. N. Hall, J. R. Swisher. 2006. Discrete-event simulation of health care systems. R. Hall, ed. *Patient Flow: Reducing Delay in Healthcare Delivery*. International Series in Operations Research & Management Science, vol. 91. Springer, Boston, MA, 211–252.

Kim, S. C., I. Horowitz, K. K. Young, T. A. Buckley. 2000. Flexible bed allocation and performance in the intensive care unit. *J. Oper. Manag.* **18**(4): 427–443.

Kim, S.-H., C. W. Chan, M. Olivares, G. Escobar. 2014. ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Sci.* **61**(1): 19–38.

Kimes, S. E. 1989. Yield management: A tool for capacity-constrained service firms. *J. Oper. Manag.* **8**(4): 348–363.

Kong, Q., C.-Y. Lee, C.-P. Teo, Z. Zheng. 2013. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Oper. Res.* **61**(3): 711–726.

Magerlein, J. M., J. B. Martin. 1978. Surgical demand scheduling: A review. *Health Serv. Res.* **13**(4): 418–433.

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

May, J. H., W. E. Spangler, D. P. Strum, L. G. Vargas. 2011. The surgical scheduling problem: Current research and future opportunities. *Prod. Oper. Manag.* **20**(3): 392–405.

Mondschein, S. V., G. Y. Weintraub. 2003. Appointment policies in service operations: A critical analysis of the economic framework. *Prod. Oper. Manag.* **12**(2): 266–286.

Patrick, J., M. L. Puterman, M. Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Oper. Res.* **56**(6): 1507–1525.

Sauré, A., J. Patrick, M. L. Puterman. 2015. Simulation-based approximate policy iteration withgeneralized logistic functions. *INFORMS J. Comput.* **27**(3): 579–595.

Shi, P., M. C. Chou, J. G. Dai, D. Ding, J. Sim. 2015. Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Sci.* **62**(1): 1–28.

Smith-Daniels, V. L., S. B. Schweikhart, D. E. Smith-Daniels. 1988. Capacity management in health care service: Review and future-research directions. *Decis. Sci.* **19**(4): 889–919.

Thompson, S., M. Nunez, R. Garfinkel, M. D. Dean. 2009. Efficient short-term allocation and reallocation of patients to

floors of a hospital during demand surges. *Oper. Res.* **57**(2): 261–273.

Truong, V.-A. 2015. Optimal advance scheduling. *Management Sci.* **61**(7): 1584–1597.

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

Zeltyn, S., Y. N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, T. Lauterman, D. Schwartz, K. Moskovitch, S. Tzafrir, F. Basis. 2011. Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Trans. Model. Comput. Simul.* **21**(4): Article 24.