Submitted to *Manufacturing & Service Operations Management*
manuscript (Please, provide the manuscript number!)

A Clustered Overflow Configuration of Inpatient Beds in Hospitals

Navid Izady
Cass Business School, City, University of London, navid.izady@city.ac.uk

Israa Mohamed
Department of Operations Research, Zagazig University, israasalem@zu.edu.eg

Problem Definition The shortage of inpatient beds is a major cause of delays and cancellations in many hospitals. It may also lead to patients being admitted to inappropriate wards, whereby resulting in a lower quality of care and a longer length of stay. **Academic/Practical Relevance** Investment in additional beds is not always feasible. Instead, new and creative solutions for a more efficient use of existing resources must be sought. **Methodology** We propose a new configuration of inpatient beds which we call the clustered overflow configuration. In this configuration, patients who are denied admission to their primary wards as a result of beds being fully occupied are admitted to overflow wards, with each designated to serve overflows from a certain subset of specialties and providing the same quality of care as in primary wards. We propose two different formulations for partitioning and bed allocation in the proposed configuration: one minimizing the sum of average daily costs of turning patients away and nursing teams, and another minimizing the numbers turned away subject to nursing cost falling below a given threshold. We heuristically solve instances from both formulations. **Results** Applying the models to real data shows that the configurations obtained from our models compare very well with the other configurations proposed in the literature, provided that patients' willingness to wait is relatively short. **Managerial Implications** The proposed configuration provides the combined advantages of the dedicated configuration, wherein patients are only admitted to their primary wards, and the flexible configuration, in which all specialties share a single ward. On the other hand, it restricts the adverse impacts of pooling and minimizes cross-training costs through appropriate partitioning and bed allocation. As such, it serves as a viable alternative to existing inpatient configurations.

Key words: Health Care Management; Queueing Theory; Stochastic Models

1. Introduction

The number of inpatient beds is the most fundamental measure of hospital capacity (Green 2004). The lack of appropriate inpatient beds for post-treatment care is a major cause of surgery cancellations. It may also lead to some patients being admitted to inappropriate wards (so-called patient 'outlying'), cared for on trolleys in the hallways or emergency rooms while waiting for a bed to become available (so-called 'trolley wait'), or discharged/transferred early in order to make room for new admissions. The shortage of inpatient beds is often cited as the single most important factor contributing to overcrowding in emergency departments and subsequent ambulance diversion (e.g. Olshaker and Rathlev 2006).

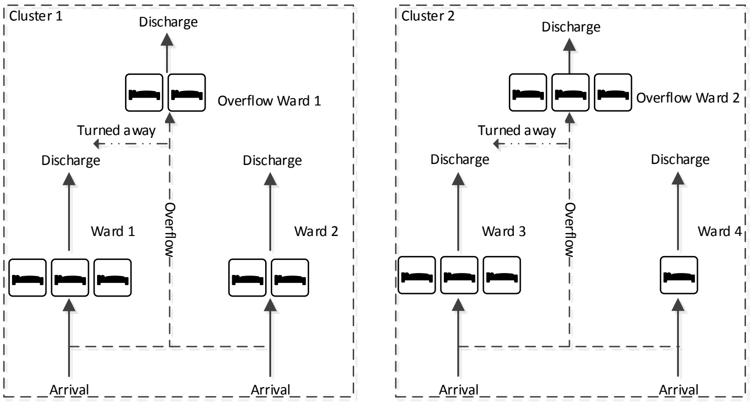
Despite advances in the medical technology that has enabled a move to day surgery, a reduced need for hospitalization and a shortened length of stay (LOS), the availability of inpatient beds is under strain worldwide. In the UK, for example, the average bed occupancy (occupied bed days divided by available bed days) reached an all-time high of 90.0% in the final quarter of 2017/18 (NHS England 2018). The same measure in the US shows an increasing trend over the last decade (National Center for Health Statistics 2016). The situation is more critical in developing countries, with an average of 0.7 beds per 1000 capita compared to 5.6 in the developed world (World Health Organization 2014). For example, in the paediatric department of a hospital in Egypt that motivated this study, more than 2000 children are denied admission every year because of a lack of inpatient beds. Investment in additional beds is not always feasible due to a range of financial, legislative or space constraints (Best et al. 2015). Instead, new and creative solutions for a more efficient use of existing resources must be sought.

In some hospitals, inpatient beds are organized into a number of clinical units, which we refer to as wards, with each having its own dedicated staff and equipment and serving patients of a certain medical specialty, e.g. neurology, cardiology or gynaecology. Once a decision is made to admit a patient, the most appropriate ward depending on the patient's diagnosis is selected, often called the primary ward of the patient. If that ward is fully occupied, two different situations might arise. In a 'dedicated' configuration, the patient is turned away (e.g. transferred to another hospital) once a waiting time threshold is passed, while in a 'dedicated with outlying' configuration, the patient is admitted to a non-primary ward selected according to some prioritization rule. Some other hospitals, on the other hand, operate a 'flexible' configuration in which most, if not all, specialties share a single ward.

The dedicated configuration provides the advantages of 'focused care' (Best et al. 2015), while the flexible configuration enjoys the benefits of 'pooling' (Green and Nguyen 2001). In this paper, we propose a new configuration combining the advantages of dedicated and flexible configurations, and compare its performance with that of the other configurations proposed in the literature. In this configuration, which we name the 'clustered overflow' configuration, specialties are partitioned into a number of clusters (see the example in Figure 1). Each cluster includes a dedicated ward for each of the specialties in the cluster as well as a single shared ward called the overflow ward. Patients who cannot be accommodated in the dedicated ward of their cluster will be admitted to the overflow ward of the cluster if it has an empty bed available. Otherwise, they will be turned away once a waiting time threshold is reached. The overflow ward of each cluster is staffed with multi-skilled nursing teams capable of caring for all of the specialties in the cluster.

The main advantages of the dedicated configuration are those associated with focused care, as a narrow and cohesive set of conditions are typically treated in each ward. This leads to '... reduced complexity, lower uncertainty, and the development of specialized expertise' (Clark and Huckman 2012, p. 708). In particular, empirical studies by KC and Terwiesch (2011) and Clark and Huckman (2012) suggest that focused care in hospitals shortens the LOS and improves the quality of care. Best et al. (2015) also report an average reduction of 7.3% in LOSs after the formation of specialized wards at the University of Chicago Medical Center, citing the staff increased sense of ownership over the beds in their ward as the main underlying reason. They further point out that

Figure 1 A Clustered Overflow Configuration with 14 Beds, 4 Specialties and 2 Clusters; Cluster 1 (2) Includes Specialties 1 and 2 (3 and 4).



the dedicated configuration gives hospitals the flexibility to reserve beds for some specialties while restricting them for others based on the utility that they provide to the hospital. The disadvantage of the dedicated configuration is that patients are turned away (or forced to wait) when their primary ward is full even if empty beds are available elsewhere. On the other hand, with the flexible configuration, patients are admitted as long as empty beds are available. The increase in the LOS due to a lack of focus and extra LOS variability induced by mixing patients from different specialties might, however, offset the advantages of pooling. The dedicated with outlying configuration performs in relatively the same way as that of the flexible configuration, particularly in the common (albeit extreme) situation in which patients are outlied in any non-primary ward. The difference is that in the flexible configuration the nursing teams are typically multi-skilled, while in the dedicated with outlying configuration they are not, thereby compromising the quality of care.

The clustered overflow configuration inherits the advantages of focused care in its dedicated wards. It also utilizes the benefits of pooling in its overflow wards, while minimizing the adverse impact of mix variability through the clustering of specialties and the appropriate distribution of beds between dedicated and overflow wards. There are also administrative advantages in the clustered overflow configuration compared to the dedicated with outlying configuration. In the latter, outlying patients may be placed in any non-primary ward, whereby making it difficult to locate and monitor them. There is also often some ambiguity as to where the ultimate responsibility of outlying patients' care lies, i.e. with the primary or non-primary wards. In the clustered overflow configuration, each cluster is responsible for serving a certain subset of specialties, in either its dedicated or its overflow wards, hence higher accountability and traceability. Note that the clustered overflow configuration proposed here is different from the general overflow configuration implemented in some hospitals in which some general wards are designated as overflow wards admitting patients of all specialties, often as a temporary measure when the hospital is under pressure. The aim here is that the quality of care delivered to patients in the overflow wards must be at the same level as in the dedicated wards. This is achieved through cross-training of overflow nursing teams.

To implement the clustered overflow configuration, one must identify the partitioning of specialties and the associated allocation of beds to different wards. Given an overall number of beds and a set of specialties, we propose two different formulations for this purpose. The first formulation, referred to as the total cost minimization (TCM), seeks to find a partition and corresponding bed allocation minimizing the sum of expected daily costs of denied admissions and nursing teams. The second formulation, referred to as the constrained blocking minimization (CBM), aims to find a solution that minimizes the total number of patients turned away subject to nursing cost falling below a given threshold. The TCM follows a conventional approach by assuming that the hospital incurs a cost for each patient turned away; see, for example, Belciug and Gorunescu (2015). The CBM, on the other hand, is more practical, giving hospital administrators the flexibility to choose a configuration depending on the additional nursing cost that they are prepared to incur. The underlying assumptions for both formulations are that A(i) beds are fully flexible, and A(ii) the waiting time threshold is zero. Following the bed allocation literature, we also initially assume that admission requests form a Poisson process and that LOSs are exponentially distributed. Motivated by our empirical observations, however, we then generalize our models in order to relax this assumption and numerically investigate its impact on results.

Assumption A(i) is commonly adopted in the bed allocation literature; see, for example, Best et al. (2015). It is typically valid when the focus of reconfiguration is on a particular level of care acuity, e.g. intensive, medium or normal care. A(ii) implies that patients are either assigned to a specific ward or turned away immediately, enabling us to model the wards as loss systems, i.e. queueing systems with no waiting provision. This is a realistic assumption in some settings, such as the hospital that has motivated this study or Dutch hospitals, wherein the fraction of transfers to other hospitals due to a lack of inpatient beds is significant while waiting times for inpatient beds are short (Bekker et al. 2016). In other settings in which patients may wait longer for bed assignment, loss systems still provide a good approximation for the relative performance of different configurations (Chevalier and Van den Schrieck 2008). For example, Tabordon (2002) shows that in delay systems where customers are allowed to wait but are impatient and their time to abandonment is short (e.g. has the same mean as service time), the rate of abandonment is proportional to the loss probability in an equivalent loss system. Numerical experiments conducted in Chevalier et al. (2005) on systems with infinitely patient customers also show that the allocation of resources among dedicated and flexible servers obtained under a zero-waiting-time assumption is nearly optimal when service quality constraints are tight.

Our solution methodology for both TCM and CBM formulations involves decomposing the full problem into an intra-cluster bed allocation problem and a partitioning and inter-cluster allocation problem. Exact methods, facilitated by fixing a sequence of specialties, are used for solving the partitioning and inter-cluster allocation problem, and a heuristic search method for solving the intra-cluster allocation problem. A performance evaluation model is embedded within our intra-cluster allocation model for estimating the performance metrics of a given cluster with specific bed allocation. This model has two novel features. Firstly, it works with non-Poisson arrival processes and general LOS distributions. Secondly, it accounts for potentially different mean LOSs in dedicated and overflow wards.

We apply our models to the data collected from the paediatric department in an Egyptian hospital, hereinafter referred to as HUS. The results show that under TCM the best configuration obtained from our models is a clustered overflow configuration. Under CBM, on the other hand, the best found configuration is typically a 'wing formation' configuration proposed in Best et al. (2015) when the impact of focus is negligible, and a clustered overflow configuration otherwise. The number of clusters in our best configurations tends to decrease (increase) with the cost of turning patients away and the nursing budget (the demand for admission). We run simulation experiments with the best configurations obtained from our models as well as the other configurations proposed in the literature. In the simulation model, we relax A(ii) and assume that patients' waiting time threshold is random following an exponential distribution. The simulation experiments show that in general the configurations obtained from our models perform better than the other configurations, provided that the mean waiting time threshold is relatively short (set to one and seven days in our experiments). In particular, we observe that under TCM our model configurations are the lowest-cost configurations in twenty nine out of the thirty two scenarios investigated. Similarly, we observe that under CBM our model configurations result in the lowest numbers of patients abandoned in twenty six out of twenty eight scenarios in which the nursing cost threshold is met. In four scenarios, however, the nursing budget is slightly exceeded with our model configurations. Our experiments with the models that permit non-Poisson arrivals show that the configurations obtained from these models could create further improvement in performance, especially under TCM or when deviation from Poisson is significant. We finally observe that in the configurations obtained from non-Poisson models, the size of overflow wards (the number of clusters) tends to increase (decrease) as admission requests become more variable, implying that the benefits of pooling rise with the variability of the arrival process.

2. Literature Review

Hospital bed planning and allocation is a classic problem studied by many researchers over the years; see the reviews in Green and Nguyen (2001) and Hall (2012). A large number of these studies seek to determine the required number of beds in a particular ward. Some of these works use discrete event simulation models, e.g. Kokangul (2008), while others rely on analytical queueing models including delay models, e.g. Green and Nguyen (2001), and loss models, e.g. de Bruin et al. (2009) and Belciug and Gorunescu (2015). The objective is typically to achieve a given bed occupancy, admission probability and/or average waiting time. Redistribution of beds to different wards in a fully dedicated configuration is investigated by, for example, Huang (1998) and Ma and Demeulemeester (2013). Bekker et al. (2016) evaluate the flexible configuration, and the dedicated with outlying configuration is modelled in Shi et al. (2015). The disadvantages of patient outlying are discussed in the medical literature; Lloyd et al. (2005) suggest that outlying patients often receive suboptimal nursing care, and Stowell et al. (2013) provide empirical evidence indicating that outlying patients are likely to have a longer LOS, higher readmission rate, and insufficient thromboembolic prevention.

To the best of our knowledge, the clustered overflow configuration of beds in hospitals has been neither investigated in the literature previously nor implemented in practice. The closest that we found in the literature is the 'earmarking' policy considered in Bekker et al. (2016), wherein a single overflow ward is shared by all specialties, each of which has its own dedicated ward. The major advantage of our proposed configuration to earmarking is that, with specialties partitioned into smaller clusters, it is much easier to provide cross-training, thus being more likely to achieve the same quality of care in the overflow wards as in the dedicated wards. Partitioning also creates an extra layer of protection against mix variability, hence further opportunities for performance improvement.

Our research builds upon the work of Best et al. (2015), and generalizes it to the clustered overflow configurations. They consider a wing formation configuration in which a cluster of specialties is allocated to each ward (referred to as a 'wing'), serving patients in a dedicated manner. Given a fixed number of beds and exponential waiting time thresholds, Best et al. (2015) develop an optimization framework to determine the number of wings to form, the number of beds to allocate to each wing, and the set of specialties to assign to each wing. They provide the first analytical model capturing the impact of focus and workload on LOSs in a ward. We use the dynamic and integer programming solution methods proposed by Best et al. (2015) in our partitioning and inter-cluster bed allocation, apply their restrictive sequencing approach in order to simplify the partitioning problem, and adopt the generalized logistic function that they propose so as to adjust the LOS based on the level of focus and the amount of workload in a ward. In contrast to Best et al. (2015), we use loss models in our performance evaluation as developing approximations for delay models (with or without abandonment) for the clustered overflow configuration would be quite challenging. As a result, our models work best when patients' waiting time threshold is relatively short, while the models in Best et al. (2015) work with long waiting time thresholds as well.

Apart from introducing a new configuration, our work is distinguished from Best et al. (2015) and the other papers cited above for the following reasons. Firstly, it is the first work to explicitly consider nursing costs in the clustering of specialties and the corresponding bed allocation. Nursing costs typically account for around half of hospital expenses (Kazahaya 2005), and there is significant evidence relating higher nursing levels to lower rates of adverse patient outcomes and mortality (see, for example, Aiken et al. 2002). Nursing levels and the associated cost, particularly the cross-training cost, are therefore an important consideration in deciding bed configuration. Secondly, we relax the widely made assumption of exponential inter-arrival and LOS distributions. Thirdly, our mathematical representation of the overflow configuration includes the other configurations considered in the literature, enabling us to undertake a thorough investigation.

Another area relevant to our work is the literature on staffing and routing in call centre and telecommunication networks, wherein handling customers blocked from specialized primary facilities in flexible overflow facilities is a well-known strategy (see Koole and Pot 2006 for a review). Evaluating loss probabilities is a major requirement for the optimal design of such hierarchical networks. Major approximations proposed in the literature include the equivalent random method (see, for example, Cooper 1981, p. 165), Hayward's approximation (Fredericks 1980), and hyper-exponential decomposition (Franx et al. 2006). Our performance evaluation model is based on Hayward's approximation, and requires evaluating a 'peakedness' measure for overflow streams. We derive an exact expression for calculating peakedness for systems with mean service time in overflow facilities potentially different from dedicated facilities (due to, for example, staff with different levels of expertise), whereby resulting in a more accurate estimation of performance metrics.

At a more strategic level, the optimal configuration of service facilities facing uncertain demand from multiple customer classes is addressed in two different streams of literature. The first deals with service facilities with naturally flexible servers, and investigates whether facilities should be dedicated or pooled (see, for example, van Dijk and van der Sluis 2008). The key finding is that a pooled configuration is superior if customer classes have the same service time distributions, but not necessarily otherwise. The second stream is about facilities in which servers' flexibility can be achieved at a cost (see, for example, Jordan and Graves 1995, Bassamboo et al. 2010). The key finding is that under high traffic a little flexibility is all that we need. A chained configuration, in particular, that uses only bi-flexible servers can achieve almost all of the benefits of full flexibility.

In contrast to the call centre and flexibility literature outlined above, we assume that there exists only one overflow ward (cross-trained resource) for each specialty (customer class); otherwise, issues of traceability and accountability similar to those associated with patient outlying would arise. This implies that the results obtained in the aforementioned literature cannot be applied in our context. For example, the tailored chaining and pairing configurations proposed in Bassamboo et al. (2010) would not be feasible in our case, as they would allow more than one overflow ward for each specialty. Furthermore, since the arrival rates for some specialties can be small, it is essential to consider the integrality of bed numbers and the resulting combinatorial complexity. This is often avoided in call centre and flexibility literature by assuming heavy-traffic limits and continuous capacity sizing. Investigating the impact of flexibility in the bed configuration context would therefore require a different treatment. Indeed, Best et al. (2015) highlight the importance of considering flexibility in bed configuration but leave it for future research due to its complexity. Our study is, in fact, the first paper to propose a methodology for such a purpose.

3. The Model

In this section, we first propose a mathematical representation for the clustered overflow configuration. We then propose two different formulations for partitioning and bed allocation, followed by an analytical approximation for performance evaluation.

3.1. Mathematical Representation

Let n be the number of medical specialties cared for at a particular acuity level, and let $\mathcal{S} = \{1, 2, \ldots, n\}$ be the corresponding index set (see Best et al. 2015, p. 167 for considerations on defining $\mathcal{S}$). Let λ^i be the arrival rate for patients of specialty $i \in \mathcal{S}$, and denote with $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$ a partition of set $\mathcal{S}$ into m clusters. For each cluster $\mathcal{C}_j$, we assume that there exists a dedicated ward i for each specialty $i \in \mathcal{C}_j$, as well as a single overflow ward j for $j = 1, \ldots, m$. The overflow ward in each cluster serves overflows from the dedicated wards in that cluster. Let $\mathbf{d} = (d^1, \ldots, d^n)$ and $\mathbf{o} = (o_1, \ldots, o_m)$ be bed allocation vectors, with $d^i \in \mathbb{Z}$ representing the number of beds in dedicated ward i for $i \in \mathcal{S}$, and $o_j \in \mathbb{Z}$ representing the number of beds in overflow ward j for $j = 1, \ldots, m$. (We use $\mathbb{Z}$ and $\mathbb{Z}_+$, respectively, to denote the set of nonnegative and positive integers; see a summary of notations in the online appendix.) For the overflow configuration illustrated in Figure 1, $\mathcal{S} = \{1, 2, 3, 4\}$, $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \mathcal{C}_2\}$ with $\mathcal{C}_1 = \{1, 2\}$ and $\mathcal{C}_2 = \{3, 4\}$, $\mathbf{d} = (3, 2, 3, 1)$, and $\mathbf{o} = (2, 3)$.

Our representation of the clustered overflow configuration captures a range of other configurations proposed in the literature as special cases. A dedicated configuration with exactly one ward allocated to each specialty is represented by $\boldsymbol{\mathcal{C}} = \{\{1\}, \ldots, \{n\}\}$, $\mathbf{d} > \mathbf{0}$, and $\mathbf{o} = \mathbf{0}$. The wing formation configuration of Best et al. (2015) is represented by $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$, $\mathbf{d} = \mathbf{0}$, and $\mathbf{o} > \mathbf{0}$, where all specialties in a cluster are assumed to be served in the overflow ward of that cluster. The earmarking configuration of Bekker et al. (2016) is represented by $\boldsymbol{\mathcal{C}} = \{\mathcal{S}\}$, $\mathbf{d} \geq \mathbf{0}$, and $\mathbf{o} = o_1 > 0$. A fully flexible configuration is represented as $\boldsymbol{\mathcal{C}} = \{\mathcal{S}\}$, $\mathbf{d} = \mathbf{0}$, and $\mathbf{o} = B$, where B is the total number of beds.

3.2. Methodology

Given a total of B beds available to specialties in $\mathcal{S}$, in this section we propose two different formulations for partitioning and bed allocation. In mathematical terms, the decisions to make in both formulations concern (i) the number of clusters $m \in \mathbb{Z}_+$ to create; (ii) the non-empty set of specialties $\mathcal{C}_j$ to assign to cluster j for $j = 1, \ldots, m$, so that the assignment results in a partition of set $\mathcal{S}$; (iii) the number of beds d^i to allocate to the ward dedicated to specialty i patients for $i \in \mathcal{S}$; and (iv) the number of beds o_j to allocate to the overflow ward of cluster j for $j = 1, \ldots, m$.

TCM In this formulation, we consider the problem

$$Z=\min_{(m,\boldsymbol{\mathcal{C}},\mathbf{d},\mathbf{o})}\left\{\sum_{j=1}^{m}T_j(m,\boldsymbol{\mathcal{C}},\mathbf{d},\mathbf{o}):\sum_{i=1}^{n}d^i+\sum_{j=1}^{m}o_j\leq B,\boldsymbol{\mathcal{C}}\text{ is a partition of }\mathcal{S},m\in\mathbb{Z}_+,\mathbf{d}\in\mathbb{Z}^n\text{ and }\mathbf{o}\in\mathbb{Z}^m\right\},\tag{1}$$

where $T_j(m,\boldsymbol{\mathcal{C}},\mathbf{d},\mathbf{o})$ is the total average daily cost of cluster $\mathcal{C}_j$, including the cost of turning patients away and the nursing cost, under feasible solution $(m,\boldsymbol{\mathcal{C}},\mathbf{d},\mathbf{o})$. Since in our configurations, clusters operate independently of each other, i.e. there is no overflow of patients or sharing of staff between them, T_j for a given cluster would depend only on the number of beds allocated to its dedicated and overflow wards, i.e. $T_j(m,\boldsymbol{\mathcal{C}},\mathbf{d},\mathbf{o})=T_j((d^i;i\in\mathcal{C}_j),o_j)$. Now, because the optimal allocation of a given number of beds among different wards of a cluster is not influenced by the allocation of beds among wards in other clusters, problem (1) can be restated as

$$Z=\min_{(m,\mathbf{b},\boldsymbol{\mathcal{C}})}\left\{\sum_{j=1}^{m}\phi_j(\mathcal{C}_j,b_j):(m,\mathbf{b},\boldsymbol{\mathcal{C}})\in\Psi\right\},\tag{2}$$

where $\mathbf{b}=(b_1,\ldots,b_m)$, $\Psi=\Big\{(m,\mathbf{b},\boldsymbol{\mathcal{C}}):\sum_{j=1}^{m}b_j\leq B,\boldsymbol{\mathcal{C}}\text{ is a partition of }\mathcal{S},m\in\mathbb{Z}_+,\text{ and }\mathbf{b}\in\mathbb{Z}^m\Big\}$, and, given $\mathcal{C}_j$ and b_j,

$$\phi_j(\mathcal{C}_j,b_j)=\min_{(d^i;i\in\mathcal{C}_j),\,o_j}\left\{T_j((d^i;i\in\mathcal{C}_j),o_j):o_j+\sum_{i\in\mathcal{C}_j}d^i\leq b_j,o_j\in\mathbb{Z},\text{ and }d^i\in\mathbb{Z}\text{ for }i\in\mathcal{C}_j\right\}.\tag{3}$$

To evaluate T_j, let c be the average cost of turning a patient away. Then $T_j((d^i;i\in\mathcal{C}_j),o_j)=cQ_j((d^i;i\in\mathcal{C}_j),o_j)+R_j((d^i;i\in\mathcal{C}_j),o_j)$, where Q_j and R_j represent the expected daily number of denied admissions and the expected daily cost of nursing staff, respectively, for a cluster $\mathcal{C}_j$ with bed allocation $(d^i;i\in\mathcal{C}_j)$, o_j. Denoting with $p_j^k((d^i;i\in\mathcal{C}_j),o_j)$ the probability of a patient of specialty $k\in\mathcal{C}_j$ being denied admission to cluster $\mathcal{C}_j$, we compute Q_j as

$$Q_j((d^i;i\in\mathcal{C}_j),o_j)=\sum_{k\in\mathcal{C}_j}\lambda^k p_j^k((d^i;i\in\mathcal{C}_j),o_j).\tag{4}$$

To evaluate R_j, we use the minimum nurse-to-patient ratio approach. This is partly because it is the most common method for establishing nursing requirements in hospitals, and partly due to the simplicity of its application in our formulation. We note, however, that there are problems associated with fixed nurse-to-patient ratios. In particular, they do not reflect variations in nursing skills, the severity of patients' illnesses, and the size of the clinical units; see, for example, Lang et al. (2004) and Kane et al. (2007). Other approaches, such as the queueing model proposed in Yankovic and Green (2011), could be used here instead but they would add to the complexity of our models. Let f^i denote the desired nurse-to-patient ratio for patients of specialty $i \in \mathcal{S}$, and $r(\mathcal{A})$ be the daily cost of a nurse working in a ward admitting patients of a subset $\mathcal{A}$ of specialties. The dependence of the daily nursing cost on $\mathcal{A}$ is to reflect different training costs and salaries for different specialties. It then follows that

$$R_j((d^i; i \in \mathcal{C}_j), o_j) = \sum_{k \in \mathcal{C}_j} r(\{k\}) \left\lceil S^d_k((d^i; i \in \mathcal{C}_j), o_j) f^k \right\rceil + r(\mathcal{C}_j) \left\lceil \sum_{k \in \mathcal{C}_j} S^o_k((d^i; i \in \mathcal{C}_j), o_j) f^k \right\rceil, \tag{5}$$

where the first (second) sum calculates the expected daily nursing cost in the dedicated wards (overflow ward) in cluster $\mathcal{C}_j$. In Equation (5), the function S^d_k (S^o_k) gives the expected number of patients of specialty $k \in \mathcal{C}_j$ in their dedicated ward (overflow ward), and $\lceil x \rceil$ is the smallest integer larger than or equal to x.

CBM In this formulation, we follow a two-stage approach. Let $\boldsymbol{\mathcal{F}}$ be the set of all partitions of $\mathcal{S}$. In the first stage, for any given partition $\boldsymbol{\mathcal{C}} \in \boldsymbol{\mathcal{F}}$, the model

$$Y(\boldsymbol{\mathcal{C}}) = \min_{\mathbf{d},\mathbf{o}} \left\{ \sum_{j=1}^{m} Q_j((d_i; i \in \mathcal{C}_j), o_j) : \sum_{j=1}^{m} o_j + \sum_{i=1}^{n} d^i \le B, \mathbf{d} \in \mathbb{Z}^n, \text{ and } \mathbf{o} \in \mathbb{Z}^m \right\} \tag{6}$$

with Q_j defined in (4), minimizes the overall expected daily number of patients turned away. Let $(\mathbf{d}^{\boldsymbol{\mathcal{C}}}, \mathbf{o}^{\boldsymbol{\mathcal{C}}})$ denote an optimal solution of problem (6). In the second stage, given $(\mathbf{d}^{\boldsymbol{\mathcal{C}}}, \mathbf{o}^{\boldsymbol{\mathcal{C}}})$,

$$X = \min_{\boldsymbol{\mathcal{C}}} \left\{ Y(\boldsymbol{\mathcal{C}}) : H(\boldsymbol{\mathcal{C}}, \mathbf{d}^{\boldsymbol{\mathcal{C}}}, \mathbf{o}^{\boldsymbol{\mathcal{C}}}) \le \tau, \boldsymbol{\mathcal{C}} \in \boldsymbol{\mathcal{F}} \right\}, \tag{7}$$

identifies the partition with minimum expected daily blocking whose nursing cost $H(\boldsymbol{\mathcal{C}}, \mathbf{d}^{\boldsymbol{\mathcal{C}}}, \mathbf{o}^{\boldsymbol{\mathcal{C}}})$ is below a threshold $\tau > 0$. Note that $H(\boldsymbol{\mathcal{C}}, \mathbf{d}, \mathbf{o}) = \sum_{j=1}^{m} R_j((d^i; i \in \mathcal{C}_j), o_j)$ with R_j defined in (5). Given the independence of clusters, following the same logic as for TCM, we recast problem (6) as

$$Y(\boldsymbol{\mathcal{C}}) = \min_{\mathbf{b}} \left\{ \sum_{j=1}^{m} \varphi_j(\mathcal{C}_j, b_j) : \sum_{j=1}^{m} b_j \le B, \text{ and } \mathbf{b} \in \mathbb{Z}^m \right\}, \tag{8}$$

where $\mathbf{b} = (b_1, \ldots, b_m)$, and, given $\mathcal{C}_j$ and b_j,

$$\varphi_j(\mathcal{C}_j, b_j) = \min_{(d^i; i \in \mathcal{C}_j),\, o_j} \left\{ Q_j((d^i; i \in \mathcal{C}_j), o_j) : o_j + \sum_{i \in \mathcal{C}_j} d^i \le b_j, o_j \in \mathbb{Z}, \text{ and } d^i \in \mathbb{Z} \text{ for } i \in \mathcal{C}_j \right\}. \tag{9}$$

3.3. Performance Evaluation

In this section, we estimate the blocking probability $p_j^k((d_i; i \in \mathcal{C}_j), o_j)$ faced by patients of specialty $k \in \mathcal{C}_j$ in the cluster $\mathcal{C}_j$ with bed allocation $(d_i; i \in \mathcal{C}_j), o_j$. For each specialty, we assume that admission requests form a Poisson process, and that LOSs are independent and identically distributed (i.i.d.) as an exponential distribution. Following Best et al. (2015), we further assume that the mean LOS of patients in a ward is focus- and workload-dependent, so it could vary depending on the number of specialties served in the ward as well as the number of beds allocated to the ward. More specifically, we denote with $\nu^i(d, \mathcal{A})$ the mean LOS for patients of specialty $i \in \mathcal{A}$ cared for in a d-bed ward shared by a subset $\mathcal{A}$ of specialties.

An exact product-form solution is provided in Bekker et al. (2016) for evaluating blocking probabilities in a cluster. However, it involves finding the integer points of a bounded polyhedral region which is computationally expensive. Furthermore, it only works if the mean LOS in the overflow ward is the same as in the dedicated ward. As such, we need to resort to approximations. To start, consider a loss system with arrival process $A(t)$ with rate $\lambda > 0$, service time cumulative distribution function (CDF) $F(t)$ with mean $\nu > 0$, and $d \in \mathbb{Z}$ servers. Hayward's approximation estimates the blocking probability in this system as (see Fredericks 1980)

$$B(A(t), F(t), d) \approx B_H(a, z, d) = B_e(a/z, d/z), \tag{10}$$

where $a = \lambda\nu$ is the offered load, z is the 'peakedness' measure, and $B_e(a, d)$ is a continuous extension of the Erlang loss function, such as $B_e(a, d) = \left[a \int_0^\infty \exp(-at)(1+t)^d dt\right]^{-1}$ proposed in Jagerman (1974). The peakedness z is defined as the variance-to-mean ratio for the steady-state number of busy servers in an equivalent infinite-server system, i.e. with arrival stream $A(t)$ and service time CDF $F(t)$. For $A(t)$ a Poisson process, $z = 1$ and, therefore, Equation (10) becomes the Erlang blocking formula; thus, it is exact. For non-Poisson arrivals, Equation (10) implies that to approximate blocking probability, it suffices to characterize the arrival and service processes through the (a, z) pair.

Equation (10) can be used directly for estimating the blocking probability faced by patients in their corresponding dedicated wards. To find the blocking faced by patients overflowing from a dedicated ward to the overflow ward of a cluster, however, one needs to estimate the peakedness of the overflow stream. To illustrate, consider a primary group of d servers with exponentially distributed service times with mean ν facing Poisson arrivals with rate λ. Customers finding all d servers busy upon arrival overflow to a secondary group of servers, also with exponential service times. When mean service times in the secondary and primary groups are the same, the well-known result $1-aB_e(a,d)+a/(d+1+aB_e(a,d)-a)$ gives the exact value of peakedness for the overflow stream (see, for example, Cooper 1981, Equation 3.2 in Chapter 4). The same equation is often used in the literature, e.g. in Chevalier and Tabordon (2003), for approximating peakedness when mean service times in the secondary and primary groups are different. In Proposition 1, we derive an exact expression for peakedness for the more general case with different means.

Proposition 1. *The peakedness of traffic overflowing a d-server loss system with i.i.d. exponential service times with mean ν facing Poisson arrivals with rate λ, relative to an infinite server system with i.i.d. exponential service times with mean ν', is given by*

$$\xi(a,d,\rho)=1-\frac{aB_e(a,d)}{\rho}+\frac{a(a+\rho)\ {}_3F_1(\rho,1-d,a+\rho+1;a+\rho;-1/a)}{\rho(a+\rho+1)\ {}_3F_1(1-d,\rho+1,2+a+\rho;a+\rho+1;-1/a)}, \tag{11}$$

where $a=\lambda\nu$ is the offered load in the loss system, $\rho=\nu/\nu'$ is the mean service ratio, and ${}_pF_q(a_1,\dots,a_p;b_1,\dots,b_q;x)$ is a generalized hypergeometric function.

Using Proposition 1, Corollary 1 provides estimates for the blocking probability as well as mean numbers of patients, enabling us to calculate Q_j and R_j given in (4) and (5), respectively.

Corollary 1. *For a cluster $\mathcal{C}_j$ with bed allocation $(d^i;i\in\mathcal{C}_j),o_j$, Proposition 1 and Hayward's approximation imply that*

$$\begin{aligned}
p_j^k((d^i;i\in\mathcal{C}_j),o_j)&\approx B_e(a^k,d^k)B_e(\alpha_j/\beta_j,o_j/\beta_j),\\
S_k^d((d^i;i\in\mathcal{C}_j),o_j)&\approx a^k(1-B_e(a^k,d^k)),\\
S_k^o((d^i;i\in\mathcal{C}_j),o_j)&\approx a^kB_e(a^k,d^k)(1-B_e(\alpha_j/\beta_j,o_j/\beta_j))/\rho^k,
\end{aligned} \tag{12}$$

for $k \in \mathcal{C}_j$*, where* $a^k = \lambda^k \nu^k(d^k, \{k\})$ *and* $\rho^k = \nu^k(d^k, \{k\})/\nu^k(o_j, \mathcal{C}_j)$*. In* (12)*,* α_j *and* β_j *are the offered load and peakedness, respectively, for the aggregate traffic overflowing dedicated wards in cluster* $\mathcal{C}_j$*, and are obtained by*

$$\alpha_j = \sum_{i \in \mathcal{C}_j} \frac{a^i}{\rho^i} B_e(a^i, d^i), \qquad \beta_j = \frac{1}{\alpha_j} \sum_{i \in \mathcal{C}_j} \frac{a^i}{\rho^i} B_e(a^i, d^i) \xi(a^i, d^i, \rho^i). \tag{13}$$

4. Solving the Models

TCM This involves solving problems (2) and (3). We start with problem (3). The plot provided in the left panel in Figure EC.1 in the appendix (based on our hospital data) indicates that the objective function for this problem, T_j, is neither convex nor differentiable. Given this and also the fact that the objective function is expensive to evaluate (due to the appearance of hypergeometric and integral functions in loss probabilities), we opt for gradient-free optimization methods (Miguel and Nikolaos 2013). We focus on direct-search algorithms due to their simplicity and also their adaptation to constrained integer programming models (Conn et al. 2009). In particular, we consider Powell's conjugate direction method (Powell 1964), Brent's principal axis method (Brent 1973), a quadratic approximation method (Gill and Murray 1974), and the conjugate direct orthogonal shift (CDOS) method (Moissev 2011). We conducted large numbers of experiments with these four methods on our hospital data, and the CDOS method proved to be the fastest. We therefore use it for finding a solution to problem (3).

For problem (2), we note that it is similar to problem (P) proposed in Best et al. (2015, p. 162). The only difference is that $\phi_j(\mathcal{C}_j, b_j)$ inside of the sum in our problem is evaluated through another problem given in (3). As such, we adopt the same solution approach as follows. Given a fixed sequence $\mathcal{N}$ of specialties, firstly, we restrict the feasible region Ψ by considering only partitions obtained by making cuts along $\mathcal{N}$ to arrive at the following restricted problem

$$Z_{\mathcal{N}} = \min_{(m, \mathbf{b}, \mathcal{C})} \left\{ \sum_{j=1}^{m} \phi_j(\mathcal{C}_j, b_j) : (m, \mathbf{b}, \mathcal{C}) \in \{\Psi \text{ and } (\mathcal{C}_j \text{ are cuts in the sequence } \mathcal{N})\} \right\}. \tag{14}$$

Secondly, a dynamic programming (DP) approach as proposed in Best et al. (2015, p. 163) is used for solving problem (14). The only difference is that the expected reward for each state-action pair in our solution is evaluated using the CDOS heuristic explained above.

CBM We start with the first stage, i.e. problems (8) and (9). As illustrated in the right panel in Figure EC.1 in the appendix, the objective function for (9), Q_j, is not convex. As such, we use the CDOS heuristic for solving this problem, too. For problem (8), we adapt the integer linear programming approach proposed in Best et al. (2015, p. 165). In particular, we define the coefficient c_{jk} in their model as the marginal improvement gained in denied admissions by allocating the $k-$th bed to cluster j, i.e. $c_{jk} = \varphi_j(\mathcal{C}_j, k-1) - \varphi_j(\mathcal{C}_j, k)$, where $\varphi_j(\mathcal{C}_j, b_j)$ is obtained using the CDOS heuristic. For the second stage, i.e. problem (7), we reduce $\boldsymbol{\mathcal{F}}$ to the set of partitions obtained by cuts in the sequence $\mathcal{N}$, denoted by $\boldsymbol{\mathcal{F}}_{\mathcal{N}}$, to arrive at the following amended problem

$$X_{\mathcal{N}} = \min_{\boldsymbol{\mathcal{C}}} \left\{ Y(\boldsymbol{\mathcal{C}}) : H(\boldsymbol{\mathcal{C}}, \mathbf{d}^{\boldsymbol{\mathcal{C}}}, \mathbf{o}^{\boldsymbol{\mathcal{C}}}) \leq \tau, \boldsymbol{\mathcal{C}} \in \boldsymbol{\mathcal{F}}_{\mathcal{N}} \right\}. \tag{15}$$

Having obtained $Y(\boldsymbol{\mathcal{C}})$ for each $\boldsymbol{\mathcal{C}} \in \boldsymbol{\mathcal{F}}_{\mathcal{N}}$, a solution to problem (15) is obtained through enumeration.

5. Case Study

HUS is a public teaching hospital serving a population of around eight million people. The paediatric department of the hospital has 160 beds allocated to seven different wards, with each serving a particular medical specialty. All patients referred for admission to the department are issued a ticket which is archived by the department at the end of each day. An admittance file including admission and discharge times is created once a patient is admitted. We collected data from the department over a four-month period from 01/08/2014 to 01/12/2014. Our discussions with department clinicians and our observations during the data collection period indicate that the current configuration of the department is a dedicated one, wherein patients are turned away if a bed is not expected to become available in the primary ward of the patients within a few hours after an admission request. Given such short waiting time thresholds and the findings from the literature as reported in Section 1, the loss assumption A(ii) applied in our models would serve as a good approximation for the system performance. A summary of the data analysis is presented in Table 1. It shows that there is an imbalance between blocking and occupancy rates across different wards. This is a typical problem observed in many hospitals, as discussed in Green and Nguyen (2001).

Table 1 Summary of Data Analysis for HUS Data.

Specialty	Nephrology (NPH)	Nutrition (NTR)	Respiratory (RSP)	Oncology (ONC)	General (GEN)	Gastroenterology (GAS)	Cardiology (CRD)
Number of beds	33	12	32	39	24	4	16
Ward occupancy	83.0%	71.6%	55.7%	94.1%	36.8%	33.4%	54.9%
Average number of arrivals per day	5.230	2.361	3.172	3.787	1.861	0.582	1.459
Average number blocked per day	3.041(58.2%)	0.385 (16.3%)	0.0 (0.0%)	2.057 (54.3%)	0.057 (3.0%)	0.0 (0.0%)	0.0 (0.0%)
LOS mean	12.510	4.347	5.614	21.212	4.900	2.293	6.024

5.1. Applying Models

We set $B = 160$ beds, and $\mathcal{S} = \{1, \dots, 7\}$ corresponding to the order of specialties given in Table 1. For the mean LOS in a d-bed ward shared by specialties in $\mathcal{A} \subset \mathcal{S}$, we use the following function proposed in Best et al. (2015),

$$\nu^i(d, \mathcal{A}) = \left(1 - \frac{\Delta\left(1 - \frac{|\mathcal{A}|}{n}\right)}{1 + e^{-\beta\left(\sum_{i\in\mathcal{A}} \frac{\lambda^i m^i}{d} - \epsilon\right)}}\right) m^i, \tag{16}$$

where $|x|$ represents the cardinality of set x, and m^i is the mean nominal LOS for specialty i patients (excluding the impact of focus and workload). In Equation (16), Δ controls the impact of focus, while β and ϵ control the impact of workload as evaluated by $\sum_{i\in\mathcal{A}} \lambda^i m^i / d$ (see Figure 3 and the corresponding description in Best et al. 2015, p. 166). Since the current configuration of the department is dedicated, the mean LOS values given in Table 1 represent the highest level of focus. To obtain estimates for m^i, we solve Equation (16) for m^i given d^i and $\nu^i(d^i, \{i\})$ values in Table 1, assuming that $\Delta = 0.05$, $\epsilon = 0.9$, and $\beta = 20$. The value selected for Δ implies a maximum reduction of 4.28% in the LOS due to focus, and the values for ϵ and β are proposed in Best et al. 2015. The resulting values for m^i yield the nominal traffic intensity of $\pi = \sum_{i\in\mathcal{S}} \lambda^i m^i / 160 = 1.22$.

For nurse-to-patient ratios, f^i, following the guidelines provided by the Royal College of Nursing (Royal College of Nursing 2013), we consider a ratio of 0.3 for children younger than two and 0.25 for children older than two in all specialties, with the mix of children determined based on the current patient population. To estimate the daily cost $r(\mathcal{A})$ of a nurse working in a ward admitting patients of specialties in $\mathcal{A}$, we consider the cost of training for the corresponding specialties discounted on a daily basis plus the daily wage. For each specialty, we use our estimate of fees for continuous professional development courses plus the cost of temporary cover during the course of training as the initial cost of training. This is then discounted daily over a five-year period with a yearly discount rate of 3.0% (as suggested in Johns et al. 2003 for healthcare projects) in order to obtain the daily cross-training cost. For daily wages, we use the average daily wage of a registered nurse in the hospital plus a 10% additional payment for each additional specialty for which the nurse cares in order to represent the higher value of multi-skilled nurses to the hospital. The resulting daily cross-training costs vary from 0.0% (for GEN) to 4.2% (for NPH and ONC) of the average daily wage.

For the sequence $\mathcal{N}$ of specialties, we use the sequence obtained by sorting specialties in terms of their mean nominal LOS, m^i. Our numerical experiments suggest that such a sequence works well with both TCM and CBM formulations. It is also consistent with the pooling literature suggesting that merging customer classes is more likely to create improvements when LOS values are similar, e.g. van Dijk and van der Sluis (2008). Furthermore, our sequence follows the sequence proposed in Best et al. (2015) wherein specialties are sorted based on the ratio of utility to the mean nominal LOS. This is because in our formulations the utility of admitting a patient/the cost of turning one patient away is assumed to be the same for all specialties.

We apply our models to HUS data over a range of parameters, as follows. We set Δ to 0.0, 0.05 and 0.10, corresponding to maximum reductions of 0%, 4.8% and 8.57%, respectively, in the LOS for both TCM and CBM. For TCM, we set c to 10^* and 20^*, representing blocking costs of 10 and 20 times, respectively, larger than the daily wage of a registered nurse. For CBM, τ is set to $10\%^+$ and $20\%^+$, representing a 10 and 20 percent, respectively, permitted rise in nursing costs as compared to the corresponding dedicated configuration (which is the cheapest in terms of the cross-training cost). Finally, in order to investigate the impact of demand we consider $\pi = 0.8, 1.0, 1.4$ (in addition to the current value $\pi = 1.22$) by scaling the arrival rates with a constant factor.

To validate our loss performance evaluation model presented in Corollary 1, we compare the blocking probabilities obtained from our model for HUS data to those obtained from our sample under the current configuration. Our model yields blocking probabilities of 50.9% (NPH), 13.0%

(NTR), 0.0% (RSP), 52.0% (ONC), 0.0% (GEN), 3.5% (GAS) and 0.9% (CRD), which show a relatively good match with the sample results given in Table 1.

The best configurations obtained from our models and the corresponding metrics are presented in Tables 2 and 3 for TCM and CBM, respectively. These results show that the best found configuration under TCM is a clustered overflow configuration in all scenarios considered; it contains more than one cluster, and at least one cluster has both dedicated and overflow wards. Under CBM, on the other hand, the best found configuration is a wing formation configuration in all scenarios except for one (the scenario with $\pi = 0.8$ and $\tau = 20\%$) when $\Delta = 0.0$, and a clustered overflow configuration when $\Delta > 0.0$. This is because when $\Delta = 0.0$, a lower occupancy as a result of more patients being turned away reduces the nursing cost in the wing formation configuration, thereby making it more likely to meet the nursing budget than the clustered overflow configuration. When $\Delta > 0.0$, however, the reduction in LOSs as a result of focus enables the clustered overflow configuration to reduce the numbers turned away without much increase in occupancy, hence performing better than the wing formation configuration.

To understand the complex dynamics of various factors influencing our models' solution, consider the spectrum of configurations evaluated by our models. On one side of this spectrum, there is the fully flexible configuration which benefits from the lowest amount of slack capacity but suffers from the lowest degree of focus, the maximum level of mix variability, and the highest cost of cross-training. On the other side of the spectrum, there is the fully dedicated configuration which enjoys the highest degree of focus, the lowest cost of cross-training, and the minimum level of mix variability, while suffering from the largest amount of slack capacity. The clustered overflow configuration lies between these two extremes. As the number of clusters in this configuration increases, the overflow wards need to care for fewer specialties. Thus, they could potentially benefit from shorter LOSs, smaller cross-training costs, and lower mix variability, while having more slack capacity. On the other hand, allocating fewer (more) beds to dedicated (overflow) wards may counter the impacts of increased cluster formation outlined above (see the detailed example in

Table 2 **Results for TCM Formulation with Poisson Arrivals and Exponential LOSs.**

(Δ, c)	π	Total	Blocked	Nurse	Occ'y	$\boldsymbol{\mathcal{C}}$	$\mathbf{d}$	$\mathbf{o}$
		Cost		Cost	(%)			
$(0.0, 10^*)$	0.80	1652	0.244	1549	77	$\{\{6,2,5,3,7,1\},\{4\}\}$	$(48,9,15,55,9,0,5)$	$(17,0)$
	1.00	2262	1.406	1669	82	$\{\{6,2,5,3,7\},\{1\},\{4\}\}$	$(60,9,17,43,8,0,8)$	$(15,0,0)$
	1.22	3061	3.409	1622	80	$\{\{6,2,5,3\},\{7,1\},\{4\}\}$	$(59,14,23,15,13,0,10)$	$(9,17,0)$
	1.40	3527	4.493	1630	80	$\{\{6,2,5,3\},\{7,1\},\{4\}\}$	$(77,13,22,0,12,0,9)$	$(14,13,0)$
$(0.0, 20^*)$	0.80	1746	0.170	1602	79	$\{\{6,2,5,3\},\{7,1,4\}\}$	$(48,4,13,55,7,0,5)$	$(15,13)$
	1.00	2825	1.159	1847	86	$\{\{6,2,5,3,7,1\},\{4\}\}$	$(48,8,12,47,8,0,4)$	$(33,0)$
	1.22	4434	3.102	1815	84	$\{\{6,2,5,3,7,1\},\{4\}\}$	$(68,12,16,19,7,0,9)$	$(29,0)$
	1.40	5401	4.315	1759	83	$\{\{6,2,5,3,7\},\{1\},\{4\}\}$	$(83,12,16,7,10,0,9)$	$(23,0,0)$
$(0.05, 10^*)$	0.80	1600	0.262	1489	75	$\{\{6,2,5,3,7\},\{1,4\}\}$	$(42,13,15,49,11,0,5)$	$(13,11)$
	1.00	2169	1.143	1688	83	$\{\{6,2,5,3\},\{7,1\},\{4\}\}$	$(55,9,13,47,3,0,8)$	$(18,7,0)$
	1.22	2976	3.103	1666	82	$\{\{6,2,5,3,7\},\{1\},\{4\}\}$	$(70,14,17,23,12,0,10)$	$(14,0,0)$
	1.40	3470	4.435	1598	78	$\{\{6,2\},\{5,3,7,1\},\{4\}\}$	$(73,13,22,0,12,0,9)$	$(12,19,0)$
$(0.05, 20^*)$	0.80	1658	0.107	1568	76	$\{\{6,2,5,3,7\},\{1,4\}\}$	$(42,12,8,48,10,0,5)$	$(15,20)$
	1.00	2615	0.980	1788	86	$\{\{6,2,5,3,7\},\{1,4\}\}$	$(49,9,13,38,8,0,4)$	$(18,21)$
	1.22	4214	2.883	1780	85	$\{\{6,2,5,3\},\{7,1\},\{4\}\}$	$(60,8,17,30,0,0,9)$	$(25,11,0)$
	1.40	5210	4.006	1829	85	$\{\{6,2,5,3,7\},\{1\},\{4\}\}$	$(73,11,19,19,7,0,4)$	$(27,0,0)$
$(0.1, 10^*)$	0.80	1526	0.883	1489	74	$\{\{6,2,5,3,7\},\{1,4\}\}$	$(42,12,15,50,11,0,5)$	$(11,14)$
	1.00	2059	0.896	1680	83	$\{\{6,2,5\},\{3,7,1\},\{4\}\}$	$(49,9,13,51,8,0,4)$	$(7,19,0)$
	1.22	2848	2.730	1696	82	$\{\{6,2,5,3\},\{7,1\},\{4\}\}$	$(61,8,16,26,7,0,9)$	$(18,15,0)$
	1.40	3334	3.838	1714	83	$\{\{6,2,5,3\},\{7,1\},\{4\}\}$	$(69,8,16,15,7,0,10)$	$(21,14,0)$
$(0.1, 20^*)$	0.80	1543	0.331	1515	74	$\{\{6,2,5,3,7\},\{1,4\}\}$	$(43,9,14,51,4,0,5)$	$(16,18)$
	1.00	2427	0.717	1822	86	$\{\{6,2,5,3,7\},\{1,4\}\}$	$(49,8,13,45,6,0,4)$	$(17,18)$
	1.22	3983	2.646	1750	83	$\{\{6,2,5,3\},\{7,1\},\{4\}\}$	$(55,8,17,30,7,0,8)$	$(18,17,0)$
	1.40	4940	3.757	1768	84	$\{\{6,2,5,3\},\{7,1\},\{4\}\}$	$(63,8,16,19,7,0,9)$	$(21,17,0)$

Table 3 **Results for CBM formulation with Poisson Arrivals and Exponential LOSs.**

(Δ, τ)	π	Blocked	Nurse	Occ'y	$\boldsymbol{\mathcal{C}}$	$\mathbf{d}$	$\mathbf{o}$
			Cost	(%)			
$(0.0, 10\%^+)$	0.80	0.164	1758	78	$\{\{6,2,5,3,7\},\{1\},\{4\}\}$	$(0,0,0,0,0,0,0)$	$(43,55,62)$
	1.00	1.336	1852	86	$\{\{6,2\},\{5,3,7\},\{1\},\{4\}\}$	$(0,0,0,0,0,0,0)$	$(15,36,57,52)$
	1.22	3.288	1851	86	$\{\{6\},\{2\},\{5,3,7\},\{1\},\{4\}\}$	$(0,0,0,0,0,0,0)$	$(3,15,44,69,29)$
	1.40	4.483	1845	86	$\{\{6\},\{2,5\},\{3,7\},\{1\},\{4\}\}$	$(0,0,0,0,0,0,0)$	$(4,29,37,78,12)$
$(0.0, 20\%^+)$	0.80	0.098	1938	79	$\{\{6\},\{2,5,3,7\},\{1,4\}\}$	$(31,0,0,0,0,0,0)$	$(4,43,82)$
	1.00	1.141	1983	87	$\{\{6,2,5,3,7\},\{1\},\{4\}\}$	$(0,0,0,0,0,0,0)$	$(48,58,54)$
	1.22	3.066	2007	88	$\{\{6\},\{2,5,3,7\},\{1\},\{4\}\}$	$(0,0,0,0,0,0,0)$	$(3,56,69,32)$
	1.40	4.300	1999	86	$\{\{6,2,5,3\},\{7\},\{1\},\{4\}\}$	$(0,0,0,0,0,0,0)$	$(53,13,78,16)$
$(0.05, 10\%^+)$	0.80	0.131	1693	76	$\{\{6,2,5,3\},\{7,1\},\{4\}\}$	$(41,3,9,61,0,0,4)$	$(25,17,0)$
	1.00	1.057	1866	86	$\{\{6\},\{2,5,3,7\},\{1\},\{4\}\}$	$(55,2,7,56,0,3,0)$	$(0,37,0,0)$
	1.22	3.135	1865	85	$\{\{6,2\},\{5,3\},\{7\},\{1\},\{4\}\}$	$(66,8,8,32,0,0,12)$	$(9,25,0,0,0)$
	1.40	4.128	1867	86	$\{\{6,2,5\},\{3,7\},\{1\},\{4\}\}$	$(73,6,11,21,1,0,0)$	$(23,25,0,0)$
$(0.05, 20\%^+)$	0.80	0.061	1873	78	$\{\{6,2,5,3,7,1\},\{4\}\}$	$(40,0,0,65,0,0,0)$	$(55,0)$
	1.00	0.975	1977	87	$\{\{6,2,5,3,7\},\{1\},\{4\}\}$	$(55,2,7,57,0,0,0)$	$(39,0,0)$
	1.22	2.742	2025	88	$\{\{6,2,5,3,7\},\{1\},\{4\}\}$	$(66,3,7,37,0,0,0)$	$(47,0,0)$
	1.40	4.005	2030	87	$\{\{6\},\{2,5,3,7\},\{1\},\{4\}\}$	$(73,3,8,22,0,4,0)$	$(0,50,0,0)$
$(0.1, 10\%^+)$	0.80	0.088	1655	75	$\{\{6,2,5\},\{3,7,1\},\{4\}\}$	$(40,6,0,62,4,0,0)$	$(12,36,0)$
	1.00	0.909	1803	84	$\{\{6,2\},\{5,3,7\},\{1\},\{4\}\}$	$(54,7,8,57,1,0,0)$	$(7,26,0,0)$
	1.22	2.868	1807	85	$\{\{6,2\},\{5\},\{3,7\},\{1\},\{4\}\}$	$(63,7,10,37,13,0,0)$	$(9,0,21,0,0)$
	1.40	3.836	1955	86	$\{\{6,2,5\},\{3,7\},\{1\},\{4\}\}$	$(70,6,11,27,1,0,0)$	$(22,23,0,0)$
$(0.1, 20\%^+)$	0.80	0.014	1779	74	$\{\{6,2\},\{5,3,7,1,4\}\}$	$(38,7,4,48,0,0,0)$	$(8,55)$
	1.00	0.741	2030	86	$\{\{6,2\},\{5,3,7\},\{1,4\}\}$	$(39,7,8,0,1,0,0)$	$(7,26,72)$
	1.22	2.586	2034	88	$\{\{6\},\{2,5,3,7\},\{1\},\{4\}\}$	$(63,3,7,41,1,3,0)$	$(0,42,0,0)$
	1.40	3.740	2140	87	$\{\{6\},\{2,5,3,7\},\{1\},\{4\}\}$	$(70,4,8,27,1,4,0)$	$(0,46,0,0)$

Section EC.5 in the appendix). Due to these conflicting impacts of clustering and bed distribution, it is difficult to predict how the structure of our models' solution would change as a result of changes in the models' parameters. However, our numerical experiments with a wide range of these parameters suggest that the number of clusters decreases (increases) in general with c and τ (π).

5.2. Comparison with Other Configurations

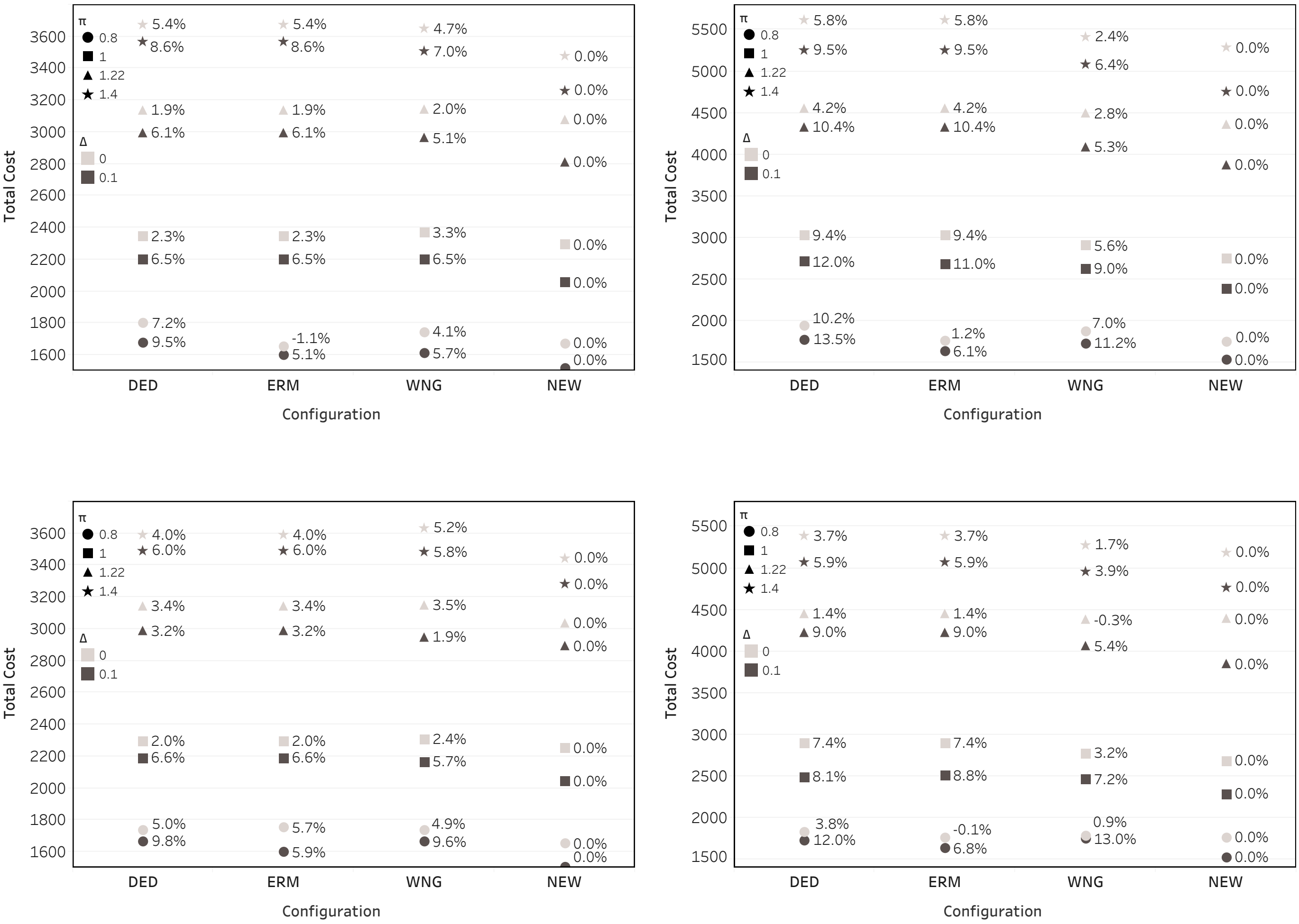
In this section, we compare the performance of configurations obtained from our models (NEW) with that of the dedicated (DED), earmarking (ERM), wing formation (WNG) and flexible (FLX) configurations. For DED (ERM), the bed allocation under TCM is found by restricting the set of actions in the DP model so that each specialty (all specialties) is (are) allocated to a cluster. Similarly, under CBM the bed allocation for DED (ERM) is obtained by solving problem (8) for $\boldsymbol{\mathcal{C}} = \{\{1\}, \ldots, \{7\}\}$ ($\boldsymbol{\mathcal{C}} = \{1, \ldots, 7\}$). For WNG, the clustering and bed allocation is found by solving problem (2) ((8)) with $\phi_j(\mathcal{C}_j, b_j) = T_j(\mathbf{0}, b_j)$ ($\varphi_j(\mathcal{C}_j, b_j) = Q_j(\mathbf{0}, b_j)$) for TCM (CBM). For FLX, the partition and bed allocation is fixed.

We compare different configurations using a simulation model instead of analytical models so that we can relax A(ii). In the simulation model, we assume that patients who find their relevant dedicated and overflow wards fully occupied upon arrival wait in dedicated queues corresponding to their specialties. Once a dedicated (overflow) bed becomes available, a patient is admitted from the corresponding specialty queue (the longest queue in the corresponding cluster). The longest queue policy is found to outperform other major policies in Jordan et al. (2004). Patients waiting in the queue are assumed to have a random willingness to wait distributed exponentially with mean q days. We simulate all five configurations with $q = 1$ day and $q = 7$ days. Although both of these thresholds are relatively short, $q = 1$ is more representative of the current department situation while $q = 7$ illustrates a less critical setting with a potentially larger workload of elective cases. We run 100 replications of the simulation model, each over 10 years, and estimate the expected wait for admitted patients in addition to the expected daily numbers abandoned and nursing cost.

Figures 2 and 3 represent the expected daily total cost and numbers abandoned for various configurations under TCM and CBM, respectively, estimated through simulation. The corresponding waiting time results are presented in Figures EC.2 and EC.3 in the appendix. We exclude the scenarios with $\Delta = 0.05$ in order to simplify illustrations, so a total of thirty two scenarios are presented for each configuration (sixteen for each value of q). The positive (negative) percentage figure next to each mark in the plots represents how much better (worse) NEW performs than the corresponding configuration. In the plots in Figure 3, there is a second figure next to each mark which shows the percentage by which the nursing cost constraint is violated under CBM. For TCM, the total cost of FLX turns out to be substantially larger than that of the other four configurations; thus, it has not been included in the plots in Figure 2. For CBM, both FLX and ERM are excluded from the plots in Figure 3, as their nursing costs go above the permitted thresholds in all scenarios.

The plots in Figure 2 reveal that under TCM, NEW has the lowest total cost in twenty nine out of thirty two scenarios, whereby resulting in potential improvements as large as 5.6% and 4.9% (9.0% and 7.2%) for $q = 1$ and $q = 7$, respectively, as compared to its closest rival when $\Delta = 0.0$ ($\Delta = 0.1$). In these twenty nine scenarios, NEW's closest rival is WNG in eighteen, ERM in five and DED in six scenarios. On the other hand, in the three scenarios in which NEW is not the lowest-cost configuration, its cost does not exceed the minimum cost by more than 1.1%. For CBM, the plots in Figure 3 reveal that NEW breaches the nursing cost target in four out of thirty two scenarios. The size of these breaches, however, does not exceed 2.2% in our experiments. The plots also show that NEW yields the smallest numbers abandoned in twenty six out of twenty eight scenarios in which the nursing cost constraint is met. On the other hand, WNG breaches the nursing budget in five scenarios, and yields the lowest numbers abandoned in twelve scenarios. (The overlap is because NEW and WNG are the same in seven out of eight scenarios with $\Delta = 0.0$, as discussed in Section 5.1.) The waiting time plots in Figures EC.2 and EC.3 in the appendix illustrate that under TCM (CBM), NEW has the shortest mean waiting time in nineteen (twenty eight) out of thirty two scenarios, with ERM (DED) typically providing the shortest wait in the remaining scenarios. Overall, the simulation experiments conducted in this section suggest that the configurations obtained with our models typically perform better than the other configurations proposed in the literature as long as patients' waiting time threshold is relatively short.

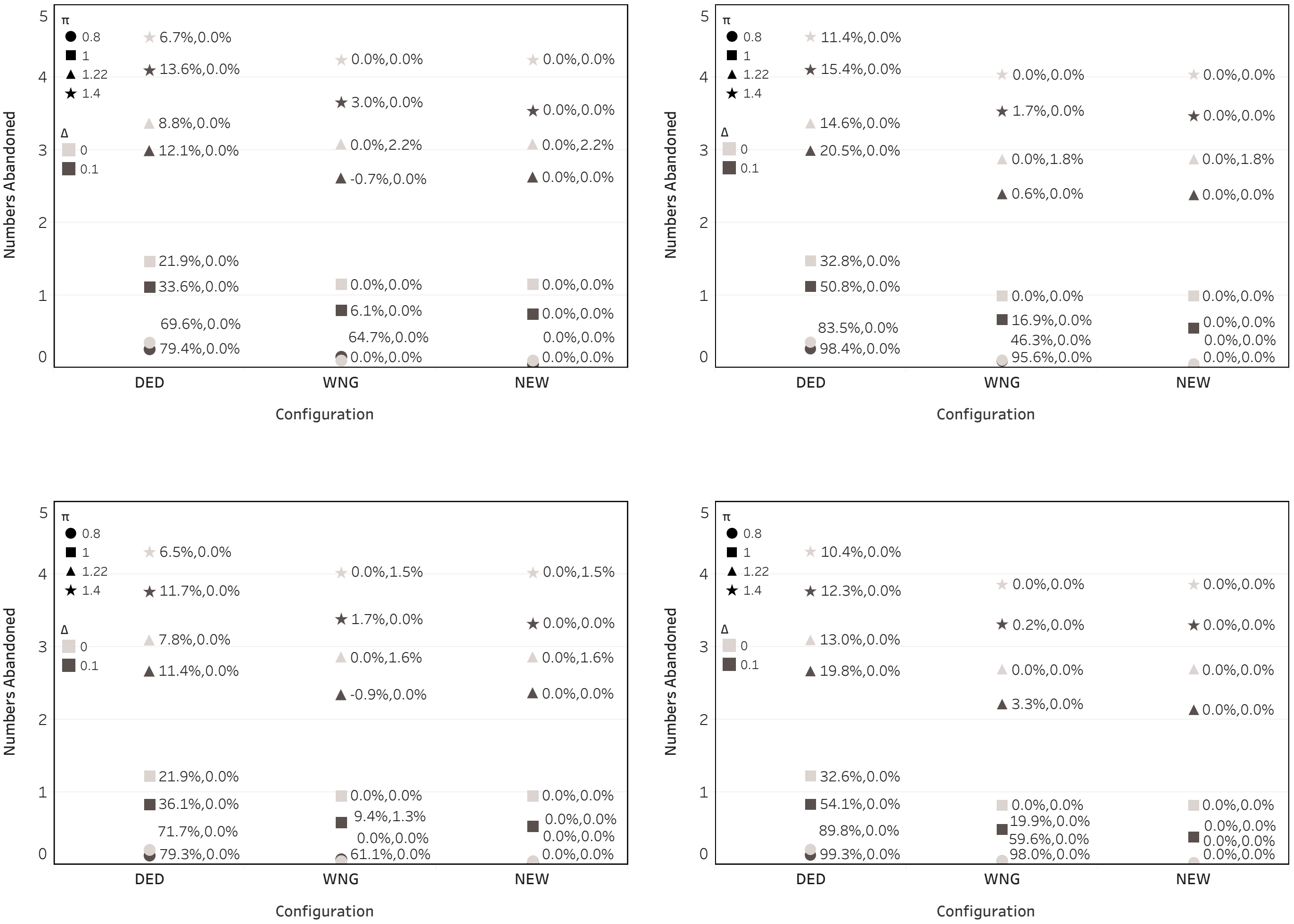
Figure 2 Expected Daily Total Cost for Different Configurations under TCM for $c = 10^*$ (left), $c = 20^*$ (right), $q = 1$ day (top) and $q = 7$ days (bottom) Obtained from Simulation.



6. Non-Poisson Admission Requests

A deeper analysis of HUS data as reported in Section EC.8 in the appendix suggests that daily admission numbers are not distributed as Poisson; they are over-dispersed for some specialties (i.e. their variance-to-mean ratios are larger than one, the theoretical ratio under Poisson), and under-dispersed for some others. The plots and statistics provided in Figure EC.4 in the appendix for LOS values also show that they are not exponentially distributed. In this section, we generalize our performance model in Section 3.3 to non-exponential inter-arrival and LOS distributions so that we can obtain the corresponding configurations for our case study data and compare them with

Figure 3 Expected Daily Numbers Abandoned for Different Configurations under CBM for $\tau = 10\%+$ (left) and $\tau = 20\%+$ (right), $q = 1$ day (top) and $q = 7$ days (bottom) Obtained from Simulation.



relevant results under the Poisson assumption. Firstly, note that for a loss system with a renewal arrival process and general service time distribution, the peakedness is evaluated by

$$z = 1 + \frac{(\kappa - 1)}{\nu} \mathbb{E}\left[\min\{L_1, L_2\}\right], \tag{17}$$

where κ is the squared coefficient of variation (SCV) of inter-arrival times, and L_1 and L_2 are independent random variables distributed as the service time (Li and Whitt 2014). Equation (17) can be applied in order to evaluate the peakedness of admission requests arriving in the dedicated wards. Secondly, in order to evaluate the peakedness of the traffic overflowing dedicated wards, we extend Proposition 1 to non-Poisson traffic in Proposition 2.

PROPOSITION 2. *The Fredericks (1980) approximation method implies that the peakedness of (a,z) traffic overflowing a d-server loss system with i.i.d. service times with mean ν, relative to an infinite server system with i.i.d. service times with mean ν', is estimated by*

$$\xi(a,z,d,\rho) \approx z - \frac{aB_e(a/z,d/z)}{\rho} + \frac{a(a+\rho z)\,{}_3F_1(\rho,1-d/z,a/z+\rho+1;a/z+\rho;-z/a)}{\rho(a+\rho z+z)\,{}_3F_1(1-d/z,\rho+1,2+a/z+\rho;a/z+\rho+1;-z/a)}, \tag{18}$$

where $\rho = \nu/\nu'$, and ${}_pF_q(a_1,\ldots,a_p;b_1,\ldots,b_q;x)$ is a generalized hypergeometric function.

Corollary 1 can now be generalized to systems with non-Poisson admission by replacing $B_e(a^i,d^i)$ with $B_e(a^i/z^i,d^i/z^i)$, as well as $\xi(a^i,d^i,\rho^i)$ with $\xi(a^i,z^i,d^i,\rho^i)$, where z^i is the peakedness for specialty i arrivals. The same solution methodologies as those explained in Section 4 can then be applied for finding the configurations under the TCM and CBM formulations.

For specialties RSP, GAS and CRD, for which (according to Table 1) blocking rates obtained from our sample are zero, we use Equation (17) to evaluate peakedness. In this equation, κ is estimated empirically from inter-arrival times, and LOSs are assumed to follow the empirical distributions obtained from our sample. For the remaining specialties with positive blocking rates, however, we cannot use the SCV obtained from the sample, since arrival times are only available for admitted patients. We follow two different methods in order to estimate peakedness for these specialties. In Method 1, inter-arrival times corresponding to periods with full ward occupancy are eliminated from our sample. This is because blocked admissions might have occurred during these intervals. In Method 2, for each specialty (with a positive blocking rate), we solve Hayward's equation in (10) for z, given the existing bed number, offered load and blocking probability for that specialty. This is because the current configuration is a dedicated one with a short waiting time threshold, so Hayward's approximation applied with the appropriate z value to each ward should produce a blocking probability close to the value obtained from our sample. The peakedness values obtained from both methods are presented in Table 4. This table shows that Method 2 produces larger peakedness values than Method 1. To test and compare the accuracy of these methods, we have included in Table 4 the blocking probabilities obtained with the estimated peakedness values using

Table 4 Peakedness for Different Specialties.

Specialty	NPH	NTR	RSP	ONC	GEN	GAS	CRD
Blocking Prob (Data)	58.2%	16.3%	0.0%	54.3%	3.0%	0.0%	0.0%
Peakedness (Method 1)	1.296	1.312	1.482	1.036	1.437	0.712	1.062
Blocking Prob (Sim)	51.8%	14.0%	0.0%	52.5%	0.1%	0.0%	0.0%
Peakedness (Method 2)	9.922	1.409	1.482	3.033	4.771	0.712	1.062
Blocking Prob (Sim)	56.2%	15.0%	0.0%	52.5%	0.3%	0.0%	0.0%

simulation. In the simulation, the LOS is sampled from the corresponding empirical distribution, and inter-arrival times are assumed to follow a log-normal distribution with parameters set so that the resulting arrival rate and peakedness match the estimated figures. The results imply that i) both methods produce more accurate blocking probabilities than those obtained with the Poisson assumption as presented in Section 5.1, and ii) Method 2 produces blocking probabilities closer to the sample results than Method 1.

We apply our generalized models to HUS data with peakedness values obtained from both methods explained above. In the generalized models for CBM, we set τ equal to the nursing cost obtained with the Poisson models so that we can have a fair comparison. The best configurations obtained from the generalized models for $\pi = 1.22$ (the current traffic) are presented in Tables EC.4 and EC.5 in the appendix. The first observation from these results is that the number of overflow beds tends to increase and the number of clusters tends to decrease as peakedness increases, suggesting that pooling becomes more beneficial when the arrival process is more variable. The second observation is that under TCM the overall occupancy may drop to as low as 66% when arrivals are highly over-dispersed, whereas with CBM it tends to be more stable. Finally, we observe that the configurations obtained with the generalized models are different from those obtained under the Poisson assumption. Given the substantial increase in computation time with the generalized models as illustrated in Table EC.6 in the appendix, it would be useful to estimate the added value of using these models. To do so, we estimate the average total cost and total patients turned away for the configurations obtained with the Poisson assumption using the performance evaluation model based on Proposition 2 as explained above, and compare them with the corresponding figures for the configurations obtained with the generalized models. The results are presented in

Table 5 **Relative Difference in Performance of Best Found Configurations with Generalized Models as Compared to Those with Poisson Models.**

Scenario	Total Cost (TCM)		Scenario	Blocked (CBM)	
	Method 1	Method 2		Method 1	Method 2
$\Delta = 0.0, c = 10^*$	-2.8%	-5.5%	$\Delta = 0.0, \tau = 1851$	-0.2%	-6.8%
$\Delta = 0.0, c = 20^*$	-2.2%	-9.2%	$\Delta = 0.0, \tau = 2007$	-0.1%	-3.3%
$\Delta = 0.1, c = 10^*$	-4.8%	-6.9%	$\Delta = 0.1, \tau = 1807$	-0.4%	-11.6%
$\Delta = 0.1, c = 20^*$	-1.7%	-4.8%	$\Delta = 0.1, \tau = 2034$	-0.4%	-8.4%

Table 5. They suggest that with CBM, the improvements obtained with the generalized models are likely to be large when over-dispersion is high. With TCM, on the other hand, the improvements could be substantial even with small over-dispersion.

7. Conclusion

The literature on flexibility suggests that the choice between specialization and pooling is not an 'all or nothing' proposition, but often an intermediate configuration with some resource sharing (Bassamboo et al. 2010). Building upon this literature, as well as on bed planning and overflow analysis, we proposed a partially pooled configuration that matches the requirements of inpatient care. Not only does this configuration provide the advantages of focus in its dedicated wards, it also leads to a better performance by pooling beds' idleness in overflow wards while keeping the mix variability and cross-training costs to a minimum. We proposed two different formulations for partitioning and bed allocation in the proposed configuration with and without the Poisson arrival assumption, and illustrated their applications using real data. Our simulation experiments suggested that the configurations obtained from our models under the zero-waiting-time assumption work well when patients' waiting time threshold is relatively short. Developing models capturing the impact of long waiting time thresholds, e.g. longer than seven days, for the clustered overflow configuration would be an area for future research.

The proposed configuration and formulations have been well received by the practitioners at HUS. They found the proposed configuration to be a viable solution to their bed shortage problem, provided that additional clinical, space, location and privacy constraints are taken into account in the formation of clusters. Such constraints can be easily incorporated into our models, e.g. by restricting the action set in the TCM formulation or the set of partitions $\mathcal{F}$ in the CBM formulation, and would, in fact, result in a substantial reduction in computation time. Appropriate estimation of the models' parameters is also an important step for successful implementation. The most challenging parameter to estimate is the cost of turning patients away, as required by the TCM formulation. In private hospitals, this cost could be estimated as the average loss of earnings associated with an admission refusal. In public systems, there are other cost elements that may need to be considered at a more strategic level, such as the cost of delay in treatment or patient transportation to another hospital. In any case, a range of values should be tried in the models before a specific choice is made. On the other hand, CBM does not require this cost estimate and, therefore, may be preferred over TCM. Nurses' cross-training cost is another element that may be difficult to estimate, especially if cross-training is provided on-the-job under the supervision of a senior nurse. A fraction of the cost of the senior nurse in this case could be used as an estimate for cross-training cost. Given these considerations, a successful execution of the clustered overflow configuration could relieve the pressure on inpatient beds.

Acknowledgments

References

Aiken, L. H., S. P. Clarke, D. M. Sloane, J. Sochalski, J. H. Silber. 2002. Hospital Nurse Staffing and Patient Mortality, Nurse Burnout, and Job Dissatisfaction. *J. Am. Med. Assoc* **288**(16) 1987–93.

Bassamboo, A., R. S. Randhawa, J. A. Van Mieghem. 2010. Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing. *Management Sci.* **56**(8) 1285–1303.

Bekker, R., G. Koole, D. Roubos. 2016. Flexible Bed Allocations for Hospital Wards. *Health Care Manag. Sci.* 1–14.

Belciug, S., F. Gorunescu. 2015. Improving Hospital Bed Occupancy and Resource Utilization through Queuing Modeling and Evolutionary Computation. *J. Biomed. Inform.* **53** 261–9.

Best, T. J., B. Sandıkçı, D. D. Eisenstein, D. O. Meltzer. 2015. Managing Hospital Inpatient Bed Capacity Through Partitioning Care into Focused Wings. *Manufacturing Service Oper. Management* **17**(2) 157–176.

Brent, R. P. 1973. *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

Chevalier, P., R. Shumsky, N. Tabordon. 2005. Routing and Staffing in Large Call Centers with Specialized and Fully Flexible Servers. Technical report, Universit catholique de Louvain, Louvain, Belgium.

Chevalier, P., N. Tabordon. 2003. Overflow Analysis and Cross-Trained Servers. *Intern. J. Production Econ.* **85**(1) 47–60.

Chevalier, P., J. Van den Schrieck. 2008. Optimizing the Staffing and Routing of Small-Size Hierarchical Call Centers. *Production Oper. Management* **17**(3) 306–319.

Clark, J. R., R. S. Huckman. 2012. Broadening Focus: Spillovers, Complementarities, and Specialization in the Hospital Industry. *Management Sci.* **58**(4) 708–722.

Conn, A. R., K. Scheinberg, L. N. Vicente. 2009. *Introduction to Derivative-Free Optimzation*. SIAM.

Cooper, R. 1981. *Introduction to Queueing Theory*. 2nd ed. Elsevier.

de Bruin, A. M., R. Bekker, L. van Zanten, G. M. Koole. 2009. Dimensioning Hospital Wards Using the Erlang Loss Model. *Ann. Oper. Res* **178**(1) 23–43.

Franx, G. J., G. M. Koole, A. Pot. 2006. Approximating Multi-Skill Blocking Systems by HyperExponential Decomposition. *Performance Evaluation* **63**(8) 799–824.

Fredericks, A. A. 1980. Congestion in Blocking Systems-A Simple Approximation Technique. *Bell Syst. Tech. J.* **59**(6) 805–827.

Gill, P. E., W. Murray, eds. 1974. *Numerical methods for constrained optimization*. Academic Press.

Green, L. V. 2004. *Operations Research and Health Care: A Handbook of Methods and Applications*, chap. Capacity Planning and Management in Hospitals. Springer US, Boston, MA, 15–41.

Green, L. V., V. Nguyen. 2001. Strategies for Cutting Hospital Beds: The Impact on Patient Service. *Health Services Research* **36**(2) 421–442.

Hall, R. 2012. *Handbook of Healthcare System Engineering*, chap. Bed Assignment and Bed Management. Springer Science+Business Media, Boston, MA, 177–200.

Huang, X. 1998. Decision Making Support in Reshaping Hospital Medical Services. *Health Care Manag. Sci.* **1**(2) 165–173.

Jagerman, D. L. 1974. Some Properties of the Erlang Loss Function. *Bell Syst. Tech. J.* **53**(3) 525–551.

Johns, B., R. Baltussen, R. Hutubessy. 2003. Cost Effectiveness and Resource Programme costs in the economic evaluation of health interventions. *Cost Eff. Resour. Alloc.* **1** 1–10.

Jordan, W. C., S. C. Graves. 1995. Principles on the Benefits of Manufacturing Process Flexibility. *Management Sci.* **41**(4) 577–594.

Jordan, W. C., R. R. In, D. E. Blumenfeld. 2004. Chained cross-training of workers for robust performance. *IIE Transactions* **36**(10) 953–967.

Kane, R. L., T. A. Shamilyan, C. Mueller, S. Duval, T. J. Wilt. 2007. The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis. *Medical Care* **45**(12) 1195–1204.

Kazahaya, G. 2005. Harnessing Technology to Redesign Labor Cost Management Reports. *Healthcare Financial Management : Journal of the Healthcare Financial Management Association* **59**(4) 94–100.

KC, D. S., C. Terwiesch. 2011. The Effects of Focus on Performance: Evidence from California Hospitals. *Management Sci.* **57**(11) 1897–1912.

Kokangul, A. 2008. A Combination of Deterministic and Stochastic Approaches to Optimize Bed Capacity in a Hospital Unit. *Comput. Methods Programs Biomed.* **90**(1) 56–65.

Koole, G., A. Pot. 2006. An overview of Routing and Staffing Algorithms in Multi-Skill Customer Contact Centers. *Working Paper* **VU University Amsterdam.** Available Online at http://www.math.vu.nl/.

Lang, T. A., M. Hodge, V. Olson, P. S. Romano, R. L. Kravitz. 2004. Nurse-patient ratios: A systematic review on the effects of nurse staffing on patient, nurse employee, and hospital outcomes. *Journal of Nursing Administration* **34**(7-8) 326–337.

Li, A. A, W. Whitt. 2014. Approximate Blocking Probabilities in Loss Models with Independence and Distribution Assumptions Relaxed. *Performance Evaluation* **80** 82–101.

Lloyd, J. M., S. Elsayed, A. Majeed, S. Kadambande, D. Lewis, R. Mothukuri, R. Kulkarni. 2005. The Practice of Outlying Patients is Dangerous: A Multicentre Comparison Study of Nursing Care Provided for Trauma Patients. *Injury* **36**(6) 710–3.

Ma, G., E. Demeulemeester. 2013. A Multilevel Integrative Approach to Hospital Case Mix and Capacity Planning. *Comp. Oper. Res.* **40**(9) 2198–2207.

Miguel, L., R. Nikolaos. 2013. Derivative-Free Optimization : A Review of Algorithms and Comparison of Software Implementations. *J. Global Optim.* **56** 1247–1293.

Moissev, S. N. 2011. Universal derivative-free optimization method with quadratic convergence.

National Center for Health Statistics. 2016. Health, United Stated, 2016. URL `http://www.cdc.gov/nchs/data/hus/hus16.pdf`.

NHS England. 2018. Bed Availability and Occupancy. URL `https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/`.

Olshaker, J. S., N. K. Rathlev. 2006. Emergency Department Overcrowding and Ambulance Diversion: The Impact and Potential Solutions of Extended Boarding of Admitted Patients in the Emergency Department. *J. Emerg. Med.* **30**(3) 351–6.

Powell, M. J. D. 1964. An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives. *Comp. J.* **7**(2) 155–162.

Royal College of Nursing. 2013. Defining Staffing Levels for Children and Young Peoples Services. URL `https://www.rcn.org.uk/professional-development/publications/pub-002172`.

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.* 1–28.

Stowell, A., P. Claret, Mu. Sebbane, X. Bobbia, C. Boyard, R. Genre Grandpierre, A. Moreau, J. de La Coussaye. 2013. Hospital Outlying through Lack of Beds and its Impact on Care and Patient Outcome. *Scand. J. Trauma Resusc. Emerg. Med.* **21** 17.

Tabordon, N. 2002. Modeling and Optimizing the Management of Operator Training in a Call Center. Ph.D. thesis, Institut DAdministration et de Gestion, Universite Catholique de Louvain, Belgium.

van Dijk, N. M., E. van der Sluis. 2008. To Pool or Not to Pool in Call Centers. *Production Oper. Management* **17**(3) 296–305.

World Health Organization. 2014. World Health Statistics. URL `http://apps.who.int/iris/bitstream/10665/112738/1/9789240692671.pdf`.

Yankovic, N., L. V. Green. 2011. Identifying Good Nursing Levels: A Queuing Approach. *Oper. Res.* **59**(4) 942–955.