Reconfiguration of Inpatient Services to Reduce Bed Pressure in Hospitals

Navid Izady[a,*], Bahar Arabzadeh[a], Nicholas Sands[b], James Adams[b]

[a] *Bayes Business School (formerly Cass), City, University of London, London, UK*
[b] *Royal Surrey County Hospital NHS Foundation Trust, Surrey, UK*

Abstract

Healthcare systems around the world are facing an inpatient bed crisis. This was highlighted more than ever during the recent COVID-19 pandemic. The consequences of bed shortage are substantial for both patients and staff. Finding innovative ways to improve the utilization of the existing bed base is therefore of significant importance. We focus on reconfiguration of inpatient services as a cost-effective solution to bed pressure in hospitals, and propose a comprehensive methodology for finding a low-cost configuration given a total number of beds, a set of specialties, and a finite or infinite waiting time threshold for patients. This involves developing novel approximations for performance evaluation of overflow delay and abandonment systems, and embedding them within heuristic search algorithms. We apply our reconfiguration methodology on inpatient data from a large UK hospital. Simulation experiments show that the configurations proposed by our methodology can result in significant savings compared to the existing configuration, and that a clustered overflow configuration is likely to produce the best results in many scenarios.

Keywords: OR in health services, Inpatient services, Bed configuration, Optimal allocation

1. Introduction

Now, more than ever before, healthcare systems around the world are facing an inpatient bed crisis. As an example, the statistics published by the National Health Service (NHS) in the UK show that in 2019, the average bed occupancy across the country was above 90% for the third year in a row (Ewbank et al., 2020). The shortage of beds was highlighted even further during the recent COVID-19 pandemic. For example, Mateen et al. (2021) report that during the first wave of the pandemic in England, many hospitals operated above safe-occupancy thresholds for significant periods. Moreover, the post-pandemic surge in demand for elective procedures has added the pressure on hospital beds (Propper et al., 2020).

*Corresponding author
Email address: navid.izady@city.city.ac.uk (Navid Izady)

Shortage of inpatient beds has many ramifications for both patients and staff. It prolongs the trolley wait, i.e., the time between a decision being made in the emergency department (ED) to admit a patient and admission to inpatient care. Trolley waits lower the quality of patient care and may result in patient fatalities. For example, a recent study of more than 5 million patients in England shows a linear increase in all-cause 30-day mortality as the admission delay increases from 5 hours after arrival to the ED up to 12 hours (Jones et al., 2022).

Trolley waits also create large backlogs in emergency departments. Congestion in emergency departments is linked to higher morbidity rates and may also lead to *ambulance diversion*; see, e.g., Olshaker & Rathlev (2006). Further, overcrowded hospitals are exposed to a higher risk of hospital acquired infections (Kaier et al., 2012). Not only does this put patients' health at risk, but also prolongs length of stay (LOS) and may also result in bed and ward closures (Goldstein et al., 2017), exacerbating the bed shortage problem. Shortage of beds for post-operative care may result in cancellation of medical procedures. Patients may also be discharged pre-maturely, only to be re-admitted later with potentially worse conditions (Maguire, 2015). Readmitted patients are reported to cost the UK's NHS £2.6 billion each year (Conroy & Dowsing, 2012).

Patient outlying, i.e., admitting patients to clinically inappropriate wards, is also a common phenomenon in hospitals which operate with high occupancy levels. Stowell et al. (2013) report a significant increase in the LOS and rate of mortality of outlied patients. In a recent study, Lim et al. (2021) show that the level and volatility of outlying will increase the LOS and the likelihood of readmission for non-outlied patients as well. Finally, the pressures that bed shortages create can have a damaging impact on staff morale and retention, which in turn impacts negatively on patient care (Robertson et al., 2017).

Expanding the bed base of a hospital is the most straight-forward solution to bed shortages. However, it is only considered as a last resort as adding a hospital bed with necessary equipment and staffing it are quite costly (Akcali et al., 2006). The focus of hospitals is therefore typically on more efficient use of existing resources. In this research, we focus on the reconfiguration of inpatient services as a cost-effective way for dealing with bed shortages. To understand the role of reconfiguration, it is important to note that inpatient care in general hospitals is typically delivered through a number of clinical units or *wards*. Each ward is equipped with a specific number of beds and a dedicated nursing team, providing care for specific medical specialties. The configuration of inpatient services identifies how beds and specialties are allocated to each ward, significantly impacting the performance and quality of inpatient care.

We identify five major configurations as illustrated in the spectrum in Figure 1. We compare these configurations in terms of their level of "focused care", the amount of "slack capacity"

and "mix variability", and "cross-training" costs. Focus in a hospital setting means delivering care for a limited set of conditions. It lowers uncertainty, reduces complexity, and gives the clinical staff the opportunity to develop specialised expertise (Clark & Huckman, 2012). Several studies suggest that focused care improves the quality of care, in particular, it reduces mean LOS (Best et al., 2015), rate of mortality (Clark & Huckman, 2012), and risk of readmission (KC & Terwiesch, 2011). Slack capacity refers to the number of available but unused beds in the hospital. It increases as the wards become more specialized. We define mix variability as the potential increase in variability of LOS caused by mixing specialties with different LOS distributions in a single ward. Queueing theory suggests that a higher variability in LOS typically leads to longer waiting times in the queue (Gross et al., 2011). Cross-training is about nurses being trained to deliver care for the specialties treated in their ward. Clearly, the cost of cross-training increases as more specialties are allocated to a ward.

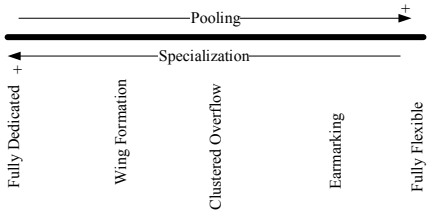


Figure 1: The spectrum of inpatient bed configurations

Starting from the far left of the spectrum in Figure 1, in a fully dedicated configuration each specialty is allocated to a single ward, which is called the primary ward of that specialty and is fed by a single queue. This configuration benefits from the maximum level of focused care and the minimum amount of mix variability as a limited number of conditions are treated in each ward. Slack capacity will, however, be at its highest level as there will be situations wherein patients are waiting for admission to their primary ward whilst beds are available in other wards.

The exact opposite of the fully dedicated configuration is the fully flexible configuration on the far right of the spectrum in Figure 1. It acts like a super ward, admitting patients of different specialties (who join a single queue) as long as a bed is available. This configuration enjoys the benefits of pooling, i.e., the slack capacity will be minimized. However, mix variability will increase as a result of mixing patients with different LOS distributions. This, along with the potential increase in mean LOS due to losing focus, may off-set the advantages of pooling. The fully flexible configuration also requires full cross-training of nursing teams which is expensive.

The other three configurations are placed in the middle of the spectrum in Figure 1 as they attempt to strike a balance between pooling and specialization. The wing formation configuration, proposed in Best et al. (2015), achieves this by partitioning the specialities into a number of

clusters and dedicating each cluster to a ward (wing). Each ward is fed by a single queue including all the patients requiring care for specialties in the corresponding cluster. Its cross-training cost, the amount of slack capacity and mix variability, and the level of focus vary depending on how specialties are partitioned.

In the earmarking configuration, introduced in Bekker et al. (2016), each speciality has a ward dedicated to it which is linked to a single queue, similar to the fully dedicated configuration. But there also exists a shared overflow ward admitting patients whose dedicated wards are full. The earmarking configuration benefits from focused care in its dedicated wards, and from pooling in its overflow ward. Since all specialities share one overflow ward, however, it requires costly full nurse cross-training. Both earmarking and wing formation configurations capture the fully dedicated and flexible configurations as special cases.

In the clustered overflow (COF) configuration, proposed in Izady & Mohamed (2021), specialities are partitioned into a number of clusters similar to the wing formation configuration. Each cluster, however, includes a dedicated ward for each of the specialities in the cluster as well a single overflow ward shared among all the specialities of the cluster. There is a single queue attached to each dedicated ward, feeding the patients first into the dedicated ward and next to the cluster overflow ward. Similar to earmarking, this configuration benefits from both pooling and focused care. The flexibility created by partitioning of specialties, however, helps reduce the cross-training cost as well as the amount of mix variability. The clustered overflow configuration captures all the other four configurations as special cases.

One of the five major configurations described above is typically adopted by hospitals. However, there always exists a degree of patient outlying in practice. While this reduces the admission wait for the patients involved, it may negatively influence the care quality of outlied and non-outlied patients as explained earlier, and may also worsen the performance of the system overall. Further, a large percentage of patient outlying is an indicator that inpatient services are not organized properly.

We aim to avoid patient outlying by choosing the right configuration of inpatient services. To achieve this, we develop a methodology that seeks to find the optimal configuration of inpatient beds in a hospital given a set of specialties, a total number of beds, and a waiting time threshold. The waiting time threshold identifies the extent to which patients wait in their queues before being transferred to another hospital. It depends on the urgency of patient condition, with emergency patients having a much shorter threshold than electives. It also varies in different healthcare systems. In Dutch hospitals, for example, the threshold is very short (Bekker et al., 2016), while in the UK it is very long with many patients waiting for several months before

admission.

We define the optimal configuration as the one minimizing the expected daily costs, including the cost of patients waiting in the queue or abandoning the queue plus the cost of nursing teams. We work with the COF configuration because it captures the other configurations as special cases. For this configuration, the specialties must be partitioned into a number of clusters, and the number of dedicated and overflow beds of each cluster must be identified. We design a heuristic methodology for this purpose, drawing on the intra-cluster bed allocation model proposed in Izady & Mohamed (2021) and the partitioning and inter-cluster bed allocation model proposed in Best et al. (2015). For the intra-cluster allocation model, we propose two novel approximation methodologies for estimating the performance metrics of a given cluster with specific bed allocation. One approximation applies to systems with an infinite waiting time threshold (IWTT), and the other works with systems with a finite waiting time threshold (FWTT). The simulation experiments indicate reliable accuracy of both approximations.

We investigate the application of our reconfiguration methodology using real inpatient data from the Royal Surrey County Hospital (RSCH) in the UK. A comprehensive data analysis shows that RSCH is operating a wing formation configuration with overlapping clusters under a high level of daily occupancy. There also exists a substantial level of patient outlying, accounting for about 30% of the total workload in the hospital. This level of patient outlying, as indicated by our analysis, prolongs the LOS of all patients in the hospital, including the non-outlying ones. There also exists a large number of ward changes within each specialty, which in addition to inconvenience for patients, results in longer LOSs in particular for older patients.

Simulation results suggest that implementing the configurations recommended by our methodology can create significant cost savings when compared to the existing configuration. Furthermore, our findings reveal that in the majority of scenarios considered, the COF configuration consistently emerges as the most cost-effective option, followed by the wing formation (earmarking) configuration when the impact of focus is small (large).

2. Literature Review

We start by reviewing the inpatient bed allocation literature in Section 2.1. We then explain the critical role played by performance evaluation models in inpatient bed allocation methodologies and explore the relevant literature in Section 2.2. This is followed by our contributions in Section 2.3.

2.1. Bed Allocation

Inpatient bed allocation has been studied by many researchers throughout the years. It means "...assigning beds to various patient categories according to medical specialty, accommodation type, and logistical considerations; presumably, patient needs, research goals, and educational requirements are taken into account, along with cost..." (Dumas, 1985, p. 44). We divide the relevant literature into three main streams as follows.

The first stream includes articles that seek to identify the number of beds for a clinical unit so as to achieve a given objective; see, for example, de Bruin et al. (2009). The second stream includes studies that seek to find the optimal allocation of a given number of beds to a set of specialties under a fully dedicated configuration; see, for example, Li et al. (2009). A comprehensive review of the first and second streams is provided in Arabzadeh (2022), Chapter 2.

The third stream is the closest to our research, and includes articles that consider clustering a given set of specialties and identifying the corresponding bed allocation simultaneously. The two prominent papers in this stream are those of Best et al. (2015) and Izady & Mohamed (2021). Best et al. (2015) propose the wing formation configuration, aiming to find the optimal level of bed pooling. Given a total number of beds, a set of specialties, and a finite waiting time threshold, they propose a methodology for finding the optimal allocation of specialties to different wings and the corresponding bed allocation such that the total utility to the hospital is maximised. Best et al. (2015) apply their model to inpatient data from an 18-specialty hospital. Their numerical experiments with varying levels of waiting time threshold and workload suggest that hospitals with a longer waiting time threshold or higher levels of demand should form more specialized wings to benefit from the advantages of focused care.

Izady & Mohamed (2021) introduce the COF configuration and propose a heuristic methodology for its clustering and bed allocation. Assuming a zero waiting time threshold (ZWTT), they propose two different formulations, a total cost minimization and a constrained blocking minimization. The former aims to minimize the total average daily cost, including the cost of turning patients away plus the cost of nursing teams, whereas the latter seeks to minimize the total number of patients turned away subject to nursing cost falling below a given threshold. The solution methodology in Izady & Mohamed (2021) involves an intra-cluster bed allocation model and a partitioning and inter-cluster bed allocation model. They apply their methodology to the data from a 7-specialty paediatric department, and report that the configurations obtained from their methodology compare very well with other major configurations as long as patients' waiting time threshold is relatively short.

2.2. Performance Evaluation

A performance evaluation model lies at the heart of all clustering and bed allocation methodologies. It evaluates the performance metrics of a given partition of specialties with a given allocation of beds. We categorize these models based on two dimensions: (i) the type of interaction between different wards: no interaction, a hierarchical interaction, or a cross-facility interaction, and (ii) the waiting time threshold of patients: zero, finite, or infinite. This leads to nine different categories of performance evaluation models as outlined in Figure 2. We review each of these categories below.

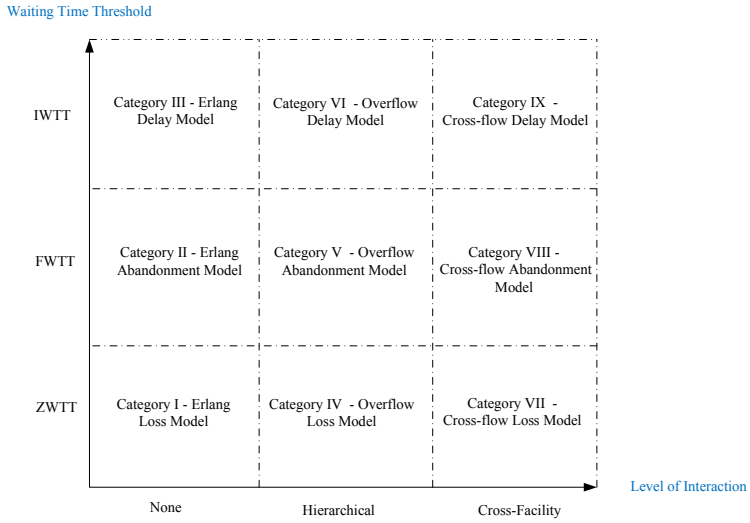


Figure 2: Different categories of performance evaluation models.

Category I: An Erlang loss queueing model, denoted by $M/GI/s/0$ — with a Poisson arrival process (the M), a general service time distribution that is independent of the arrival process (the GI), s servers and 0 waiting space — is generally applied. An example is the work of de Bruin et al. (2009).

Category II: An Erlang abandonment queueing model, denoted by $M/M/s/\infty + M$ — with an Exponential service time distribution (the second M), infinite waiting space (the ∞), and an Exponential waiting time threshold (the $+M$) — is typically applied. For an example in the bed planning literature, see the work of Best et al. (2015).

Category III: An Erlang delay model, denoted by $M/M/s/\infty$, is generally applied. See Green & Nguyen (2001) for an example.

Category IV, V, and VI: We refer to the relevant models as "overflow loss", "overflow abandonment", and "overflow delay" models, respectively, which can be applied for performance evaluation of earmarking as well as clustered overflow configurations under the ZWTT, FWTT, and IWTT assumptions. These categories are the most relevant to our research. Overflow mod-

els represent multi-class hierarchical systems which consist of a collection of dedicated primary facilities and a flexible overflow facility. An exact product-form performance evaluation methodology is proposed in Bekker et al. (2016) for a special case of overflow loss models. For more general cases, approximations have been proposed in the literature. These include the equivalent random method (Cooper, 1990, p.165), Haywards' approximation (Fredericks, 1980), and the Hyper-Exponential decomposition (Franx et al., 2006). Izady & Mohamed (2021) extend the Hayward's model to situations in which mean service time of a customer class in the overflow facility is potentially different from the corresponding mean in the primary facility. This is useful for capturing the impact of focus on LOS in bed allocation models. We did not find any analytical methodology for performance evaluation of overflow abandonment models in the literature. An approximation methodology is proposed in Chevalier & Van den Schrieck (2009) for the special case of overflow delay models where the mean LOS in dedicated and overflow wards are the same, i.e., no impact of focus.

Categories VII, VIII and, IX: We refer to the relevant models as "cross-flow loss", "cross-flow abandonment", and "cross-flow delay" models, respectively. These models are relevant for evaluating the performance of inpatient services with a substantial level of patient outlying. Simulation models are typically applied for performance evaluation of cross-flow models; see, for example, Shi et al. (2016).

2.3. Contributions

We adapt the dynamic programming (DP) methodology proposed in Best et al. (2015) for our inter-cluster bed allocation and partitioning. We also apply the analytical function they propose for representing the impact of focus and workload on LOS. Our methodology differs to that of Best et al. (2015) in the following ways: i) ours is based on the COF configuration, capturing the other four configurations as special cases, while theirs is restricted to the wing formation configuration, which does not capture COF and earmarking; ii) ours takes nursing costs explicitly into account as it can have a substantial impact on the optimal configuration, while theirs does not; and iii) ours works with both IWTT and FWTT assumptions, whereas theirs is restricted to FWTT.

We utilize the COF configuration proposed in Izady & Mohamed (2021) as well as their total cost minimization formulation. We generalize their methodology and experimentation in the following ways. First, we relax the ZWTT assumption made in Izady & Mohamed (2021), which implies that patients are either admitted to their dedicated or overflow wards upon arrival, or turned away immediately. This is not a realistic assumption in many inpatient departments as patients do wait for admission. The relaxation is achieved by developing two

novel performance evaluation approximations, one for overflow delay models, and the other for overflow abandonment models. Exact analysis of both models is extremely challenging due to high dimensionality of the required state vector. Second, Izady & Mohamed (2021) test their methodology on a limited dataset from a paediatric department with 7 specialties, whereas we conduct a comprehensive case study using a large inpatient dataset from a hospital with 16 different specialties.

3. The Optimization Model

We consider the COF configuration proposed in Izady & Mohamed (2021), and adapt their total cost minimization formulation to our setting. Suppose there is a total of B inpatient beds providing care for a total of n specialties. Let $\mathcal{S} = \{1, \ldots, n\}$ be the index set of specialties, and denote by $\boldsymbol{\mathcal{C}} = \{\mathcal{C}^1, \ldots, \mathcal{C}^m\}$ a partition of set $\mathcal{S}$ into $m \in \mathbb{Z}_+$ clusters. We use $\mathbb{Z}$ and $\mathbb{Z}_+$ to denote the set of non-negative and positive integers, respectively. For every cluster $\mathcal{C}^j \in \boldsymbol{\mathcal{C}}$ in the COF configuration, there exists a ward dedicated to patients of each speciality $i \in \mathcal{C}^j$, and an overflow ward j admitting overflowing patients of specialties in the cluster, for $j = 1, \ldots, m$. Let $\boldsymbol{d} = (d_1, \ldots, d_n)$ and $\boldsymbol{o} = (o^1, \ldots, o^m)$ be the dedicated and overflow bed allocation vectors, respectively, with $d_i \in \mathbb{Z}$ representing the number of beds in the ward dedicated to speciality i for $i \in \mathcal{S}$, and $o^j \in \mathbb{Z}$ the number of beds in the overflow ward of cluster j for $j = 1, \ldots, m$.

We assume patients of each specialty request admission according to a stationary Poisson process, independently from other specialties, and their LOSs are independent and identically distributed (i. i. d.) as Exponential random variables. Both assumptions are followed in bulk of the bed allocation literature; see, for example, Best et al. (2015) and Bekker et al. (2016). We denote the rate of admission request for specialty i patients by λ_i. To capture the impact of focus and workload on LOS, we represent the mean LOS for specialty i patients admitted to a $d-$bed ward shared by a subset $\mathcal{A} \ni i$ of specialties by function $\nu_i(d, \mathcal{A})$. For systems with an IWTT, we assume patients wait in their queues until they are served. For systems with a FWTT, we assume waiting time thresholds are i. i. d. according to an Exponential distribution with rate γ for all specialties. This is also the assumption followed in Best et al. (2015). We assume arrival, service, and abandonment processes are mutually independent.

Identifying the optimal configuration and the corresponding bed allocation for the COF configuration requires 4 sets of decision variables, namely, m, $\boldsymbol{\mathcal{C}}$, $\boldsymbol{d}$, and $\boldsymbol{o}$. We define the optimal configuration as the one minimizing the mean total daily cost including the cost of patients abandoning (waiting in) the queue for FWTT (IWTT) assumption, and the cost of nursing teams. Cost minimization formulations are common in the bed allocation literature;

see, for example, Izady & Mohamed (2021) and Wu et al. (2019) . The optimization problem is formulated as

$$Z = \min_{(m,\boldsymbol{\mathcal{C}},\boldsymbol{d},\boldsymbol{o})} \left\{ \sum_{j=1}^{m} T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) : \sum_{i=1}^{n} d_i + \sum_{j=1}^{m} o^j \leq B, \right.$$
$$\left. \boldsymbol{\mathcal{C}} \text{ is a feasible partition of } \mathcal{S},\, m \in \mathbb{Z}_+, \boldsymbol{d} \in \mathbb{Z}^n \text{ and } \boldsymbol{o} \in \mathbb{Z}^m \right\}, \quad (1)$$

where $T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j)$ gives the mean total daily cost of cluster $\mathcal{C}^j$ with $(d_i;\, i \in \mathcal{C}^j)$ dedicated beds and o^j overflow beds. Following Izady & Mohamed (2021), the problem in (1) can be restated as

$$Z = \min_{(m,\boldsymbol{b},\boldsymbol{\mathcal{C}})} \left\{ \sum_{j=1}^{m} \phi(\mathcal{C}^j, b^j) : (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) \in \Psi \right\}, \quad (2)$$

where $\boldsymbol{b} = (b^1, \ldots, b^m)$,

$$\Psi = \left\{ (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) : \sum_{j=1}^{m} b^j \leq B, \boldsymbol{\mathcal{C}} \text{ is a feasible partition of } \mathcal{S}, m \in \mathbb{Z}_+, \text{and } \boldsymbol{b} \in \mathbb{Z}^m \right\},$$

and

$$\phi(\mathcal{C}^j, b^j) = \min_{(d_i; i \in \mathcal{C}^j),\, o^j} \left\{ T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) : \right.$$
$$\left. o^j + \sum_{i \in \mathcal{C}^j} d_i \leq b^j, o^j \in \mathbb{Z}, \text{and } d_i \in \mathbb{Z} \text{ for } i \in \mathcal{C}^j \right\}. \quad (3)$$

The problem in (2) is the partitioning and inter-cluster bed allocation problem, while the one in (3) is the intra-cluster bed allocation problem. To evaluate the cost function T in problem (3), we assume a cost of c_w is incurred each day a patient waits in the queue for systems with an IWTT assumption, and a cost of c_a is incurred for each patient abandoning the queue (as a result of her waiting time exceeding the threshold) for systems with an FWTT assumption. This gives

$$T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = c_w Q(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) + R(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j), \quad (4)$$

and

$$T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = c_a B(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) + R(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j), \quad (5)$$

for IWTT and FWTT systems, respectively, where R represents the mean daily cost of nursing, Q the mean number of patients waiting in the queue, and B the mean daily number of patients

abandoning the system, all for cluster $\mathcal{C}^j$ with bed allocation $(d_i; i \in \mathcal{C}^j), o^j$. Note that c_w and c_a are assumed to be the same across all specialties to simplify the analysis. However, Equations (4) and (5) can be readily adapted to capture different cost parameters for different specialties. It is easy to show that

$$Q(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = \sum_{k \in \mathcal{C}_j} \lambda_k W_k(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j), \tag{6}$$

$$B(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = \sum_{k \in \mathcal{C}_j} \lambda_k A_k(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j), \tag{7}$$

where W_k and A_k give the mean waiting time and probability of abandonment, respectively, of speciality $k \in \mathcal{C}^j$ patients.

To evaluate the nursing cost, $R(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j)$, we apply the minimum nurse-to-patient ratio approach following Izady & Mohamed (2021). As they explain, this approach makes analytical calculations easier, and is also the most common method for establishing nursing requirements in hospitals. Denote by f_i the desired nurse-to-patient ratio for specialty $i \in \mathcal{S}$ patients, and let $r(\mathcal{A})$ be the daily cost of a nurse working in a ward admitting patients of a subset $\mathcal{A}$ of specialties. We then have

$$R(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = \sum_{k \in \mathcal{C}_j} r(\{k\}) \left\lceil S_k^d(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j) f_k \right\rceil + r(\mathcal{C}^j) \left\lceil \sum_{k \in \mathcal{C}^j} S_k^o(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j) f_k \right\rceil, \tag{8}$$

where functions $S_k^d(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$ and $S_k^o(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$ give the mean numbers of beds occupied by specialty $k \in \mathcal{C}^j$ patients in their dedicated and overflow wards, respectively, and $\lceil x \rceil$ gives the smallest integer larger than or equal to x.

Equation (8) employs the expected number of patients in each ward to derive the base staffing levels associated with each cluster. This approach is in line with the common practice in healthcare institutions, where base staffing levels, reflective of average occupancy rates, are utilized for mid- to long-term decision making. As explained in Malaki et al. (2023), these base staffing levels are typically supplemented by temporary staffing adjustments made several hours or days prior to each shift to effectively respond to short-term fluctuations in patient demand. We note that the nurse numbers in Equation (8) are rounded up to ensure the allocation of an integer number of nurses to each ward.

In the next section, we show how performance metrics W_k, A_k, S_k^d, and S_k^o can be estimated for a cluster with a given bed allocation.

4. The Performance Evaluation Models

In this section, we develop approximation methodologies for estimating performance metrics of overflow delay systems as well as overflow abandonment systems. We define an overflow queueing system as a hierarchical multi-class queueing system with two types of server pools: i) dedicated pools, each specialized to serve a single class of customers; and ii) an overflow pool, cross-trained to serve all classes of customers.

Upon arrival, customers will be served by their dedicated pool if it has an idle server available, and the overflow pool otherwise. If both pools are busy, customers wait in dedicated queues corresponding to their classes; see Figure 3. We assume that once a server in a dedicated pool becomes available, it serves the next customer in its dedicated queue according to a first-come first-served (FCFS) discipline. Once a server in the overflow pool becomes available, on the other hand, we assume that it serves the next customer from the longest queue following a FCFS discipline. The longest queue policy is found to outperform the other major policies in Jordan et al. (2004). There exists no transfer of customers between dedicated and overflow pools.

In overflow delay systems, customers are infinitely patient and wait in their queues until they are served. In overflow abandonment systems, customers are impatient and abandon the queue once their waiting time threshold is reached. A cluster in the COF configuration with an IWTT can therefore be represented as an overflow delay system, while a cluster with an FWTT can be represented as an overflow abandonment system.

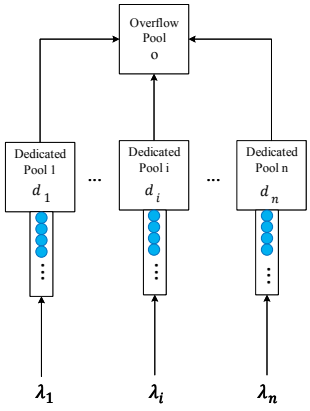


Figure 3: The schematic diagram of an overflow system.

Suppose there are n customer classes with $\mathcal{S} = \{1, 2, \ldots, n\}$ the corresponding index set. Let d_i be the number of servers in the pool dedicated to class i customers for $i \in \mathcal{S}$, and $\boldsymbol{d} = (d_1, \ldots, d_n)$ be the corresponding vector. Let o be the number of serves in the overflow pool. Suppose class i customers arrive to the system according to a Poisson process with rate

λ_i, and their service times are i. i. d. as an Exponential distribution with mean ν_i if service is provided by the corresponding dedicated pool, and ν_i' if service is provided by the overflow pool, for $i \in \mathcal{S}$. For the overflow abandonment system, in additions to the assumptions above, we assume times to abandon are i. i. d. following an Exponential distribution with rate γ for all customer classes. Arrival, service and abandonment processes are assumed to be mutually independent. Following the formulation in Section 3, we need to evaluate mean waiting time in the queue for the overflow delay system, and the probability of abandonment for the overflow abandonment system, for each class of customers. To evaluate nursing costs, we also need to evaluate mean busy servers in dedicated and overflow pools for both systems.

To demonstrate the complexities of performance evaluation in overflow systems, an exact analysis is conducted in Section 4.2 of Arabzadeh (2022) for a simplified overflow delay system with only two customer classes and single-server primary and overflow service facilities. Their analysis indicates that a four-dimensional state vector is required and the total size of the state space would be $2(N+1)^2+4$, where N is the maximum number of customers in the system for each class excluding the one in the overflow server. As shown in numerical experiments in Arabzadeh (2022), the computation time exceeds one hour with $N = 20$, implying that an exact performance evaluation would not be practical for finding the optimal bed allocation within a cluster, which may potentially include more than two specialties and require a larger N. We therefore propose approximation methodologies for performance evaluation of hierarchical queues.

Our approximations includes two main steps. For both delay and abandonment systems, the first step involves estimating the blocking probabilities of different customer classes in an equivalent overflow loss system. For overflow delay systems, the second step of our approximation involves converting the estimated blocking probabilities in the overflow loss system to mean waiting times in the overflow delay system using the exact relation between loss probability in single-class $M/M/s/0$ loss queues and mean waiting time in single-class $M/M/s/\infty$ delay queues. For overflow abandonment systems, the second step involves converting the estimated blocking probabilities in the overflow loss system to abandonment probabilities in the overflow abandonment system using the exact relation between loss probability in single-class $M/M/s/0$ loss queues and abandonment probability in single-class $M/M/s/\infty + M$ abandonment queues.

Section 4.1 explains how loss probabilities are estimated in overflow loss systems. The approximations for evaluating performance in overflow delay and abandonment systems are elaborated in Sections 4.2 and 4.3, respectively. The accuracy of approximations are tested against simulation results in Section 4.4.

4.1. Overflow Loss Systems

In this section, we utilize the approximation method proposed by Izady & Mohamed (2021), which accounts for differences in mean service times between dedicated and overflow pools. This distinction is crucial for capturing the impact of focus on LOS. The approximation estimates the blocking probability faced by customers of class i in an overflow loss system (with the same arrival and service processes as the overflow delay or abandonment system) as follows:

$$L_i = B_e(a_i, d_i)L \approx B_e(a_i, d_i)B_e(\alpha/\beta, o/\beta), \tag{9}$$

where $B_e(a, d)$ is a continuous extension of the Erlang loss function (such as $B_e(a, d) = [a\int_0^\infty \exp(-at)(1+t)^d dt]^{-1}$ proposed in Jagerman, 1974), $a_i = \lambda_i \nu_i$ is the offered load of class i customers, and L is the blocking probability experienced by the aggregate stream overflowing dedicated pools. By Hayward's approximation, L can be estimated as:

$$L \approx B_e(\alpha/\beta, o/\beta). \tag{10}$$

Here, α is the offered load of the aggregate overflow stream, computed as:

$$\alpha = \sum_{i\in\mathcal{S}} a_i B_e(a_i, d_i)/\rho_i, \tag{11}$$

where $\rho_i = \nu_i/\nu_i^{'}$ is the mean service ratio, and β represents the "peakedness" (see Fredericks, 1980 for its definition) of the aggregate overflow stream, evaluated by:

$$\beta = \frac{1}{\alpha}\sum_{i\in\mathcal{S}} \frac{a_i}{\rho_i} B_e(a_i, d_i)\xi(a_i, d_i, \rho_i). \tag{12}$$

In Equation (12), $\xi(a_i, d_i, \rho_i)$ denotes the peakedness of the stream overflowing from dedicated pool i. As per Proposition 1 in Izady & Mohamed (2021), this peakedness is computed using the following formula:

$$\xi(a_i, d_i, \rho_i) = 1 - \frac{a_i B_e(a_i, d_i)}{\rho_i} + \frac{a_i(a_i+\rho_i)\,{}_3F_1(\rho_i, 1-d_i, a_i+\rho_i+1; a_i+\rho_i; -1/a_i)}{\rho_i(a_i+\rho_i+1)\,{}_3F_1(1-d_i, \rho_i+1, 2+a_i+\rho_i; a_i+\rho_i+1; -1/a_i)}, \tag{13}$$

where ${}_pF_q(a_1, ..., a_p; b_1, ...b_q; x)$ represents the generalized hypergeometric function.

4.2. Overflow Delay Systems

In this section, we estimate the probability of delay and mean waiting time experienced by class i customers, denoted by P_i and W_i, respectively, in an overflow delay system using the estimate of L_i provided in Equation (9). To start, consider an $M/M/s/\infty$ delay queue with offered load $a < s$. The probability of delay, $P_{M/M/s/\infty}$, in this queue is related to the loss probability, $L_{M/M/s/0}$, in an equivalent $M/M/s/0$ loss queue through the equation (see Cooper, 1990, Chap. 10, Equation 5.31):

$$P_{M/M/s/\infty} = \frac{sL_{M/M/s/0}}{s - a(1 - L_{M/M/s/0})}. \tag{14}$$

Motivated by this equation, Chevalier & Van den Schrieck (2009) propose

$$P_i \approx \frac{s_i L_i}{s_i - \delta_i(1 - L_i)}, \tag{15}$$

for estimating the delay probability faced by class i customers in an overflow delay system. In Equation (15), L_i is estimated through Equation (9), s_i is a suitably defined number of servers allocated to class i customers as we shall explain below, and δ_i is the offered load of class i customers in the overflow delay system evaluated by:

$$\delta_i = \lambda_i \bar{\nu}_i, \tag{16}$$

where $\bar{\nu}_i$ is the weighted average of class i mean service times in dedicated and overflow pools obtained by the following equation:

$$\bar{\nu}_i = (1 - \chi_i)\nu_i + \chi_i \nu_i'. \tag{17}$$

Here, χ_i is the fraction of class i customers served by the overflow pool in the overflow delay system. We propose estimating this fraction by the corresponding fraction in the associated overflow loss system, i.e.,

$$\chi_i \approx B_e(a_i, d_i)(1 - L), \tag{18}$$

with L given in Equations (10). This loss-based approximation of χ_i is more accurate than the fluid approximation utilized in Chevalier & Van den Schrieck (2009) as it captures the uncertain nature of inter-arrival and service times.

To evaluate s_i, we first define I_i^d as the idle service capacity in dedicated pool i, and I_i^o as the portion of idle service capacity in the overflow pool allocated to class i customers. Both metrics

have the same dimension as arrival and service rates, i.e., they are measured in customers per time unit. To estimate I_i^d and I_i^o, we combine the fluid approximation proposed by Chevalier & Van den Schrieck (2009) with the results from the overflow loss system. In particular, we estimate I_i^d as

$$I_i^d \approx (d_i/\nu_i - (\lambda_i - \lambda_i'))^+, \tag{19}$$

where $x^+ = \max\{x, 0\}$, and λ_i' is the rate of customer overflow from dedicated pool i estimated using the overflow loss system results as:

$$\lambda_i' \approx \lambda_i \chi_i. \tag{20}$$

The overall idle capacity at the overflow pool is therefore

$$I^o \approx (o/\bar{\nu} - \sum_{i \in \mathcal{S}} \lambda_i')^+, \tag{21}$$

where $\bar{\nu}$ is the weighted average of mean service times of different customer classes in the overflow pool given by

$$\bar{\nu} = \sum_{i \in \mathcal{S}} \lambda_i' \nu_i' / \sum_{i \in \mathcal{S}} \lambda_i'. \tag{22}$$

Assuming customer classes with higher loss probabilities in the equivalent overflow loss system will have a higher proportion of the overflow pool idle capacity, we have

$$I_i^o \approx \frac{\lambda_i L_i}{\sum_{i \in \mathcal{S}} \lambda_i L_i} I^o, \tag{23}$$

with L_i given in Equation (9). Note that our estimation of I_i^d and λ_i' given in Equations (19) and (20), respectively, are more accurate than those of Chevalier & Van den Schrieck (2009), i.e., $I_i^d \approx (d_i/\nu_i - \lambda_i)^+$ and $\lambda_i' \approx (\lambda_i - d_i/\nu_i)^+$, which are purely based on fluid approximation. We now propose

$$s_i = I_i^o \bar{\nu} + \lambda_i' \nu_i' + d_i. \tag{24}$$

Equation (24) implies that s_i equals the size of dedicated pool i plus the sum of idle and used capacity of the overflow pool allocated to class i customers. We use Equation (24) instead of the conservation equation $s_i = (\lambda_i + I_i^d + I_i^o)\nu_i$ proposed by Chevalier & Van den Schrieck (2009) as it does not account for different mean service times in dedicated and overflow pools. Substituting s_i and δ_i in Equation (15) with their corresponding values given in Equations (24) and (16),

respectively, and simplifying, we obtain

$$P_i = \frac{\left(I_i^o \bar{\nu} + \lambda_i^{'} \nu_i^{\prime} + d_i\right) L_i}{I_i^o \bar{\nu} + \lambda_i^{'} \nu_i^{\prime} + d_i - \lambda_i \bar{\nu}_i (1 - L_i)}. \tag{25}$$

The next step is to convert the delay probability P_i to mean waiting time W_i. For this, we use the relation $W_{M/M/s/\infty} = P_{M/M/s/\infty}/(s/\nu - \lambda)$ between delay probability, $P_{M/M/s/\infty}$, and mean waiting time, $W_{M/M/s/\infty}$, in an $M/M/s/\infty$ queue with arrival rate λ and mean service time ν (see, e.g., Cooper, 1990, Chap. 10, Equation 5.31). The denominator of this relation is in fact the idle service capacity. Motivated by this, we have the following approximation

$$W_i \approx \frac{P_i}{I_i^d + I^o}. \tag{26}$$

Note that if $I_i^d + I^o = 0$, the system is unstable for class i customers and so $P_i = 1$ and $W_i \to \infty$. Applying Little's law on server pools, we obtain the mean number of class i customers in the dedicated and overflow pools by

$$S_i^d \approx \lambda_i (1 - \chi_i) \nu_i, \tag{27}$$

and

$$S_i^o \approx \lambda_i \chi_i \nu_i^{'}, \tag{28}$$

respectively, where the fraction of customers served by the overflow pool is estimated by the corresponding fraction in the overflow loss system.

4.3. Overflow Abandonment Systems

Consider an $M/M/s/\infty + M$ abandonment queue with arrival rate λ, mean service time ν, and abandonment rate γ. The abandonment probability $A_{M/M/s/\infty+M}$ in this queue is related to the loss probability $L_{M/M/s/0}$ in an equivalent $M/M/s/0$ loss queue through the following equation (see Equation 5.22 in Zhang, 2010)

$$A_{M/M/s/\infty+M} = \frac{s(1 + f(c,\eta)(\eta/c - 1))L_{M/M/s/0}}{\lambda\nu\left(1 + (f(c,\eta) - 1)L_{M/M/s/0}\right)}, \tag{29}$$

where

$$f(c,\eta) = \sum_{i=0}^{\infty} \frac{\Gamma(c+1)\eta^i}{\Gamma(c+i+1)},$$

with $c = s/\nu\gamma$, $\eta = \lambda/\gamma$, and $\Gamma(x) = \int_0^\infty y^{x-1}e^{-y}dy$ is the Gamma function. Motivated by Equation (29), we propose the approximation

$$A_i \approx \frac{s_i(1 + f(c_i, \eta_i)(\eta_i/c_i - 1))L_i}{\lambda_i\bar{\nu}_i\,(1 + (f(c_i, \eta_i) - 1)L_i)}, \tag{30}$$

for the abandonment probability experienced by class i customers in the overflow abandonment system, where s_i is a suitably defined number of servers for class i customers, $c_i = s_i/\bar{\nu}_i\gamma$, $\eta_i = \lambda_i/\gamma$, $\bar{\nu}_i$ is the average mean service time of class i customers given in Equation (17), and L_i is estimated through Equation (9). For s_i, we use the same equation as for the overflow delay system, i.e.,

$$s_i = I_i^o\bar{\nu} + \lambda_i^{'}\nu_i' + d_i, \tag{31}$$

with I_i^o, $\bar{\nu}$, and $\lambda_i^{'}$ given in Equations (23), (22), and (20), respectively. Once abandonment probabilities are obtained, we can use Equations (27) and (28) to estimate mean numbers of class i customers in the dedicated and overflow pools, respectively, by replacing arrival rate λ_i with effective arrival rate $\lambda_i(1 - A_i)$.

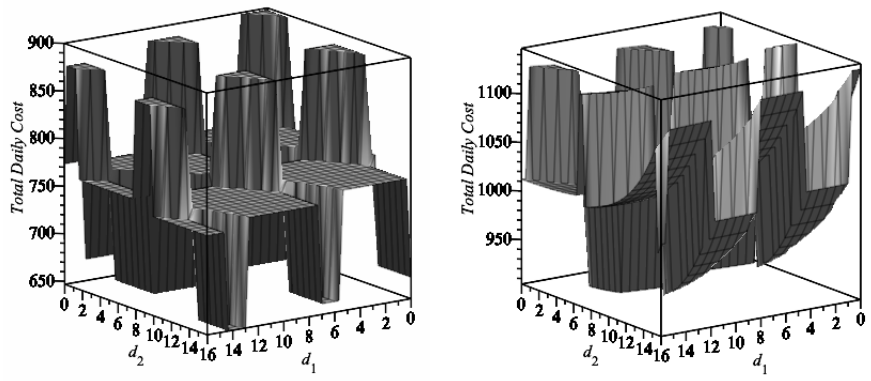
4.4. Simulation Experiments

We perform an extensive series of simulation experiments to assess the accuracy of our approximation methodologies, along with the one proposed by Chevalier & Van den Schrieck (2009). The details of these experiments and their corresponding outcomes can be found in Section A of the e-companion. The results show a practical level of accuracy for our approximation methodologies. They also show that our approximation for overflow delay systems consistently outperforms the methodology proposed by Chevalier & Van den Schrieck (2009). Consequently, for a cluster with a given bed allocation, we can apply our approximation methodologies to estimate the requisite performance metrics for computing the mean total daily cost of the cluster, as given in Equations (4) and (5) for ITWTT and FWTT systems, respectively.

5. Solving the Models

We start with the intra-cluster allocation problem in (3). Considering a hypothetical cluster $\mathcal{C} = \{1, 2\}$ with $b = 70$ beds, in Figure 4 we plot the objective function of this problem, i.e., $T(\mathcal{C}, (d_i;\, i \in \mathcal{C}), o)$, with a given set of parameters as a function of d_1 and d_2 under IWTT as well as FWTT assumptions. The plots in this figure demonstrate that the objective function is neither convex nor differentiable for either assumption. This implies that we need to apply a gradient-free heuristic optimization method for finding a good solution. Izady & Mohamed

(2021) propose the CDOS heuristic developed by Moiseev (2011) for finding a good bed allocation in a cluster under the ZWTT assumption, i.e., assuming the cluster works as an overflow loss system. The experiments presented in Section B of the e-companion show that CDOS performs relatively well under the FWTT and IWTT assumptions too. The simulation optimization experiments conducted in the same section also illustrate that CDOS combined with our performance evaluation methodologies typically results in the correct allocation of beds. As such, we will use it for intra-cluster bed allocation.



(a) (b)

Figure 4: Surface plots for the total mean daily cost of a 70-bed cluster with 2 specialties assuming $f_1 = f_2 = 0.149$ and $\rho_1 = \rho_2 = 1.0$. Panel (a) is for an IWTT system with $\lambda_1 = 5.2$, $\lambda_2 = 3.9$, $\nu_1 = 4.8$, $\nu_2 = 2.9$, and $c_w = 1030$, and panel (b) is for an FWTT system with $\lambda_1 = 6.2$, $\lambda_2 = 4.9$, $\nu_1 = 4.8$, $\nu_2 = 2.9$, $\gamma^{-1} = 5$ and $c_a = 1030$

For the partitioning and inter-cluster allocation problem in (2), following Best et al. (2015), we first restrict the feasible region Ψ by focusing only on partitions obtained by making cuts along a fixed sequence $\mathcal{N}$ of specialties. We then solve the restricted model using the DP approach proposed in Best et al. (2015), with the difference that expected reward for each state-action pair is evaluated using the CDOS heuristic explained above. See Section 5.3 in Arabzadeh (2022) for further details.

6. Case Study

We started a collaborative project with RSCH in January 2019. RSCH is a general NHS hospital located in Surrey, UK, providing emergency and general hospital services to a population of more than 330,000 people living across south west Surrey. As we shall demonstrate in Section 6.1, the hospital was experiencing a high level of bed occupancy even before the start of the COVID-19 pandemic. The aim of our collaboration was therefore to find innovative ways to reduce the bed pressure on hospital with a focus on reconfiguration of inpatient services. Here we report the interim results of our collaboration with RSCH.

6.1. Current Status

Data Coverage. Admission data covering a three-year period starting from 01/10/2015 is provided by the hospital. The data captures a wide range of information for each patient admitted to the hospital. In particular, it contains information about different episodes of care within each hospital spell, including (but not limited to) their specialty, primary and secondary diagnoses, and the procedures conducted. In the NHS jargon, a hospital spell is defined as "... the total continuous stay of a patient using a hospital bed on premises controlled by a health care provider during which medical care is the responsibility of one or more consultants ...", while an episode of care means "...the time a patient spends in the continuous care of one consultant..." (NHS Data Model and Dictionary, 2021). For each episode of care, the data also captures the sequence of beds the patient has visited, the start and end dates and times of each bed visit, and the corresponding ward. For example, the spell illustrated in Table 1 has 4 episodes of care, 2 of which belong to Geriatric Medicine, one to Rheumatology, and one to Respiratory specialties. The first episode, for example, involves 4 bed visits, three of which occur in the same ward.

Episode No	Specialty	Bed Visit	Visit Start Date & Time	Visit End Date & Time	Ward
1	Geriatric Medicine	1	01/10/2015 02:58	01/10/2015 10:07	EAU
		2	01/10/2015 10:07	01/10/2015 17:58	EAU
		3	01/10/2015 17:58	01/10/2015 18:19	M2
		4	01/10/2015 18:19	02/10/2015 10:40	EAU
2	Rheuma-tology	5	02/10/2015 10:40	02/10/2015 19:24	EAU
		6	02/10/2015 19:24	02/10/2015 21:30	EAU
		7	02/10/2015 21:30	03/10/2015 07:50	M2
3	Geriatric Medicine	8	03/10/2015 07:50	05/10/2015 15:10	M2
4	Respiratory Medicine	9	05/10/2015 15:10	07/10/2015 23:00	M1
		10	07/10/2015 23:00	10/10/2015 14:40	M1

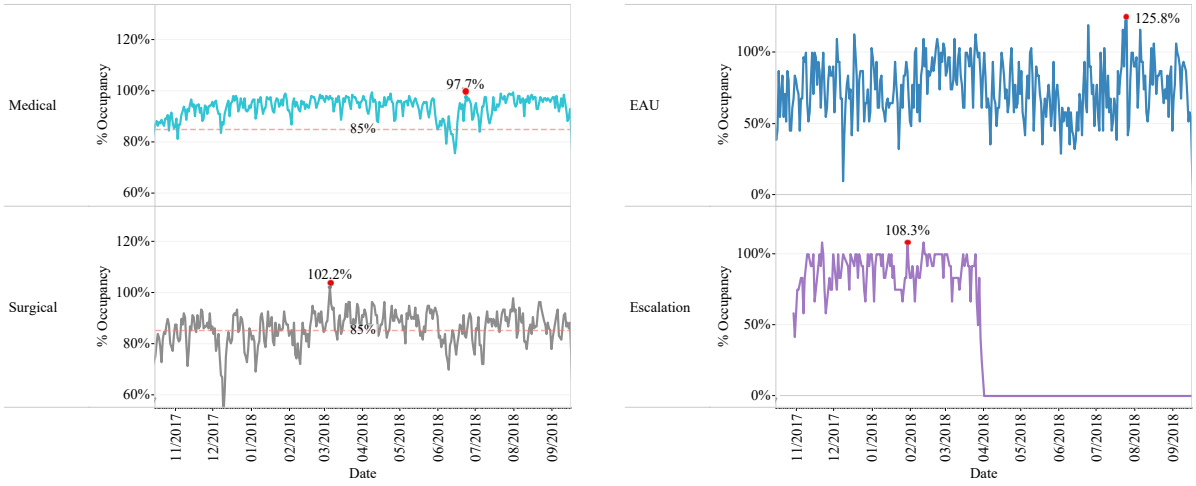
Table 1: An example of episodes of care and corresponding bed visits within a hospital spell at RSCH.

There exists a total of 394 inpatient beds in the hospital divided among 13 inpatient wards, including 8 medical wards (215 beds), 5 surgical wards (136 beds), an escalation ward (12 beds), and an emergency assessment unit (EAU; 31 beds). The escalation ward is mainly used during Winter when demand for inpatient beds is at its peak. EAU is a short stay specialist assessment and admission facility specifically for patients whose LOSs are expected to be less than 48 hours. Episodes of care spent entirely in a paediatric, rehabilitation, mental health, or intensive care ward are excluded from our analysis as these wards often have dedicated resources, and so are not included in the reconfiguration exercise. We also exclude all episodes spent entirely in a day-case ward, e.g., in the Cardiac Day Ward or the Endoscopy Unit, so as to keep the focus on inpatient care. This leaves a total of 73,466 unique hospital spells with at least a period of care spent in an inpatient ward. The rest of our analysis is restricted to these spells.

Bed Occupancy. The numbers of beds and average occupancy levels of inpatient wards at RSCH are given in Table 2. The occupancy figures given in this table demonstrate the significant strain on hospital beds. This is highlighted further in Figure 5 which represent the daily occupancy levels over the last year of our data coverage period. In particular, we observe in this figure that daily occupancy exceeds 85% for significant periods of time for both medical and surgical wards, and that EAU and escalation wards' occupancies hit 100% frequently.

Wards	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	E	EAU
No. Beds	30	30	16	30	24	30	31	24	28	19	30	30	29	12	31
Occupancy (%)	93	96	87	96	92	90	92	95	80	85	88	87	92	88	76

Table 2: Bed numbers and average occupancy levels of different inpatient wards at RSCH. We use "M" for medical wards, "S" for surgical wards, "E" for the escalation ward, and "EAU" for the emergency assessment unit.



(a) (b)

Figure 5: Daily occupancy levels at RSCH over the period 01/10/2017 to 01/10/2018 for medical and surgical wards (a), and EAU and escalation wards (b). Occupancy rates in excess of 100% indicate times at which patients were accommodated in non-inpatient wards.

Re-defining Specialties. For the purpose of reconfiguration, we need to identify the specialty for each episode of care as accurately as possible. This is also important for assessing the extent as well as impact of patient outlying in the hospital. However, the "Specialty" field provided in the hospital data (see Table 1) represents the specialty of the consultant in charge of the corresponding care episode, which may not necessarily represent the clinical requirement of the patient, in particular when the patient is outlied. To identify the correct specialty for each episode of care, we use "Finished Consultant Episode - Healthcare Resource Group" (FCE-HRG) codes provided in the data. HRG codes "... are standard groupings of clinically similar treatments which use common levels of healthcare resources ...", and are used for costing purposes. They consist of 5 parts each referring to a specific characteristic (NHS Data Model and

Dictionary, 2021). The FCE-HRG code in our data identifies the HRG code for each episode of care. We derive the specialty of each episode from the first two letters of this code as explained in Section C of the e-companion. Setting the specialties in this way, we observe that RSCH provides inpatient care for a total of 18 specialties. Removing Radiology specialty, and combining Gynaecology and Obstetrics as well as Haematology and Oncology, we obtain 16 specialties, including 9 medical and 7 surgical ones, as listed along with their abbreviations in Table 3.

Specialities	Acronym	Division
Cardiology	CRD	Medical
Endocrinology	END	Medical
Ear, Nose and Throat	ENT	Surgical
Gastroenterology	GAS	Medical
General Surgery	GSR	Surgical
Geriatric Medicine	GRT	Medical
Gynaecology & Obstetrics	GYN	Surgical
Neurology	NRO	Medical
Oncology & Haematology	ONC	Medical
Ophthalmology	OPL	Surgical
Oral & Maxillo Facial	ORM	Surgical
Palliative Medicine	PAL	Medical
Respiratory Medicine	RSP	Medical
Rheumatology	RUM	Medical
Trauma & Orthopaedics	ORT	Surgical
Urology	URO	Surgical

Table 3: Specialties at RSCH, and their corresponding divisions and acronyms.

To gain a better understanding of patients' journeys in inpatient services, a breakdown of a hospital spell is provided in Figure 6. This diagram shows that each hospital spell contains one or more *specialty spells*, which we define as the continuous amount of time a patient spends within one specialty. The change in specialty during a hospital spell is often due to different medical needs of a patient. Each specialty spell, in turn, includes one or more episodes of care, each of which contains one or more *bed visits*. We define a bed visit as a period of time a patient occupies a specific bed in a specific ward.

The data provided by hospital already includes timings and the other relevant information for hospital spells, episodes of care, and bed visits (see Table 1). We create specialty spells by linking together the episodes of care within a hospital spell that have the same specialty. Our analysis of 73,466 hospital spells show that each spell includes an average (maximum) of 1.08 (5.00) specialty spells, each specialty spell includes an average (maximum) of 1.29 (8.00) episodes of care, and each episode of care includes an average (maximum) of 2.79 (25.00) bed visits. We obtain the length of each specialty spell by adding up the lengths of its constituent episodes of care, and refer to it as the specialty LOS.

Patient Outlying and its Impact. To estimate the extent of patient outlying at RSCH, we measure the percentage of specialty spells admitted to a non-primary ward. To do this, we first

Figure 6: The breakdown of a hospital spell.

identify the primary ward(s) of each specialty as given in Table 4. This table is provided by our hospital partners based on their perception of what happens on the ground as well as the skill-set of nursing teams in different wards. We then count a specialty spell as an outlying spell if it has at least one bed visit in a non-primary ward of the corresponding specialty. The results indicate that about 49% of all specialty spells are outlying, with the breakdown given in panel (a) of Figure 7. This panel shows that for 8 specialties, half or more of specialty spells are outlying.

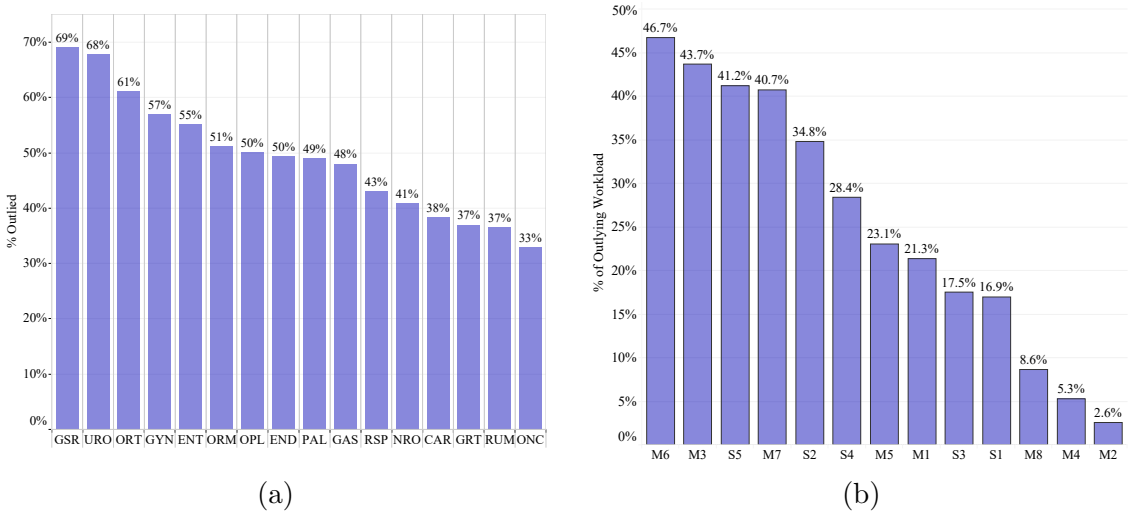
To estimate the contribution to workload by outlying patients, we count a bed visit as an outlying visit if it occurs in a non-primary ward of its corresponding specialty. We then add up the lengths of outlying bed visits and divide it by the sum of all bed visits. The corresponding percentages are presented in panel (b) of Figure 7 for different wards. This figure suggests that the contribution of outlying patients varies significantly from ward to ward, with M6 (dedicated to GAS) having the highest contribution and M2 (dedicated to GRT) the lowest. Overall, outlying patients account for 27.7% of inpatient workload at RSCH.

Specialty	CRD	END	ENT	GAS	GSR	GRT	GYN	NRO	ONC	OPL	ORM	PAL	RSP	RUM	ORT	URO
Primary Wards	M5	M3	S2	M6	S5	M2 M8 M3 M4	S3 S5	M8	M7	S2	S2	M1 M7	M1	M5 M7	S1 S4	S3

Table 4: The primary wards of different specialties at RSCH.

As stated in Section 1, studies such as Stowell et al. (2013) report that outlying patients usually have a longer LOS. We investigate this on GSR and URO specialties which have the largest percentage of outlying spells. We observe that outlying GSR and URO specialty spells are 0.92 and 2.05 days, respectively, longer than the corresponding non-outlying spells in our data.

Given the large number of outlying specialty spells at RSCH, it would also be interesting to see their impact on non-outlying spells, i.e., those occurring in primary wards of corresponding specialties. In a recent study, Lim et al. (2021) demonstrate empirically that mean LOS of non-outlying patients is longer in wards that receive a larger number of outlying patients. To



(a) (b)

Figure 7: Percentage of outlying specialty spells for each specialty (a), and contribution to workload by outlying patients for each ward (b).

investigate this impact on RSCH data, we focus on GRT and ORT specialties, which account for 15% and 11% of specialty spells, respectively, at RSCH. According to Table 4, wards M4 and M2 are both primary wards for GRT specialty. However, our analysis show that mean LOS of this specialty in ward M2, in which the number of outlying spells account for 26.0% of all spells, is 1.25 days longer than in ward M1, wherein outlying spells stand at 16.4%. Similarly, wards S4 and S1 are both primary wards for ORT. But mean LOS of this specialty in ward S4, with 73% outlying spells, is 2.72 days longer than in ward S1, with 30% outlying spells. These analyses highlight the impact of outlying patients on both outlying and non-outlying spells.

Ward Change and its Impact. The analysis presented in Section D of the e-companion reveals a high number of ward changes within a specialty spell. Additionally, the analysis suggests that as the number of ward changes increases, the mean LOS of the specialty increases too, with the impact being more significant for elderly patients.

Overall Situation. The analyses conducted above portrays a hospital under immense pressure throughout the year. While a dedicated configuration was originally intended for the inpatient services at RSCH, a wing formation configuration with overlapping clusters is currently operating in the hospital. This is evident from the partitioning of specialties, derived from Table 4, as below

$$\begin{aligned}\mathcal{C}_{current} = \{&\{\text{ORT}\},\{\text{ENT},\text{OPL},\text{ORM}\},\{\text{URO},\mathbf{GYN}\},\{\text{GSR},\mathbf{GYN}\},\{\mathbf{PAL},\text{RSP}\},\\ &\{\mathbf{GRT}\},\{\text{END},\mathbf{GRT}\},\{\text{GRD},\mathbf{RUM}\},\{\text{GAS}\},\{\mathbf{PAL},\mathbf{RUM}\},\{\mathbf{GRT},\text{NRO}\}\},\end{aligned}$$

wherein highlighted specialties are allocated to more than one cluster. There also exists a

substantial amount of patient outlying and ward changes in the hospital. Outlying patients not only experience a longer LOS, but may also negatively influence the LOS of other patients. This creates a vicious circle, wherein some patients are admitted to non-primary wards due to bed unavailability. This results in a longer LOS for both outlying and non-outlying patients, exacerbating the bed shortage problem, which in turn leads to more patients being outlied. A new configuration of inpatient services, in which beds are pooled in a structured way so as to reduce the number of patients outlying while minimizing the negative impacts of losing focus and increase in mix variability, is therefore likely to create some improvements. We investigate this in the next section.

6.2. Parameter Estimation

For estimating the input parameters of our models, we focus on specialty spells occurred within the last year of our data coverage period, i.e., from 01/10/2017 to 01/10/2018. The arrival rate, λ_i, of each specialty is evaluated by dividing the total number of spells of that specialty within this period by 365. The mean LOS of each specialty, m_i, and its coefficient of variation (CV; standard deviation divided by mean), κ_i, are obtained using the LOSs of that specialty within the coverage period. Nurse-to-patient ratio of each specialty, f_i, is provided by the hospital. See Table 5 for the set of specialties and corresponding λ_i, m_i, f_i, and κ_i values.

Specialty	CRD	END	ENT	GAS	GSR	GRT	GYN	NRO	ONC	OPL	ORM	PAL	RSP	RUM	ORT	URO
λ_i	5.29	1.55	4.47	9.29	6.22	11.33	3.18	2.29	1.98	0.21	0.13	0.08	6.33	1.90	6.99	5.84
m_i	3.50	3.35	2.59	3.78	5.28	8.90	1.93	4.68	5.54	2.29	2.82	7.73	6.36	5.22	5.80	3.08
f_i	0.15	0.15	0.15	0.20	0.2	0.2	0.15	0.15	0.2	0.20	0.15	0.20	0.20	0.15	0.15	0.16
κ_i	1.88	1.65	2.10	1.81	1.92	1.43	2.13	1.77	1.71	2.59	1.27	1.16	1.36	1.60	1.40	1.83

Table 5: Input parameters for our reconfiguration models. Time unit is one day.

To apply our methodology for bed reconfiguration, we set the total number B of beds to 394. Dividing the total offered load, obtained as the sum-product of λ_i and m_i values given in Table 5, by this number of beds yields an overall traffic intensity of 86.7%. Note that the beds in the EAU and escalation wards are included in the total number of beds as without them the traffic intensity would rise to 97.5%, making the queues extremely long. We use the functional relation

$$\nu_i(d, \mathcal{A}) = \left(1 - \frac{\Delta\left(1 - \frac{|\mathcal{A}|}{n}\right)}{1 + e^{-\beta\left(\sum_{i\in\mathcal{A}} \frac{\lambda_i \tau_i}{d} - \epsilon\right)}}\right) \tau_i, \tag{32}$$

as proposed in Best et al. (2015) for estimating the mean LOS of specialty i patients admitted to a $d-$bed ward shared by a subset $\mathcal{A} \ni i$ of specialties. In Equation (32), $|x|$ represents the cardinality of set x, τ_i is the nominal mean LOS for specialty i patients (excluding the impact of

focus and workload), Δ controls the impact of focus, and β and ϵ control the impact of workload as evaluated by $\sum_{i\in\mathcal{A}} \lambda_i \tau_i / d$.

Equation (32) indicates that, under a given workload, the maximum reduction in LOS is achieved when only one specialty, denoted as $|\mathcal{A}| = 1$, is assigned to the ward. However, as more specialties are assigned, the reductions gradually decline, eventually reaching 0% when all n specialties are included. On the other hand, when a fixed set of specialties $\mathcal{A}$ is considered, the reduction in LOS is initially modest for smaller workloads but increases as the workload grows until it approaches the asymptotic value of $\Delta\,(1 - |\mathcal{A}|/n)$. The presence of the exponential function in the denominator of the equation reflects the increasing (decreasing) pace of reductions in LOS with respect to workload when workload is smaller (larger) than ϵ; see Figure 3 in Best et al. 2015, p. 166.

Given the substantial amount of patient outlying in the current configuration, we treat the existing specialty mean LOS values as nominal mean LOS values required for Equation (32). Specifically, we set τ_i in Equation (32) equal to the mean LOS value obtained from the data, i.e., m_i reported in Table 5. Following Best et al. (2015), we also set $\epsilon = 0.9$ and $\beta = 20.0$.

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 average daily salary of a band 5 nurse equal to £103.03 per day plus a 10% additional payment for each additional specialty the nurse cares for in order to represent the higher value of multi-skilled nurses to the hospital.

Three additional constraints are included to make the configurations proposed by our methodology viable. The first constraint is that medical and surgical specialties cannot be mixed in a cluster. The second constraint requires that the RSP must not be mixed with any other specialty to reduce the risk of in-hospital transmission of respiratory diseases. The last constraint is that for privacy reasons, GYN must also not be mixed with any other specialty. These three constraints are implemented by returning large cost values for clusters involving the specialties that cannot be mixed together in the intra-cluster allocation model. Note that the number of beds for RSP and GYN specialties must still be determined by our methodology.

The sequence $\mathcal{N}$ of specialties is created as follows: i) medical specialties excluding RSP are sorted in terms of their mean nominal LOS; ii) surgical specialties excluding GYN are sorted in terms of their mean nominal LOS, and inserted at the end of the sequence created in step (i); and iii) RSP and GYN are added to the end of the sequence created in step (ii). These steps ensure that specialties are sorted in terms of their mean nominal LOS as proposed in Izady & Mohamed (2021), while taking the three constraints mentioned above into account.

6.3. Best-Found Configurations with IWTT Assumption

Following Izady & Mohamed (2021), we set $c_w = 1030$, i.e., 10 times larger than the daily salary of a nurse. In all the experiments we conduct, we consider two values for Δ: 0.0 and 0.1. $\Delta = 0.0$ represents no impact of focus. Given $n = 16$, $\Delta = 0.1$ implies a maximum reduction of 9.6% in mean LOS due to focus (which occurs in a ward dedicated to a single specialty). This is consistent with reductions observed in mean LOSs of specialties following the creation of specialized wings in the study of Best et al. (2015). The best configurations obtained from our methodology are illustrated in Figure 8 for $\Delta = 0.0$ and $\Delta = 0.1$. These figures show that the best-found configuration is a COF configuration with either value of Δ as there exists at least one cluster with both dedicated and overflow wards.

For $\Delta = 0.0$, as illustrated in panel (a) of Figure 8, the best-found configuration has 4 clusters, including one cluster for each of GYN and RSP specialties, one cluster for all surgical specialties, and one cluster for all medical specialties. The cluster involving all the surgical (medical) specialties, i.e., cluster 3 (cluster 4), has a total of 121 (214) beds allocated to it, 55 (71) of which are overflow beds. Overall, there exists a total of 268 dedicated beds and 126 overflow beds in the best-found configuration with $\Delta = 0.0$. We also observe that ORM, ENT, URO, OPL, and END specialties have not been allocated a dedicated ward and are treated in overflow wards of their clusters.

When Δ increases to 0.1, as illustrated in panel (b) of Figure 8, a new cluster is formed by separating PAL from the cluster of medical specialties and dedicating 3 beds to it. This is because this specialty has a very small arrival rate but a very long LOS, which increases further in overflow wards when $\Delta = 0.1$. Our model separates these infrequent but long-staying patients to minimize their negative impact on frequent but short-staying patients, which is consistent with the findings from the pooling literature. We also observe in panel (b) of Figure 8 that URO is allocated a dedicated ward with 13 beds. Further, the number of overflow beds in clusters 3 and 4 decreases by 17 each compared to the scenario with $\Delta = 0.0$.

To compare the best configurations obtained from our methodology with the current configuration, we evaluate the performance of each configuration using a discrete-event simulation model. For the configurations generated by our methodology, we simulate each cluster, independently from other clusters, using the specialty parameters given in Table 5. For the current configuration, we simulate each ward, independently from other wards, using the wards' parameters as given in Table 6 instead of those of specialties. This is because there exists a significant level of patient outlying in the current configuration as demonstrated in Section 6.1. Hence, simulating the wards with their overall arrival rates and LOSs would enable us to represent

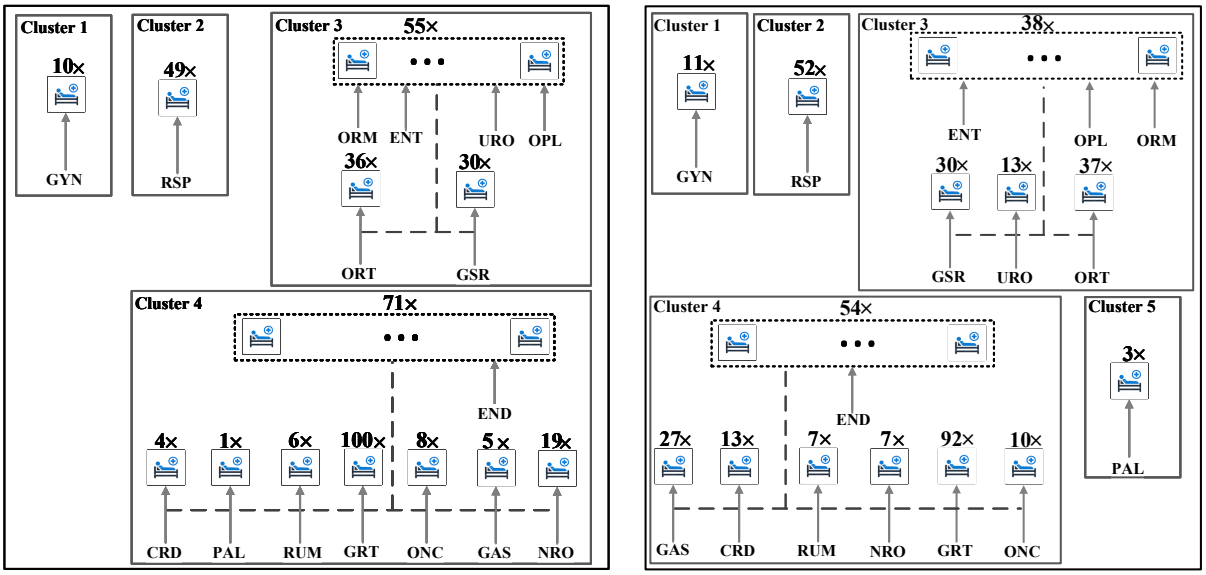


Figure 8: The best-found configurations under IWTT assumption with $\Delta = 0.0$ (a), and $\Delta = 0.1$ (b).

the current situation more accurately without having to capture the complex flows of patients between wards. However, the ward parameters given in Table 6 yield a traffic intensity of 87.6% which is slightly higher than the traffic intensity of specialties, 86.7%, as reported earlier. To have a fair comparison, we multiply all ward arrival rates given in Table 6 by 0.991 so that both systems have exactly the same workload.

Wards	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	E	EAU
Arrival Rate	3.40	2.52	2.47	2.31	3.21	3.90	3.96	1.29	4.69	5.36	9.80	4.21	4.58	0.80	41.00
LOS Mean	8.11	11.51	5.61	12.52	6.72	6.74	7.20	18.02	4.78	3.00	2.68	6.18	5.78	5.27	0.60
LOS CV	2.11	3.11	1.43	2.34	1.50	1.25	3.40	2.10	1.50	2.30	1.80	2.10	1.17	3.05	2.45

Table 6: Arrival rates and LOS mean and CV for inpatient wards at RSCH. Time unit is one day.

The inter-arrival times and LOSs are assumed to be Exponential and Log-Normal, respectively, in all of our simulation experiments. We replicate each simulation model 10 times, with each replication running for $100,000$ days. The simulation provides estimates of mean numbers of patients waiting in the queues and mean numbers of occupied dedicated and overflow beds, using which cost functions are evaluated.

We obtain the mean total daily costs of £$11,840$ (£$7,840$) for our best-found configuration with $\Delta = 0.0$ ($\Delta = 0.1$) as compared to £$85,665$ for the existing configuration. These figures indicate significant cost savings with our proposed configurations. These savings result from a 96% (99%) reduction in the number of patients waiting for admission at the expense of a 39% (19%) increase in staffing costs for $\Delta = 0.0$ ($\Delta = 0.1$); see Table 7 for queue size and costing figures. The substantial reduction in the numbers waiting for admission highlight the

inefficiency of the current bed configuration, which was also evident in the large numbers of outlying patients as reported in Section 6.1.

It is important to note that the saving figures presented above are contingent upon the assumption that each additional day a patient spends in the admission queue incurs costs, both for the patient and society at large, that are tenfold higher than the daily expense of employing a nurse. Additionally, staffing costs for certain specialties may exceed the average values incorporated into our model. Hence, the savings would fluctuate depending on the precise cost estimations applied. However, the central takeaway remains clear: the implementation of our configurations has the potential to significantly reduce admission queues. Achieving this does not necessitate investments in additional beds but rather in nursing staff (and cross-training them).

Configuration	Average Total Queue Size	Average Total Nursing Cost	Average Total Cost
Current configuration	77.19	6137	85665
New configuration with $\Delta = 0.0$	3.24	8500	11840
New configuration with $\Delta = 0.1$	0.52	7310	7840

Table 7: Simulation-based comparisons between best-found configurations and the current configuration.

We now compare the performance of our best-found configurations (presented in Figure 8) with those of fully dedicated (DED), fully flexible (FLX), wing formation (WNG), and earmarking (ERM) configurations using simulation. To obtain the best allocation of beds and specialties for each of these configurations, we amend the inter- and intra-cluster allocation models in the following ways. For DED, we revise the inter-cluster allocation model so that only one specialty is allocated to each cluster. The intra-cluster allocation model is replaced with a performance evaluation routine which uses the exact results for $M/M/s$ queues to evaluate the mean total daily cost of a single-specialty cluster. A similar routine is used for WNG configuration to evaluate the mean total cost of a cluster, which may include more than one specialty. The partitioning and inter-cluster model for this configuration works in the same way as for COF. For ERM, the inter-cluster allocation model is revised so that only 4 clusters, one involving all medical specialties except RSP, one involving all surgical specialties except GYN, one involving only RSP, and one involving only GYN, are considered. The intra-cluster allocation model for this configuration works in exactly the same way as for COF. For FLX, the inter-cluster model is changed in the same way as for ERM. The same performance evaluation routine used for WNG configuration is used for evaluating mean total cost of this configuration.

The savings in total, waiting, and staffing costs obtained from using the best-found configurations as compared to the other configurations are illustrated in Table 8. The results in this table suggest that WNG, FLX, ERM, and DED (ERM, WNG, FLX, and DED) rank 2nd, 3rd,

4th, and 5th, respectively, in terms of mean total daily cost following our best-found configuration, i.e., the COF configuration, for $\Delta = 0.0$ ($\Delta = 0.1$). DED has the lowest staffing cost but also the highest waiting cost for both values of Δ. WNG and FLX, on the other hand, have the joint-lowest waiting cost but also the highest staffing costs for $\Delta = 0.0$. For $\Delta = 0.1$, COF has the lowest waiting cost and second-lowest staffing cost following DED.

Scenario	Configuration	Total Cost	Staffing Cost	Waiting Cost
$\Delta = 0.0$	DED	60.6%	−25.7%	85.3%
	ERM	10.7%	1.8%	27.4%
	FLX	6.5%	20.7%	−72.3%
	WNG	6.4%	20.6%	−72.3%
$\Delta = 0.1$	DED	56.3%	−15.2%	95.5%
	FLX	31.1%	29.8%	44.6%
	WNG	25.4%	8%	79.5%
	ERM	11%	3.1%	58.2%

Table 8: The savings obtained from the best-found configuration (COF) as compared to other configurations under IWTT assumption.

6.4. Best-Found Configurations with FWTT Assumption

We run our methodology with $\gamma^{-1} \in \{1, 15, 30\}$ days, assuming the cost of a patient abandoning the queue, c_a, is £1030 per patient. Figure 9 depicts the best-found configurations obtained for $\Delta = 0.0$ and $\Delta = 0.1$, with $\gamma^{-1} = 1$ day. This figure shows that COF remains the lowest cost configuration returned by our methodology. In comparison with best configurations obtained under IWTT, we observe that medical and surgical specialties are each split into two clusters for both focus scenarios, increasing the number of clusters to 7. The total number of overflow beds also reduces (rises) to 79 (93) for $\Delta = 0.0$ ($\Delta = 0.1$) as compared to the corresponding scenarios under IWTT. Further, no bed is allocated to PAL given its small arrival rate and long LOS.

In Table 9, we compare the performance of our best-found configuration with the other configurations under FWTT assumption with $\gamma^{-1} = 1$ day using simulation. The results in the table show that WNG and ERM are the 2nd-best configurations for $\Delta = 0.0$ and $\Delta = 0.1$, respectively, in terms of mean total daily cost. On the other hand, FLX is the worst configuration for both focus scenarios. DED provides the lowest staffing costs and the highest abandonment costs for both values of Δ. ERM, on the other hand, produces the lowest abandonment cost with $\Delta = 0.0$, and the same abandonment cost as COF with $\Delta = 0.1$. The results for $\gamma^{-1} = 15$ and $\gamma^{-1} = 30$ days are presented and discussed in Section E of the e-companion. The main observation is that COF remains the lowest cost configuration for $\Delta = 0.0$ and $\Delta = 0.1$ when $\gamma^{-1} = 15$ days, and for $\Delta = 0.10$ when $\gamma^{-1} = 30$ days. When $\gamma^{-1} = 30$ and $\Delta = 0.0$, however, WNG becomes the lowest cost configuration. Overall, COF yields the lowest cost in seven out of

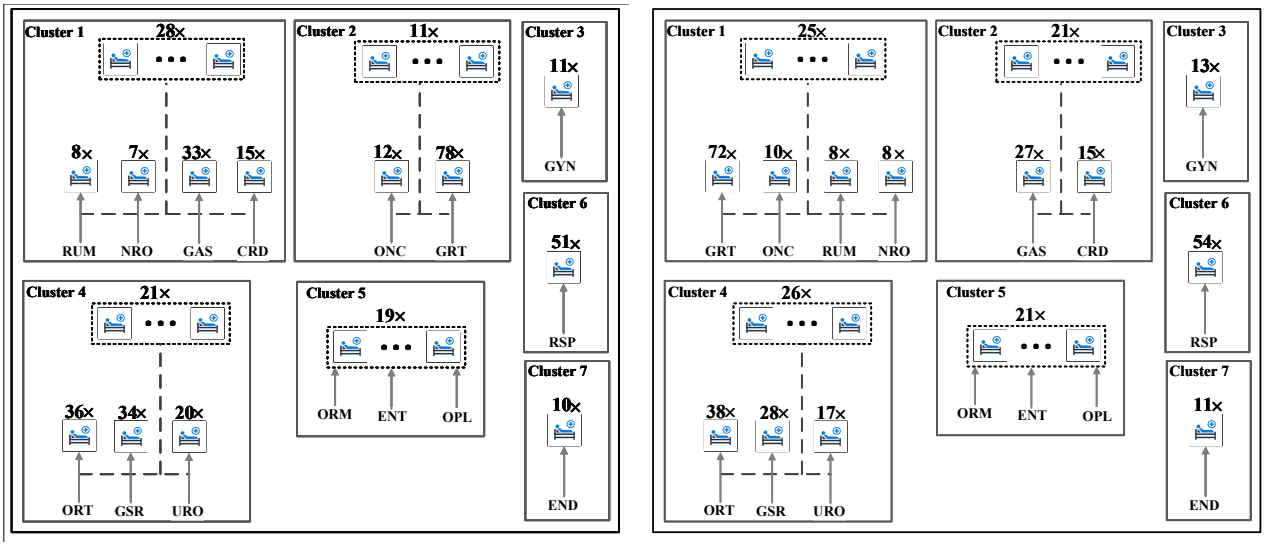


Figure 9: The best-found configurations under FWTT for $\gamma^{-1} = 1$ day with $\Delta = 0.0$ (a), and $\Delta = 0.1$ (b).

eight scenarios we considered. For all these seven scenarios, the second lowest cost is delivered by WNG when $\Delta = 0.0$, and ERM when $\Delta = 0.1$. A sensitivity analysis is conducted in Section F of the e-companion on the impact of cost parameters on the number of clusters formed as well as the number of overflow beds in the best configurations produced by our methodology.

Scenario	Configuration	Total Cost	Staffing Cost	Abandonment Cost
$\Delta = 0.0$	FLX	26.5%	31.6%	0%
	DED	9.5%	−4.2%	64.9%
	ERM	4.0%	7.6%	−82.1%
	WNG	3.5%	1.0%	26.4%
$\Delta = 0.1$	FLX	33.3%	34.3%	0%
	DED	14.9%	−1.5%	84.5%
	WNG	9.4%	4.0%	64.4%
	ERM	1.4%	4.2%	0%

Table 9: The savings obtained from the best-found configuration (COF) as compared to other configurations under FWTT assumption with $\gamma^{-1} = 1$ day.

7. Conclusions

We proposed the most versatile methodology to date for finding a low cost configuration of inpatient services in a hospital. Given a total number of beds, a set of specialties, and a specific finite or infinite waiting time threshold, this methodology uses two search algorithms combined with novel performance evaluation approximations to find a configuration with a low total mean daily cost, considering all major configurations proposed in the literature. We demonstrated how our methodology can be modified so that the best partitioning and bed allocation of a given configuration other than COF, i.e., dedicated, flexible, wing formation, or

earmarking, is obtained. This is useful for situations where practical constraints limit the range of configurations that can be implemented in the hospital.

Using inpatient data from a large hospital, we illustrated how our reconfiguration methodology can reduce the bed pressure on hospital without expanding the bed base, and with a moderate increase in daily nursing costs. Our simulation experiments suggest that the savings obtained from our methodology could be significant, and that the COF configuration is likely to yield the lowest cost in most situations, followed by WNG (ERM) when the impact of focus is negligible (significant). While our methodology may require several hours to complete when a large number of specialties and/or beds are involved, this time frame is generally not a significant concern, given that the reconfiguration process is typically conducted once every few years.

To implement the configurations recommended by our methodology, the hospital may need to fill the shortages in size and skill-set of its nursing teams, compared to the requirements of the proposed configurations, through a combination of cross-training of existing nurses and recruitment of new nurses. Furthermore, adjustments to the layout of beds and wards will be required. It is important to note that there may be additional considerations, including privacy concerns, clinical requirements, and spatial constraints, that need to be factored in. Fortunately, many of these constraints can be seamlessly integrated into our methodology, as demonstrated in our case study.

We proposed novel performance evaluation approximations for overflow delay and overflow abandonment systems with potentially different mean service times in dedicated and overflow servers. These approximations produce practically accurate results in a short time. In addition to inpatient bed planning, these can be applied to other hierarchical systems such as those observed in telecommunication and computer networks

In addition to developing algorithms, we contributed to inpatient data analysis in the following ways. First, we proposed identifying the specialty of each episode of care based on the HRG codes used for costing purposes. This provides a more accurate representation of patients' clinical needs than the specialty of the consultant in charge of the episode, which could be misleading in case of patient outlying. Second, we proposed linking together the episodes of care within a hospital spell that have the same specialty to create what we referred to as specialty spells. The percentage of outlying specialty spells can then be evaluated for each specialty as well as the hospital as a whole. This would give an indication of the frequency at which outlying occurs in the hospital. Third, we proposed measuring the workload contribution of outlying patients in a ward or the entire hospital by adding up the lengths of outlying bed visits and dividing it

by the sum of all bed visits in that ward or the entire hospital. The frequency and workload measures together provide an accurate picture of the extent of outlying in the hospital. Finally, we proposed measuring the number of ward changes within a specialty spell, and assessing its impact on mean LOS of each specialty.

Improving the accuracy of our performance evaluation approximations, as well as the accuracy and speed of the intra-cluster bed allocation heuristic, are clear directions for future research. Such improvements would improve the overall accuracy and speed and thus the applicability of our methodology.

References

Akcali, E., Coˆté, M. J., & Lin, C. (2006). A network flow approach to optimizing hospital bed capacity decisions. *Health Care Management Science*, *9*, 391–404.

Arabzadeh, B. (2022). *Reconfiguration of inpatient services to reduce bed pressure in hospitals*. PhD dissertation City, University of London. URL: `https://openaccess.city.ac.uk/id/eprint/28994/1/` Accessed: 2023-10-06.

Bekker, R., Koole, G., & Roubos, D. (2016). Flexible bed allocations for hospital wards. *Health Care Management Science*, (pp. 1–14).

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

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

Chevalier, P., & Van den Schrieck, J.-C. (2009). Computing the performance of multiclass queueing systems based on equivalent loss systems. Unpublished Manuscript.

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

Conroy, S., & Dowsing, T. (2012). What should we do about hospital readmissions? *age and aging*, *41*, 702–704.

Cooper, R. B. (1990). Queueing theory. *Handbooks in Operations Research and Management Science*, *2*, 469–518.

Dumas, M. B. (1985). Hospital bed utilization: an implemented simulation approach to adjusting and maintaining appropriate levels. *Health services research*, *20*, 43.

Ewbank, L., Thompson, J., Mckenna, H., & Siva, A. (2020). NHS hospital bed num-

bers: past, present, future. URL: https://www.kingsfund.org.uk/publications/nhs-hospital-bed-numbers Accessed: 2023-10-06.

Franx, G. J., Koole, G., & Pot, A. (2006). Approximating multi-skill blocking systems by HyperExponential Decomposition. *Performance Evaluation*, *63*, 799–824.

Fredericks, A. A. (1980). Congestion in Blocking Systems-A Simple Approximation Technique. *Bell System Technical Journal*, *59*, 805–827.

Goldstein, N. D., Ingraham, B. C., Eppes, S. C., Drees, M., & Paul, D. A. (2017). Assessing occupancy and its relation to healthcare-associated infections. *infection control & hospital epidemiology*, *38*, 112–114.

Green, L. V., & Nguyen, V. (2001). Strategies for cutting hospital beds: the impact on patient service. *Health services research*, *36*, 421–442.

Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2011). *Fundamentals of Queueing Theory* volume 627. John Wiley & Sons.

Izady, N., & Mohamed, I. (2021). A clustered overflow configuration of inpatient beds in hospitals. *Manufacturing & Service Operations Management*, *23*, 139–154.

Jagerman, D. L. (1974). Some Properties of the Erlang Loss Function. *Bell System Technical Journal*, *53*, 525–551.

Jones, S., Moulton, C., Swift, S., Molyneux, P., Black, S., Mason, N., Oakley, R., & Mann, C. (2022). Association between delays to patient admission from the emergency department and all-cause 30-day mortality. *Emergency Medicine Journal*, *39*.

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

Kaier, K., Mutters, N., & Frank, U. (2012). Bed occupancy rates and hospital-acquired infections—should beds be kept empty? *Clinical microbiology and infection*, *18*, 941–945.

KC, D. S., & Terwiesch, C. (2011). The Effects of Focus on Performance: Evidence from California Hospitals. *SSRN Electronic Journal*, .

Li, X., Beullens, P., Jones, D., & Tamiz, M. (2009). An integrated queuing and multi-objective bed allocation model with application to a hospital in china. *Journal of the Operational Research Society*, *60*, 330–338.

Lim, J. M., Song, H., & Yang, J. (2021). The spillover effects of capacity pooling in hospitals. URL: https://ssrn.com/abstract=3800351 Accessed: 2023-10-06.

Maguire, D. (2015). Premature discharge: is going home early really a Christmas gift? URL: https://www.kingsfund.org.uk/blog/2015/12/premature-discharge-from-hospital Accessed: 2023-10-06.

Malaki, S., Izady, N., & de Menezes, L. M. (2023). A framework for optimal recruitment of temporary and permanent healthcare workers in highly uncertain environments. *European Journal of Operational Research*, *308*, 768–781.

Mateen, B. A., Wilde, H., Dennis, J. M., Duncan, A., Thomas, N., McGovern, A., Denaxas, S., Keeling, M., & Vollmer, S. (2021). Hospital bed capacity and usage across secondary healthcare providers in england during the first wave of the covid-19 pandemic: a descriptive analysis. *BMJ Open*, *11*.

Moiseev, S. (2011). Universal derivative-free optimization method with quadratic convergence. URL: `https://arxiv.org/abs/1102.1347` Accessed: 2023-10-06.

NHS Data Model and Dictionary (2021). Healthcare resource group. URL: `https://datadictionary.nhs.uk/supporting_information/healthcare_resource_group.html` Accessed: 2023-10-06.

Olshaker, J. S., & Rathlev, N. K. (2006). Emergency Department overcrowding and ambulance diversion: the impact and potential solutions of extended boarding of admitted patients in the Emergency Department. *The Journal of emergency medicine*, *30*, 351–6.

Propper, C., Stoye, G., & Zaranko, B. (2020). The wider impacts of the coronavirus pandemic on the NHS. *Fiscal Studies*, *41*, 345–356.

Robertson, R., Wenzel, L., Thopson, J., & Charles, A. (2017). Understanding NHS finanical pressure. How are they affacting patient care? URL: `https://www.kingsfund.org.uk/publications/understanding-nhs-financial-pressures` Accessed: 2023-10-06.

Shi, P., Chou, M. C., Dai, J. G., Ding, D., & Sim, J. (2016). Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time. *Management Science*, *62*, 1–28.

Stowell, A., Claret, P.-G., Sebbane, M., Bobbia, X., Boyard, C., Genre Grandpierre, R., Moreau, A., & de La Coussaye, J.-E. (2013). Hospital out-lying through lack of beds and its impact on care and patient outcome. *Scandinavian journal of trauma, resuscitation and emergency medicine*, *21*.

Wu, X., Li, J., & Chu, C.-H. (2019). Modeling multi-stage healthcare systems with service interactions under blocking for bed allocation. *European Journal of Operational Research*, *278*, 927–941.

Zhang, Z. (2010). *Call centres with balking and abandonment: from queueing to queueing network models*. PhD dissertation University of Saskatchewan. URL: `https://harvest.usask.ca/bitstream/handle/10388/etd-06222010-103338/ZhidongZhangThesisFinal.pdf` Accessed: 2023-10-06.