Journal of
Health & Medical Informatics

Lee et al., J Health Med Informat 2018, 9:2
DOI: 10.4172/2157-7420.1000308

Research Article Open Access

Strategies for Inpatient Bed Management

Eva K. Lee*, Zixing Wang and Andriy Shapoval

Center for Operations Research in Medicine and Health Care, NSF I/UCRC Center for Health Organization Transformation, Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Abstract

We consider the problem of partitioning clinical services in hospitals into groups with the goal of efficiently allocating existing inpatient beds. At the strategic level, there are two major possibilities: pooling versus focusing. Pooling the bed capacity allows one to achieve an overall high occupancy level for a fixed number of beds. On the other hand, focusing by dividing the capacity into groups with restricted access may offer increased efficiency and better resource utilization. We first derive a 2-stage approach to address the 3-fold problem: 1) how many groups of services to form; 2) how many beds to allocate to each group; and 3) how to partition services among the groups. Specifically, Stage 1 uses cluster analysis utilizing the similarity principle for possible advantages of economies of scale. Stage 2 then incorporates utility/benefit functions to optimize the partitions and allocation of beds. To contrast the results, we combine the two stages into a single mixed integer nonlinear program. Three full-scale examples demonstrate the flexibility and diverse application of our framework with managerial insights for different utility optimization goals and queuing systems. The resulting modeling framework is not computationally sensitive to the number of beds, making it more practical for usage by any hospitals.

Keywords: Pooling; Resource allocation; Queuing systems; Bed capacity management; Clustering; optimization; Nonlinear mixed integer program

Introduction

Healthcare expenditures continue to rise around the world. Developed countries see their healthcare costs grow faster than the gross domestic product (GDP). Increasing demand for healthcare from an aging population is supplemented by inefficient use of available resources and insufficient supply of medical facilities and personnel.

Hospitals play an important role in healthcare industry. They especially struggle with meeting the growing patient demand. Inefficiencies in hospital operations can lead to decline in quality and unnecessarily high cost of care. One essential part of these operations is related to capacity planning and management. Capacity planning activities can be classified according to time: strategic (long-term), tactical (intermediate), and operational (short-term). This includes capacity decisions for allocating equipment, rooms, personnel (especially nurses) and determining the proper number of inpatient beds to meet the changing demand.

Bed planning is complex since the actual number of occupied beds follows a stochastic process based on patient arrivals and service times. It is not a trivial task to determine the trade-offs between the requirement for meeting the peak demand versus efficient utilization of resources. All three levels of planning can be present: for example, strategic level involves defining the size of the hospital units using the number of staffed beds; tactical planning involves bed re/allocation and reservation; and operational planning covers admission issues.

Bed capacity planning relates to strategic decisions, and historically it is strongly influenced by powerful stakeholders, for example, regulators, healthcare policy makers, and insurance companies [1,2]. "Hospital bed capacity decisions have traditionally been made based on target occupancy levels-the average percentage of occupied beds. Historically, the most commonly used occupancy target has been 85%... Until recently, the number of hospital beds was regulated in most states under the Certificate of Need process, under which hospitals could not be built or expanded without state review and approval. Target occupancy levels were the major basis for these approvals [3]. "These demonstrate the necessity of taking into account external restrictions for hospitals in capacity decision making.

In the United States, a new hospital bed costs more than "$1 million, and the average cost per day for a hospital stay is thousands of dollars" [4,5]. Most hospitals are non-profit and are under constant pressure to cut costs. Their revenue is tightly regulated and depends on compensation. Their social and saving-life missions may have direct conflicts with financial realities. Quality in healthcare has dual requirements: it includes not only a set of performance characteristics for any service (for example, short waiting time), but also the medical outcome, which may affect the admission decisions. There is also the issue of accessibility, which may conflict with price and quality, and may influence many decisions.

Traditionally, hospital inpatient services are organized into care units/types, clinical services, and specialties (e.g., cardiology, gynaecology, neurology). Upon admission to the hospital each patient is assigned a diagnostic related group (DRG). There are 467 DRGs, and they are organized into 25 Major Diagnostic Categories (MDCs). The diagnoses in each MDC are associated with a particular medical specialty.

We consider hospitals that accept all types of insured and uninsured individuals. Patients are differentiated by diagnoses only, and we assume that each patient has only one diagnosis, or only one major diagnosis is in consideration.

Possibilities of consolidating services lead to a trade-off between

***Corresponding author:** Eva K Lee, Center for Operations Research in Medicine and Health Care, NSF I/UCRC Center for Health Organization Transformation, Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA, Tel: (404) 894-4962; E-mail: eva.lee@gatech.edu

Received April 01, 2018; **Accepted** April 07, 2018; **Published** April 12, 2018

Citation: Lee EK, Wang Z, Shapoval A (2018) Strategies for Inpatient Bed Management. J Health Med Informat 9: 308. doi: 10.4172/2157-7420.1000308

pooling and focusing. Pooling decisions assume fair unrestricted access for patients and allow one to gain from the economy of scale [6], while focusing ("specialist systems") may provide convenient restrictions for hospitals, effect of experience for personnel, and increased efficiency of care as shown in some empirical studies [7,8].

The problem of care unit consolidation and partitioning has tight connections with capacity sizing. Best et al. (2015), in their study of University of Chicago Medical Center and started from a known overall number of medical / surgical beds and clinical services, raised three specific questions simultaneously in a single complex decision: 1) how many groups of services to form, 2) how many beds to allocate to each group, and 3) how to partition services among the groups. Its combinatorial nature makes solutions to realistic instances with hundreds or thousands of beds computationally prohibitive. Best provided a [9] solution approach by restricting the feasible region, and solved a 300-bed problem using dynamic programming. Their model employs a Markovian queueing system for arrival, service, and abandonment, and requires exogenous parameters for financial utilities and service time. The essential point of their work is to generate financial advantages by grouping clinical services while restricting beds for "not very profitable" services.

In this paper, we propose a 2-stage framework for solving this three-fold inpatient bed capacity management problem. In this approach, we first restrict the feasible region by using similarity/dissimilarity principle from cluster analysis. This is followed by optimization of the partitions and allocation of beds. We contrast the 2-stage approach with a direct approach where the group partition process is integrated within the optimization problem. In the direct theme, the optimization searches through the entire feasible region and thus will be computationally more expansive.

Starting with n clinical services, we look for m possible groups, where minimal m=1 means "pooling" (centralized, i.e. no separate groups), and maximal m=n means totally focused "specialist" system (completely decentralized). It is easy to observe that the number of possible group configurations equals the number of partitions of the set with n elements. This is given as the Bell number Bn,

Where $B_{n+1} = \sum_{m=0}^{n} \binom{n}{m} B_m$

Hence, the number of potential partitions is exponential.

The number of all possible feasible solutions in the case of 300 beds exceeds 1030. Using simulation-based optimization approach in [9], "it would take about six months" just to evaluate all groups (more than 50,000) for a fixed instance in their heuristic. The problem is computationally intractable.

In this study, we provide a holistic, flexible, and fast computational framework to tackle this problem. We apply a decomposition scheme to manage the "curse of dimensionality" in the number of beds using the fact that the number of clinical services (and possible groups) is relatively small. The resulting system is not computationally sensitive to the number of beds, making it more practical for usage by any type of hospitals. All possible outcomes (pooling, focusing, and partial pooling) can be obtained as solutions. Choosing and checking various objective functions over the same restricted feasible set give flexibility and insight for decision making. Different queueing sub models can be used. Specifically, in Stage 1, n-2 linear optimization problems of moderate size are solved to obtain candidates for service partitioning (i.e. partial pooling modes). Using this approach, the feasibility region is reduced to a limited number. These candidate solutions along with "complete pooling" and "focusing" are input into Stage 2, where utility goals are incorporated. Utility here represents the state of being useful, profitable or beneficial. The overall number of beds is known, and an underlying queueing model thus defines the structure of the optimization problems in Stage 2.

The clustering approach allows multiple criteria of similarity in Stage 1. The Stage 1 heuristic has both advantages and drawbacks: it is simple and fast, but it does not observe an objective function (of Stage 2) and may cut too deep into the feasible region. So, it may be weaker than other heuristics that utilize information about the objective function and constraints of particular optimization problems directly. Stages 1 and 2 are bundled and allow users to compare results with other heuristics alternatives. The drawback can be compensated by the fact that the optimization problems in both stages are open to adding any customized constraints and group-candidate solutions.

In section II, we present the 2-stage and the combined approaches. Section III illustrates our approaches using three hospital cases. Discussions and conclusions are presented in Section IV.

Methodology and Mathematical Models

The 2-stage scheme

In the 2-stage approach, we solve the bed management problem based on i) cluster analysis and ii) solution to a sequence of queueing-model-based mixed integer nonlinear programs (MINLP). Specifically, Stage 1 clusters the services based on similarities and determines a potential number of groups with a list of partition-candidates of the services. Stage 2 optimizes the actual bed allocation over the feasible set returned by Stage 1 guided by some desirable objective function(s).

Stage 1: Clustering specialties to establish a limited number of group-service candidates Let n ne the number of entities (services) each represented by a numeric vector, qi, i=1, …, n. Given m groups (m<n), Sj, j=1, …, m, we determine the group membership of each service by minimizing the within-group sum of squares, f, where

$$f = \sum_{j=1}^{m} \sum_{i \in s_j} \left\| q_i - c_j \right\|^2$$

Here $ccjj$ is the center of $SSjj$, i.e. $c_j = \frac{1}{|s_j|} \sum_{i \in s_j} q_i$. When partitioning the entities, we adopt the "condition with a leading entity." That is, in an optimal solution one entity, the leader of the group, should have the distance with any out-of-group entity no less than the distance between the leader and any entity inside its group. This problem is first investigated by Rao [10]. Although the condition is weak, it offers an attractive 0-1 integer programming formulation. Let binary variable $yypp$=1 if group p is formed, 0 otherwise. We establish a restricted set partition problem:

$$z = \min \left\{ h^T y \middle| AY = 1, 1^T Y \le m, y \in \{0,1\} \right\} \quad (1)$$

Where A is an n by m 0-1 incidence matrix, $aaiii$ =1 means entity i belongs to group j, and 1 is a column vector of all ones with appropriate dimensions. For convenience, we assume the distances are distinct. This allows us to arrange all distances with any entity i as a leader in an increasing order. Thus, for each leading entity, there are n-1 possible groups, not counting the group of all entities. Because there are n entities and each of them can be a candidate for leadership, thus there are at most (n-1)+1 groups. In this case, the matrix A has at most (n-1)+1columns. The problem size can further be reduced by deleting identical groups.

The vector h defines the within-group sum of squares. In particular, h*pp* denotes the within-group sum of squares of group p. Assume that group p contains the entities i through j, then

$$h_p = h_i^j = \sum_{k=1}^{j}(q_k - q_{ij})^2, with q_{ij} = \frac{\sum_{k=i}^{j} q_k}{j-i+1}$$ (the group mean),

and *qqii* the numerical characteristics of entity *i*. Note that h*iiii*=0. With our approach, there are at most (*n*-1)/2 nonzero values in h. The constraint $1^T y \le m$ ensures that there are at most m total number of groups. Our formulation allows us to manually assign some variables to 0 (without getting infeasibility), if they are not desirable to be in the solutions.

Solving the sequence of restricted set partition problems for m=2,…, n-1 establishes a limited number of candidates for the number of groups along with partition-candidates of services.

Stage 1 returns sets S^m of candidate assignment vectors Si of specialties: $S^m = \{S_1^m, S_2^m, ..., S_m^m\}$, with $U_i S_i^m = S_m, 1 \le m \le n$. The superscript denotes the number of groups, and the subscript is the index of groups. Two possible solutions are trivial: "the complete pooling" mode $S_1 = \{S_1^1\}, where S_1^1 = (1, 2, ...n)$, and "the complete focused" mode $S^n = \{S_1^n, S_2^n, ..., S_n^n\}$, where each assignment vector contains exactly one specialty, $S_i^n = i$, *i*=1,2,…*nn*. For each *mm*=2,3,… *nn*-1. Only one instance-candidate set *SSmm* with fixed entries is kept. Consider the example with three specialties: for *m*=1, $S_1 = \{S_1^1\}$, where S_1=(cardiology, gynaecology, neurology) as one group; for m=3,S^3= $\{S_1^3, S_2^3, S_3^3\}$, where S^1_3= cardiology, S^2_3=gynaecology, S^3_3=neurology; for m=2, $S^2 = \{S_1^2, S_2^2\}$ is a solution of the restricted set partition problem, say S^1_2= (cardiology, gynaecology), S^2_2= neurology, which outperforms the other configurations consisting of two groups and is kept for Stage 2.

Using the 'leading entity' approach, the feasible region is greatly restricted. Thus the resulting set partitioning problem in Stage 1 is relatively small and can be easily solved to optimality. We contrast the leading entity approach with k-means++ clustering by running it on the data set 100 times and select the best results [11].

Stage 2: Optimizing bed allocation and finalizing partitioning services into groups.

Recall n denote the number of clinical services. In Stage 2, the main problem initially is transformed to a sequence of n optimization problems with respect to the fixed sets S1, S2, …, Sn to find the optimal objective function values z1, z2, …zn. Let N denote the number of groups (i.e. the fixed value of m in Stage 1). Then each of the n problems contains constraints from the underlying queueing system along with the bed allocation constraints $\sum_{i=1}^{N} c_i = C, c_i \ge 1, i = 1,$ 2,…N, where C denotes the total number of beds. The best optimal values zN* among these n problems gives the optimal bed allocation vector $c^* = (c_1, c_2, ...c_N)$, and solves the main problem with optimal values m=N*, c*, and SN*. Thus N* answers how many groups of services to form, c* reports the number of beds allocated to each group, and SN* shows the partition of services.

Thus, Stage 2 i) manages the sub problem of bed allocation by taking into account the utility goals; and ii) finalizes partitioning services into groups. A sequence of mixed integer nonlinear programs (MINLP) is involved to optimize a particular objective function. Queueing sub models are essential parts of the Stage 2 and they are defined by the problem structure.

An illustrative theoretical-computational framework for stage 2

Let λ be the arrival rate, τ be the mean service time, $\rho = \frac{\lambda\tau}{c} < 1$ be the utilization and $v2$ be the service time squared coefficient of variation (SCV).

Assume that a hospital with C beds has the information about its patients for all clinical/surgical services i=1, …, n in the form of parameters λ_i, τ_i and v^i_2. Let $\Lambda = \sum^n_{i=1}\lambda_i$. We use M/G/c queue to model the patient accommodation process here, i.e. the patient arrival flow for the i^{th} service is a Poisson process with arrival rate λ_i, and the service time is an arbitrary distribution with mean τ_i and SCV v^i_2. The hospital can operate stably in the pooling mode, i.e. for all services,

$$\rho_{total} = \frac{\sum_{all}\lambda_i\tau_i}{c} < 1.$$

Suppose that the objective of the hospital is to minimize the total waiting time among all patients. For pool *SSkk* (here we omit the superscript N for simplification), the expected waiting time among all patients in this pool can be written as

$$W_K = \frac{1 = v_K^2}{2\sum_{i\in S_k}\lambda_i} \frac{\rho_k^{\sqrt{2(c_k+1)}}}{1-\rho_k}$$

Where

$$\rho_k = \frac{\sum_{i\in S_k}\lambda_i\tau_i}{c_k}$$

The results above were first presented in [12]. And group SCV is given by the formula in [13]:

$$v_k^2 = \frac{\left(\sum_{i\in S_k}\lambda_i\right)\left(\sum_{i\in s_k}\lambda_i\tau_i^2\left(v_i^2+1\right)\right)}{\left(\sum_{i\in s_k}\lambda_i\tau_i\right)^2} - 1.$$

Suppose that the group number is N, then the total waiting time among all groups could be written as: $W = \frac{\sum_{k=1}^{N}\lambda_k W_K}{\sum_{k=1}^{N}\lambda_k}$

$$= \frac{\sum_{k=1}^{N}\lambda_k W_k}{\Lambda}$$

where $\lambda_k = \sum_{i\in S_k}\lambda_i$.

We can compare the optimal objective function values of these sequences of nonlinear mixed integer programming instances. In particular, the MINLP for Stage 2 can be formulated as:

$$\min \frac{1}{\Lambda}\sum_{k=1}^{N}\frac{\left(1+v_k^2\rho_k^{\sqrt{2(c_k+1)}}\right)}{2\left(1-\rho_k\right)}$$

subject to $\frac{\sum_{i\in S_K}\lambda_i\tau_i}{c_k}\le 1-\in \quad \forall k=1,...,N$ (2)

$$\sum_{k=1}^{N} c_k - C$$

where ε is a very small positive number, e.g., e-6. The decision variables in this formulation are *cckk*. All the parameters can be calculated by the equations above after Stage 1 is completed. Suppose that we have n clinical services in a hospital, then we need to solve (2) for N=1, …, n to determine the optimal solutions.

The idea here also applies to other queueing models. Suppose that patients would leave the hospital (instead of joining a queue) if they find that all the beds are occupied. These patients are called 'blocked' patients. An M/G/c/c queue could also be applied here.

Suppose the hospital wants to minimize the maximal blocking probability among all groups. Let *BBkk* denote the blocking probability for group *SSkk*, and offered load $a_k=\lambda_k\tau_k$. We use keep the same notation as in the M/G/c queue case. The group offered load can be calculated as:

$$a_k = \sum_{i\in S_k}\lambda_i * \frac{\sum_{i\in S_k}\lambda_i\tau_i}{\sum_{i\in S_k}\lambda_i} = \sum_{i\in Sk}\lambda_I\tau_i$$

Using the continuous Markov chain, we can see the blocking probability is given by the Erlang-B formula:

$$B_k = \frac{(a_k)^{c_k}/C_k!}{\sum_{i=0}^{c_k}(a_k)^i/i!}. \quad (3)$$

To facilitate the solution process, this high-degree polynomial formula can be approximated by [14]:

$$B_k = \frac{a_k - c_k - 2\rho_k + \sqrt{(a_k-c_k)^2+4a_k}}{2a_k - 2\rho_k} \qquad k=1,..., N\cdot$$

The Stage 2 formulation for M/G/c/c queue is given by:

$$\underset{subject\ to}{Min}\ B_k \le B_w \qquad k=1,...,N$$

$$B_k = \frac{a_k - c_k - 2\rho_k + \sqrt{(a_k-c_k)^2+4a_k}}{2a_k - 2\rho_k} \qquad k=1,..., N \quad (4)$$

$$\sum_{k=1}^{N} c_k = C$$

$$a_k(1-B_k) \le C_k$$

$$0 \le B_k \le 1$$

$$C_k \in Z^+, B_w, B_k \in R, \forall k = 1,..., N$$

Where *BBww* represents the maximal blocking probability among all groups. Note that the stability constraint $\frac{\sum_{i\in}\lambda_i\tau_i}{c_k}<1$ from (2) is replaced by $a_k(1-B_k)\le c_k$ since M/G/c/c queue has finite state space and thus does not have stability requirement; but normally the hospitals would prefer the unblocked offered load to be lower than the bed capacity.

Section III illustrates the 2-stage approach in detail for three examples.

The Direct approach

The 2-stage scheme attempts to reduce the computational effort by restricting the feasible space. We will measure its solution quality by contrasting it to the one obtained via a direct mixed integer nonlinear programming approach.

Consider the M/G/c model, let x_{ij}=1when service j is assigned to group i, and 0 otherwise, *i*=1,…„=1,…,*n*.

Then the group offered load *AAii* for group *SSii* is given by $A_i = \sum_{j-1}^{n} a_j x_{ij}$. And the group SCV V^i_2 is:

$$\nu_i^2 = \frac{\left(\sum_{j=1}^{n}\lambda_j x_{ij}\right)\left(\sum_{j=1}^{n}\lambda_j\tau_j^2\left(\nu_j^2+1\right)x_{ij}\right)}{(a_i)^2}$$

The MINLP for the M/G/c model can be formulated as:

$$\min \frac{1}{\Lambda}\sum_{i-1}^{N}\frac{\left(1+V_i^2\right)\rho_i^{\sqrt{2(c_i+1)}}}{2(1-\rho_i)}\ 1$$

Subject to: $\sum_{j=1}^{n} x_{ij} \ge 1 \qquad \forall i = 1,..., N$

$$\sum_{i=1}^{N} x_{ij} = 1 \qquad \forall i = 1,..., n$$

$$A_i = \sum_{j=1}^{n} a_j x_{ij} \qquad \forall i = 1,..., N \quad (5)$$

$$V_i^2 = \frac{\left(\sum_{j=1}^{n}\lambda_j x_{ij}\right)\left(\sum_{j=1}^{n}\lambda_j\tau_j^2\left(\nu_j^2+1\right)x_{ij}\right)}{(A_i)^2} - \forall i = 1,..., N$$

ρρiiccii=AAii ∀ii=1,…,

$$\rho_i c_i = A_i \qquad \forall i = 1,..., N$$

$$0 \le \rho_i \le 1 \qquad \forall i = 1,..., N$$

The first constraint ensures that every group has at least one service. The second constraint assigns each service to exactly one group. Compared to the 2-stage scheme, this formulation adds *NN* 0/1 decision variables *xxiii* and 2*NN* derived variables *AAii* and *VVii*2. By nature, MINLP is difficult to solve. This is harder than the optimization in Stage 2 since it searchers over a much larger feasible region. From the examples below, we can see that in practice due to the limitation of non-convex optimization solution tools, this formulation may not generate better solutions than the 2-stage scheme.

For M/G/c/c model, the MINLP formulation is similar:

Min *BBww*

subject to *BBii*≤*BBww ii*=1,…,*NN*

Σ*xxiii* ≥1*nnjj*=1 *∀ii*=1,..,

$$\sum_{i=1}^{N} x_{ij} = 1 \qquad \forall j = 1,..., N$$

$$\sum_{i=1}^{N} x_{ij} = 1 \qquad \forall j = 1,..., n$$

$$A_i = \sum_{j=1}^{n} a_j x_{ij} \qquad \forall i = 1,..., N$$

$$B_i = \frac{A_i - c_i - 2\rho_i + \sqrt{(A_i - c_i)^2 + 4A_i}}{2A_i - 2\rho_i}, \qquad \forall i = 1,\ldots,N$$

$$A_i(1 - B_i) \le c_i \qquad \forall i = 1,\ldots,N$$

$$\rho_i c_i = A_i$$

$$\sum_{i-1}^{N} c_i = C$$

$$0 \le B_i \le 1, \qquad \forall i = 1,\ldots,N$$

$$c_k \in Z^+, B_w, B_i, \rho_i, A_i \in R, \forall i = 1,\ldots,N$$

Application of our Framework

We use three real-world examples to demonstrate the two modeling frameworks. For each approach, we describe the necessary ingredients and steps to setup the model and the solution approach. Specifically, for the 2-stage framework, this includes 1) choosing the similarity criteria and solving the restricted set partition sub problems; 2) formulating the utility goals in terms of service quality or finance, and 3) solving the MINLPs based on the queuing sub models. For the direct approach, we show the solution of the MINLP.

All the MINLP are solved using LINGO 17.0 solver, 104 multi start are applied to each program for better performance. The solver is based on generalized reduced gradient algorithm for each start point. Successive linear programming is also used where applicable.

Example 1. An Urban U.S. hospital

Consider an urban hospital with 16 departments/services. Figure 1 shows the normalized bed demand $\frac{\lambda_i \tau_i}{\sum_{alli} \lambda_i \tau_i}$ (the columns) and daily utility (i.e., the normalized return/benefits $\frac{\frac{u_i}{\tau_i}}{\sum_{alli} \frac{u_i}{\tau_i}}$ for the hospital from serving these patients for service i (the curve with markers) for the 16 services. Here, ui is the expected utility gain for service i. qualitatively; this example is similar to the example in [9]. It shows a standard misbalance between popular services (with high demand, e.g. general medicine) versus the most profitable ones (e.g. different types of surgery). This figure describes situations commonly faced by many hospitals in the United States. Intuitively, it may be advantageous to group services with large demand and low utility so as to limit the number of beds allocated to them. The respective utilization constraints (i.e. ρ <1 in M/G/c queueing) have to be included explicitly.

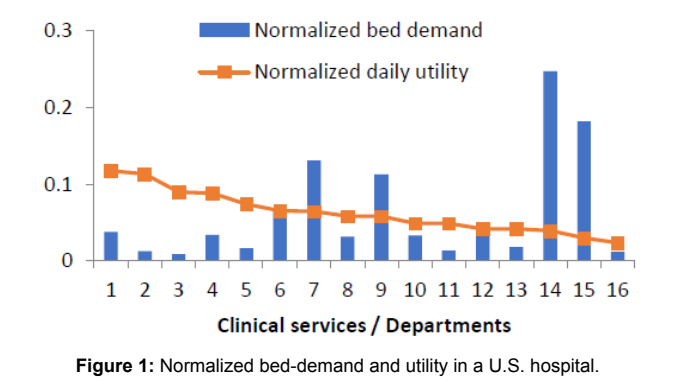

Figure 1: Normalized bed-demand and utility in a U.S. hospital.

In the 2-stage framework, we consider these two characteristics (normalized bed demand and daily utility) and test the similarity in the two-dimensional space.

Applying these data to the restricted set partition problems (1), we first note that matrix A has 140 columns (<the discussed upper bound n(n-1)+1=241). Moreover, the resulting linear relaxation returns an integer optimal solution within seconds (Table 1).

Solving the sequence of restricted set partition problems with increasing m, we obtain all candidate groups for each m.

We also perform K-means++ clustering for each m, observe that all clustering returns the same results. This implies that it is very likely that the local optimum is the global optimum. The clustering results for all m =1, 2, 3,…, 15, 16 are the same results as those from the restricted set partition problem approach.

We use these 16 candidates in Stage 2. In this example, we use formulation (2), i.e. the queueing system of M/G/c queue. The total number of beds is 504 [9].

Table 2 shows the input parameters for our model. Figure 2 shows the results. The SCV for each service is assumed to be 1. Note that the curve is monotonically increasing, meaning that the partition with the shortest expected waiting time across all services would be pooling all services together.

Let $\Omega \subseteq \{1,2\ldots N\}.card(\Omega) = k$. Tekin et al. proposed that with some strict assumptions, a sufficient condition for pooling to be beneficial is when $k^3 \ge \sum_{i\in\Omega}\sum_{j\in\Omega} T_i / T_j$, where $TTii$ is the mean service time for each clinical service [13]. From this we conject that the minimal distances between each pair of offered load ($aakk$) need to be large enough to make pooling unfavourable. Otherwise, it's always beneficial to pool services.

Our initial analysis focuses on similarity of two criteria. These criteria can be weighed to reflect their relative importance. Dimension reduction occurs when the weight is set to zero. For example, if nominal

m	S^m
1	all 16 pooled: (1,2,…16)
2	(7, 9, 14, 15), (1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 13, 16)
3	(7, 9), (14, 15), (1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 13, 16)
4	(7, 9), (14, 15), (1, 2, 3, 4, 5), (6, 8, 10, 11, 12, 13, 16)
5	(7, 9), (14), (15), (1, 2, 3, 4, 5), (6, 8, 10, 11, 12, 13, 16)
6	(7, 9), (14), (15), (1, 2, 3, 4, 5), (6, 8, 10, 12), (11, 13, 16)
7	(7, 9), (14), (15), (1, 2), (3, 4, 5), (6, 8, 10, 12), (11, 13, 16)
8	(7, 9), (14), (15), (1, 2), (3, 4, 5), (6), (8, 10, 12), (11, 13, 16)
9	(7, 9), (14), (15), (1), (2), (3, 4, 5), (6), (8, 10, 12), (11, 13, 16)
10	(7, 9), (14), (15), (1), (2), (3, 4, 5), (6), (8, 10, 12), (11, 13), (16)
11	(7, 9), (14), (15), (1), (2), (4), (3, 5), (6), (8, 10, 12), (11, 13), (16)
12	(7, 9), (14), (15), (1), (2), (4), (3, 5), (6), (8, 10), (12), (11, 13), (16)
13	(7), (9), (14), (15), (1), (2), (4), (3, 5), (6), (8, 10), (12), (11, 13), (16)
14	(7), (9), (14), (15), (1), (2), (4), (3), (5), (6), (8, 10), (12), (11, 13), (16)
15	(7), (9), (14), (15), (1), (2), (4), (3), (5), (6), (8), (10), (12), (11, 13), (16)
16	all 16 specialized

Table 1: Partitions obtained in Stage 1 for example 1.

Service index	λ: the arrival rate	τ: the mean service time
1	1.7	9.5
2	1.02	5
3	0.82	4.43
4	3.3	4.3
5	1.69	4
6	8.95	3.05
7	9.15	6.1
8	2.4	5.6
9	6.4	7.5
10	1.65	8.5
11	2.29	2.5
12	4.19	5
13	4.78	1.6
14	23.47	4.5
15	14.9	5.2
16	0.75	6.5

Table 2: Stage 2 parameters for example 1.

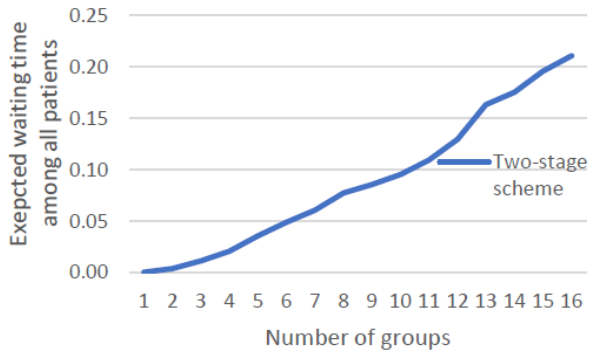

Figure 2: Expected waiting time among all patients in example 1.

daily utility is eliminated from the model, missing opportunities may occur because some services can have simultaneously high values in both demand and utility criteria, for example, due to excellent reputation of the specialists or local monopolistic status of the hospital in particular services, but are grouped with lower-paid entities. On the other hand, if only nominal daily utility is selected, the "price of instability" (i.e. difficulty with getting ρ<1 for groups) may create extra challenges. Intuitively, "relaxation of stability" creates burden for low-profitable services, but helps in protecting the beds for high-profitable ones.

For the direct approach, we formulate it using the MINLP formulation (5). However, LINGO failed to solve it because of instability in the model. Notice that the objective function in (5) has the variables on both base and exponent, which makes the large-scale problem relatively difficult to solve.

Example 2. A hospital in europe

This example is based on the case study in de Bruin et al. [15]. They use the Erlang loss queueing model M/G/c/c to determine the number of required beds for wards (when the blocking probability is pre-specified) for a typical situation "in most Dutch hospitals where ward sizes are relatively small and dispersed and where the 85% target occupancy rate is considered a golden standard." It follows from the Erlang loss formula that the percentage of refused (blocked) admissions given the fixed occupancy rate is declining monotonically and asymptotically to zero as a function of the number of beds. On the other hand, the "mirror" picture of the occupancy rate monotonically growing close to 100% as a function of the number of beds can be observed (given the fixed refused admissions percentage). Proper use of economies of scale in merging departments may help with getting an acceptable balance. De Bruin explored the potential benefit of merging three special departments. We supplement their analysis by considering 15 stationary departments (after combining Internal medicine units 1 and 2, as well as Paediatric units 1 and 2 in their settings). We use the Gini-coefficient for the length of stay (which is the same as the service time in de Bruin's paper) as the criterion of similarity (Figure 3). This coefficient is known in economics and other sciences as a measure of inequality in income and wealth distribution and has values between 0 and 1. Low or high Gini-coefficient in our content indicates that variability in LOS is low or high respectively.

In Stage 1, matrix A of the restricted set partition problem has 106 columns $\left(<n(n-1)+1=211\right)$. Solving the sequence of such problems (Table 3) gives 15 group-candidates: all 15 pooled, [(1,2,4,5,6,7,8,9,10,12,13,14,15),(3,11)],… [(9,13),1,2,3,4,5,6,7,8,10,11,12,14,15], all 15 focused.

The candidate partitions obtained from K-mean++ algorithm is given below (with trivial partitions excluded):

When m=14, $f_{k-means}-f_{IP}=-3.714e-17$, where f is the within-group sum of squares. This indicates that K-means++ returns better candidate partition than the Figure 4 restricted set partition

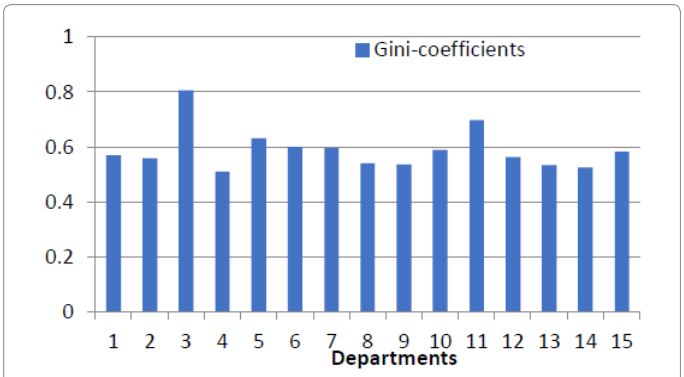

Figure 3: Gini-coefficients for 15 departments (derived from data in de Bruin et al.).

m	S^m
2	(3,11), (1,2,4,5,6,7,8,9,10,12,13,14,15)
3	(3,11), (1,2,5,6,710,12,15), (4,8,9,13,14)
4	(3), (11), (1,2,5,6,710,12,15), (4,8,9,13,14)
5	(3), (11), (1,2,6,710,12,15), (5), (4,8,9,13,14)
6	(3), (11), (1,12), (2,6,7,10,15), (5), (4,8,9,13,14)
7	(3), (11), (1,12), (2,6,7,10,15), (5), (4), (8,9,13,14)
8	(3), (11), (1,12), (2,15), (6,7,10), (5), (4), (8,9,13,14)
9	(3), (11), (1,12), (2,15), (6,7,10), (5), (4), (8,9,13), (14)
10	(3), (11), (1,12), (2,15), (6,7), (10), (5), (4), (8,9,13), (14)
11	(3), (11), (1), (12), (2,15), (6,7), (10), (5), (4), (8,9,13), (14)
12	(3), (11), (1), (12), (2,15), (6,7), (10), (5), (4), (8), (9,13), (14)
13	(3), (11), (1), (12), (2,15), (6), (7), (10), (5), (4), (8), (9,13), (14)
14	(3), (11), (1), (12), (2,15), (6), (7), (10), (5), (4), (8), (9), (13), (14)

Table 3: Partitions obtained in Stage 1 for example 2.

problem. In Stage 2, we use the results returned by the K-means++ method.

In Stage 2, we minimize the worst (highest) blocking probability using MINLP (4). Table 4 shows the optimal partition for each m returned by Stage 1; and Table 5 shows the parameters we use for Stage 2. The total number of beds is set at 150, which is less than the total offered load because the blocking probability proposed by Harel [8] performed poorly when the probability is near 0 or 1. If we set too many or too few beds, the program might not reflect the actual probability accurately.

The optimal configuration is pooling all the departments. This might also due to the close distance between *aakk* as we discussed in example 1. We observe that the direct approach returns lower maximal blocking probabilities than the 2-stage approach for every group numbers. This example illustrates that when the number of clinical services is relatively small, direct approach can return better solution (in reasonable computational time) since it searches over a larger feasible region.

Example 3. A Hospital in Asia

This example is based on the case study in Li et al. [16]. Li et al. proposed a goal programming approach to analyze the trade-off between the number of beds required to achieve a targeted probability of admissions and the number of beds needed to optimize daily profits. We use this example to show how to extend our framework to more general cases. In our analysis, we use its setting of 11 departments and apply the M/G/c/c queueing model. The objective is to maximize the total profits of the hospital.

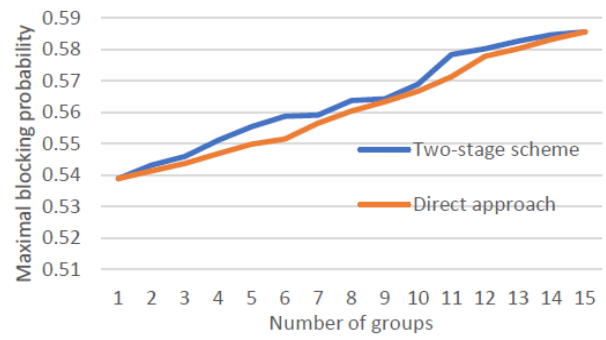

Figure 4: Maximal blocking probability for each group in example 2.

m	S^m
2	(1), (2-15)
3	(9), (7,11), (others)
4	(3,11,12), (1,4,5,6,9,10,14), (7,13), (2,8,15)
5	(10), (6,8,13), (1,14), (9), (2,3,4,5,7,11,12,15)
6	(7,11,15), (2,3,9), (4,5,8,13), (1,6,12), (14), (10)
7	(5,8), (13), (10), (6), (15), (1,3,4,9,14), (2,7,11,12)
8	(13), (15), (11,12), (1,2,4,5,6,9,14), (3), (7), (10), (8)
9	(4), (3), (15), (2,9), (13), (1,5,6,12,14), (10,11), (8), (7)
10	(7), (8,13), (2,9), (15), (3), (12), (4,5),(10),(11), (1,6,14)
11	(2), (3), (10), (7), (15), (5,6,14), (1,9), (11), (4), (8,13), (12)
12	(6,14), (8,13), (4,5), others separated
13	(4,10), (6,14), others separated
14	(4,6), others separated

Table 4: Partitions obtained using the direct approach.

Service index	Λ	τ
1	5.62	4.347
2	6.84	3.172
3	7.57	2.763
4	3.55	6.448
5	7.07	5.468096
6	8.52	3.766035
7	3.8	4.362
8	3.32	4.633
9	4.26	5.256
10	3.29	5.533
11	11.14	1.501
12	3.86	4.527
13	5.18	1.583
14	3.97	6.833
15	3.19	6.487

Table 5: Stage 2 Parameters for example 2.

Li et al. [16] calculated the profits as:

$$\rho = r\lambda\tau\left(1-B\right)-\pi\lambda B-\eta\left(C-\lambda\tau\left(1-B\right)\right)$$

where p=the average profit per day; r= revenue per day generated from each admitted patient; π=penalty cost for each patient being turned away; η=the holding cost per day per idle bed; and c,λ,τ,B follow the same meanings as an Example 1. These variables can be applied to both single service and groups.

For the 2-stage framework, we pick two similarity criteria, as reflected in Figure 5: normalized targeted number of beds with respect to 95% of patient admission (columns), and normalized targeted average profit per day for each department (Tables 1 and 2 respectively in [16]).

In Stage 1, matrix A of the restricted set partition problem has 56 columns. Solving the sequence of problems yields 11 group-candidates. Table 6 shows the best partitions obtained by solving the restricted set partitioning problem (1) versus the K-mean++ approach.

The only difference is when m=5, with $f_{K-means} - f_{ip} = 1.301e-18$. This shows that the restricted set-partitioning approach returns s slightly better solution. We use these partitions in Stage 2.

Before we establish the Stage 2 problem, it is necessary to calculate the group parameters for r, π and $\eta\eta$. We use R,Π,H to represent respectively these parameters for the group. By definition, the revenue generated by each admitted patient per day is solely determined by the services that the patient receives, and it does not relate to the length of stay nor other variables such as the number of beds and the blocking probability, etc. Thus, the revenue for group Si, *RRii*, can be estimated as $R_i = \dfrac{\sum_{j\in S_i}\lambda_j\gamma_j}{\sum_{j\in S_i}\lambda_j}$

We employ similar logic to calculate Π and H as the weighted average. We also introduce a way to calculate group offered load *AAii* and group arrival rate *ΛΛii*, which is the sum of service arrival rates in the group.

In Stage 2, we maximize the average profit per day. We modify formulation (4) as follows:

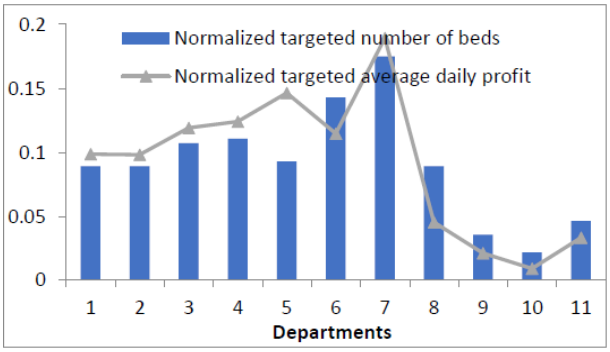

Figure 5: Normalized targeted bed distribution and average profit per day (derived from the table data in Li et al.).

m	Sm obtained by solving (1)	Sm obtained by K-mean++
2	(1,2,3,4,5,6,7), (8,9,10,11)	(1,2,3,4,5,6,7), (8,9,10,11)
3	(1,2,3,4,5,6), (7), (8,9,10,11)	(1,2,3,4,5,6), (7), (8,9,10,11)
4	(1,2,3,4,5,6), (7), (8), (9,10,11)	(1,2,3,4,5,6), (7), (8), (9,10,11)
5	(1,2,3,4,5),6,7,8,(9,10,11)	(1,2), (3,4,5,6), (7), (8), (9,10,11)
6	(1,2), (3,4,5), (6), (7), (8), (9,10,11)	(1,2), (3,4,5), (6), (7), (8), (9,10,11)
7	(1,2), (3,4), (5), (6), (7), (8), (9,10,11)	(1,2), (3,4), (5), (6), (7), (8), (9,10,11)
8	(1,2), (3,4), (5), (6), (7), (8), (9,11), (10)	(1,2), (3,4), (5), (6), (7), (8), (9,11), (10)
9	(1,2), (3,4), (5), (6), (7), (8), (9), (10), (11)	(1,2), (3,4), (5), (6), (7), (8), (9), (10), (11)
10	(1,2), (3), (4), (5), (6), (7), (8), (9), (10), (11)	(1,2), (3), (4), (5), (6), (7), (8), (9), (10), (11)

Table 6: Partitions returned by Stage 1 via the restricted set partitioning problem (1) *versus* the K-means++ approach for example 3.

$$\min \sum_{i=1}^{N} \rho_i$$

s.t.

$$\rho_i = R_i A_i (1-B_i) - \prod_i \Lambda_i B_i - H_i (c_i - A_i (1-B_i)) \forall i$$

$$B_i = \frac{A_i - c_i - 2\rho_i + \sqrt{(A_i - C_i)^2 + 4A_i}}{2A_i - 2\rho_i} \qquad \forall i$$

$$\rho_i c_i = A_i \qquad \forall_i$$

$$A_i (1-B_i) \le c_i \qquad \forall_i$$

$$\sum_{i=1}^{N} c_i = c$$

$$B_i \le b \qquad \forall i$$

$$\rho_i \ge \rho_i \qquad \forall i$$

$$c_i \in Z^+, B_i, \rho_i \in R \forall i = 1, \ldots, N$$

Here, bb is the value for the maximal blocking probability for any group set by the hospital. $ppii$ is the minimal profits for each group, which is assumed to be the sum of original profits of each service in the group, since it is undesirable to the hospital if the pooled services are less profitable than before. We choose bb to be 0.4 and the total number of beds C to be 200. Other parameters are shown in Table 7.

For the direct approach, we establish the following MINLP:

Service index	R	π	H
1	109.08	78.97436	11.217
2	110	82.05128	19
3	106.9	62.85714	15.856
4	105.896	62.4	9.7094
5	150.42	78.04878	15.61
6	72.52	47.33728	8.35
7	90.75086	108.3187	8.86171
8	56.4	81.02564	21.2
9	76.98143	3.737468	7.63253
10	81.16582	82.93014	11.64
11	80.6651	50	6.41363
Service index	**λ**	**τ**	**Original profits**
1	9.75	2	932.4118
2	9.75	2	858.4205
3	12.25	2	1108.404
4	12.5	2	1151.723
5	10.25	2	1364.731
6	16.9	2	1076.631
7	20	2.15	1155.748
8	9.75	2	282.8132
9	1.2	5.175	39.5974
10	1.48	2	52.09813
11	1	8.8	44.54212

Table 7: Stage 2 parameters for example 3.

Subject to:

$$\min \quad \sum_{i=1}^{N} p_i$$

$$\sum_{j=1}^{n} x_{ij} \ge 1 \qquad \forall i = 1, \ldots, N$$

$$\sum_{i=1}^{N} x_{ij} = 1 \qquad \forall j = 1, \ldots, n$$

$$A_i = \sum_{j=1}^{n} a_j x_{ij} \qquad \forall i = 1, \ldots, N$$

$$R_i = \frac{\sum_{j=1}^{n} \lambda_j r_j x_{ij}}{\sum_{j=1}^{n} \lambda_j x_{ij}} \qquad \forall i = 1, \ldots, N$$

$$\Pi_i = \frac{\sum_{j=1}^{n} \lambda_j \pi_j x_{ij}}{\sum_{j=1}^{n} \lambda_j x_{ij}} \qquad \forall i = 1, \ldots, N$$

$$H_i = \frac{\sum_{j=1}^{n} \lambda_j \eta_j x_{ij}}{\sum_{j=1}^{n} \lambda_j x_{ij}} \qquad \forall i = 1, \ldots, N$$

$$B_i = \frac{A_i - c_i - 2\rho_i + \sqrt{(A_i - c_i)^2 + 4A_i}}{2A_i - 2\rho_i}, \qquad \forall i = 1, \ldots, N$$

$$A_i (1-B_i) \le c_i \qquad \forall i = 1, \ldots, N$$

$$\rho_i c_i = A_i \qquad \forall i = 1, \ldots, N$$

$$\sum_{i=1}^{N} c_i = C$$

$$B_i \le b \qquad \forall i=1,...,N$$

$$\rho_i \ge \rho_i \qquad \forall i=1,...,N$$

$$c_k \in Z^+, B_i, \rho_i, A_i, R_i, \Pi_i, H_i, p_i \in R, \forall i = 1,..., N$$

Figure 6 shows that pooling all services offer the optimal partition. This corresponds to the results in example 2 that lower maximal blocking probability implies higher profits. In this example, the direct approach does not perform as well as the 2-stage approaches, which counters the theory the solution space of the direct approach contains that of the 2-stage. The inferior performance is resulted from the limitation of the commercial solver in tackling these highly nonlinear mixed integer programming instances. In practice, the resources planner should apply both approaches to obtain the best possible results.

Each of these examples has a specific illustrative purpose. They demonstrate how diverse the modeling ingredients and the outcomes can be.

Discussions and Conclusions

The efficient use of hospital beds is critical as financial pressure and demand grow. Our 2-stage and direct modeling frameworks offer broad opportunities for hospital managers to make strategic decisions about this fundamental resource. Using our methodology, hospitals of different sizes, missions, and visions can find reasonable practical solutions to efficiently manage their bed allocation.

We provide balanced models for the trade-off between pooling capacity and focused care while incorporating utility/gain objectives to drive the solution process. Flexibility can be suggested by economies of scale. The previous practice and experience of managers can suggest suitable ingredients for similarity criteria. Existing performance metrics (clinical, operational, financial, organizational) offer diverse options for applying our framework.

Three structured examples illustrate how different models can be embedded within our proposed schema. Widely used financial and operational metrics (related to average patient length of stay, admissions, discharges, transfers, utilizations of beds, etc.) coupled with achievements in queueing theory and optimization help in understanding, modeling, and solving problems of vital processes. Hard decisions of compromising between fair access for patients and a hospital's performance (and even its financial survival) may receive simple solutions in combining or not combining clinical services. Further managerial insights can be obtained depending on particular goals.

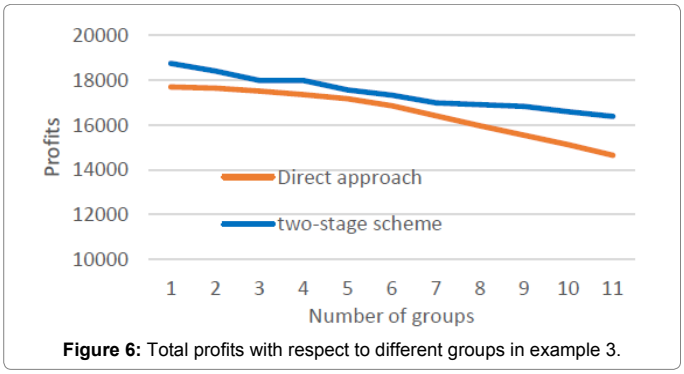

Figure 6: Total profits with respect to different groups in example 3.

Although the three-fold inpatient bed capacity management problem is intrac in general and has not been solved to global optimality, both of our frameworks provide idea to reach local optimality. If there are large amount of services or the objective incurs many variables in the program, the 2-stage scheme could be applied. Otherwise, direct approach could be better since it searches through the entire feasible region.

We only briefly touch on some properties of clustering analysis, which usually deals with a fixed number of clusters. If the number is unknown, two most common ways to proceed are as follow [5]: The first is to solve the problem repeatedly for different numbers, then "compare some criterion for each cluster", the value of a gap suggests the number of clusters. A second approach is to define a threshold for the creation of a new cluster. The leading entry and the K-mean++ discussed in Section II combines both of these approaches.

In the M/G/c and M/G/c/c models, pooling appears to be more beneficial than any other grouping when the minimal distances between the offered loads are far apart. That partly explains the results for example 1 and 2.

An observation is that when some groups have significantly more entities than the others, the offered load would be farther apart from each other because the group offered load is the sum of the individual offered load in the group. In this case, the objective function may not be strictly monotonic with respect to the number of groups. However, in our examples all the curves are monotonic, since the offered loads are not too diverse.

The proposed framework is generalizable and can be applied and extended to other underlying queueing models (e.g. limited space waiting M/G/c/k, rush hour/seasonal, abandonment, or trying simulation-based optimization), similarity criteria (e.g. nurse training, equipment, location), and goals (in terms of service quality or finance). Its flexible structure is open for experiments in designing new hospitals and hospital chains. New developments in queueing theory and hospital performance metrics and standards can establish new horizons in the choice of modeling components for the framework's practical use.

We caution that our analysis is sensitive to the input data from the hospital. This includes estimates for the arrival time, the service time, the revenue generated from each admitted patient, the penalty for each patient being turned away, and the holding cost per daily idled bed for each clinical service. These input will affect the partitions and the resulting objective function values. Care should be exercised in data collection to ensure that the results obtained are meaningful to the clinical service and patient demand patterns.

Acknowledgment

The work is partially supported by grants from the National Science Foundation, IIP- 0832390 and IIP-1361532. Findings and conclusions in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Green LV (2008) Using Operations Research to Reduce Delays for Healthcare. Informat Int Age 2008: 1-16.
2. Van Essen JT, Houdenhoven MV, Hurink JL (2015) Clustering clinical departments for wards to achieve a prespecified blocking probability. OR spectrum 37: 243-271.
3. Keegan AD (2010) Hospital bed occupancy: more than queuing for a bed. Med J Aust 193: 291-293.
4. Hall R (2012) Handbook of healthcare system scheduling. Springer 2012: 177-200.

5. Duda RO, PE Hart, Stork DG (2012) Pattern classification.
6. Vanberkel PT (2010) Efficiency evaluation for pooling resources in health care. OR Spectrum 34: 371-390.
7. Clark JR, Huckman RS (2012) Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. Manag Sci 58: 708-722.
8. Kc DS, Terwiesch C (2011) The Effects of Focus on Performance: Evidence from California Hospitals. Manag Sci 57: 1897-1912.
9. Best TJ (2015) Managing Hospital Inpatient Bed Capacity through Partitioning Care into Focused Wings. Manufact Ser Operat Manag 17: 157-176.
10. Rao M (1971) Cluster analysis and mathematical programming. J Americ Stat Assoc 66: 622-626.
11. Arthur D, Vassilvitskii S (2007) K-means++: the advantages of careful seeding. Stanford 2007: 1027-1035.
12. Whitt W (1999) Partitioning Customers into Service Groups. Manag Sci 45: 1579-1592.
13. Tekin E, Hopp WJ, Van Oyen MP (2009) Pooling strategies for call center agent cross-training. IIE Transactions 41: 546-561.
14. Harel A (2009) Sharp and simple bounds for the Erlang delay and loss formulae. Que Sys 64: 119-143.
15. De Bruin AM (2010) Dimensioning hospital wards using the Erlang loss model. Annals Oper Res 178: 23-43.
16. Li X (2008) An integrated queuing and multi-objective bed allocation model with application to a hospital in China. J Operat Res Soc 60: 330-338.