Memorandum 1947 (July 2011). ISSN 1874-4850. Available from: http://www.math.utwente.nl/publications
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Tactical Resource Allocation and Elective Patient Admission Planning in Care Pathways

Peter J.H. Hulshof*†, Richard J. Boucherie, Erwin W. Hans†, Johann L. Hurink
Center for Healthcare Operations Improvement and Research (CHOIR), University of Twente, Enschede, the Netherlands

Tactical planning of resources in hospitals concerns elective patient admission planning and the intermediate term allocation of resource capacities. Its main objectives are to achieve equitable access for patients, to meet production targets/to serve the strategically agreed number of patients, and to use resources efficiently. This paper proposes a method to develop a tactical resource allocation and elective patient admission plan. These tactical plans allocate available resources to various care pathways and determine the selection of patients to be served that are at a particular stage of their care pathways. Our method is developed in a Mixed Integer Linear Programming framework and copes with multiple resources, multiple time periods and multiple patient groups with various uncertain care pathways, thereby integrating decision making for a chain of hospital resources. Computational results indicate that our method leads to a more equitable distribution of resources and provides control of patient access times, the number of patients served and the faction of allocated resource capacity. Our approach is generic, as the base MIP and the solution approach allow for including various extensions to both the objective criteria and the constraints. Consequently, the proposed method is applicable in various settings of tactical hospital management.

Key words: Care pathways, tactical planning, resource capacity planning, patient admission planning, Mixed Integer Linear Programming (MIP)

1. Introduction

Tactical planning of resources is a key element of hospital planning and control that concerns elective patient admission planning and the intermediate term allocation of resource capacities. The main objectives are to achieve equitable access and treatment duration for patient groups, to serve the strategically agreed target number of patients (i.e., production targets or quota), to maximize resource utilization and to balance workload.

Hospital management aims to provide equitable access and treatment duration for patient groups by controlling access times. Access time is the time a patients spend on the waiting list before being served, and controlled access times ensure quality of care for the patient and prevents patients from seeking treatment elsewhere [30]. Also, in some financial systems, hospitals receive payments only after patients have completed their health care delivery process. Hence, waiting patients can be costly, as resources and materials have already been invested, but revenues are still to come. Furthermore, hospital management may have agreed with insurers or government to serve a target number of patients. Therefore, evaluation and control of the number of patients served helps to ensure that strategic objectives are being reached.

From a clinician's perspective, tactical resource and admission plans subdivide the clinician's time into segments (e.g., consultation time and surgical time) and determine the number of patients to serve from a particular patient group at a particular stage of their care pathway (e.g, consultation

* This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

† Corresponding authors: University of Twente, Peter J.H. Hulshof and Erwin W. Hans, P.O. Box 217, 7500 AE Enschede, the Netherlands. E-mail: {p.j.h.hulshof,e.w.hans}@utwente.nl.

or surgery). Because care pathways connect multiple departments and resources into a network, fluctuations in both patient arrivals (e.g., seasonality) and resource availability (e.g., holidays) result in bullwhip effects [22] in the care chain. From a patient's perspective, this means access times strongly fluctuate during their care pathways. From a hospital's perspective, this means that resource utilizations and service levels fluctuate. To cope with these fluctuations, intermediate-term re-allocation of hospital resources, taking into account a care chain perspective [6, 16, 25], is required. For example, only optimizing the outpatient clinic capacity may lead to waiting times and congestion downstream at the operating rooms. Likewise, optimizing operating room utilization without considering admission planning in the outpatient clinic may lead to underutilized operating room capacity.

The available approaches on the development of tactical resource and admission plans in the Operations Research and Management Science (OR/MS) literature are myopic, focus on developing long-term cyclical plans, or are not able to provide a solution for real-life sized instances; see Section 2 for details. This paper presents a method to determine intermediate term tactical resource and admission plans to cope with fluctuations in patient arrivals and resource availability. These plans are developed for multiple resources and multiple patient groups with various care pathways, thereby integrating decision making for a chain of hospital resources. The method incorporates available knowledge about the state of the waiting lists and the available resource capacities. Our computational results show that our method can be used to develop tactical resource and admission plans for real-life sized instances and that it improves compliance with strategically set targets for access times, care pathway duration and the number of patients served. The presented method can also be used to develop tactical plans in other service industries and in manufacturing. However, we restrict the presentation of the model and results in the terms of health care.

This paper is organized as follows. Section 2 discusses tactical resource and admission planning in health care and industry. Section 3 presents our method for tactical resource and admission planning. Section 4 discuss our approach to generate instances, based on examples from practice, that are used to run computational experiments. Section 5 presents the results of these computational experiments, and Section 6 concludes this paper.

2. Background

Due to increasing demand for health care and increasing expenditures [24], health care organizations are trying to re-organize processes more efficiently and effectively. Planning and control in health care has received an increased amount of attention over the last ten years [4], both in practice and in the literature. Health care planning and control can be subdivided in the hierarchical levels of strategic, tactical and operational planning [18]. While strategic planning addresses the dimensioning of resource capacities, tactical planning subdivides the settled resource capacities among patient groups to reach strategically set targets and to facilitate operational planning, and operational planning involves the short-term decision making related to the execution of the health care delivery process. In this section, we discuss approaches in the literature for tactical planning in health care and in industry.

Tactical resource and admission planning approaches are static or dynamic. Static approaches result in long-term plans that are often cyclical. Dynamic approaches result in intermediate-term plans in response to the variability in demand and supply. These approaches are compared in [27], and their simulation results indicate that the dynamic approach results in lower access times and higher resource utilization.

Tactical resource and admission planning approaches in health care are often myopic, which means that they do not consider multiple departments and resources along a care pathway for patient

groups. For example, they focus on the outpatient clinic [7, 11], diagnostic services [15, 27] or operating rooms [1, 2, 8, 10, 26]. Although the benefits of an integrated approach are often recognized [6, 16, 25], relatively few articles integrate decision making for a chain of resources or departments along the patient's care pathway. To support integrated tactical resource and admission planning, [21] models care pathways as Markov chains to derive resource requirements for each stage of a patient's care pathway. Similar approaches for evaluation of resource requirements are taken in [9, 12, 19, 28]. In order to calculate optimal static, elective patient admission plans for multiple resources and multiple patient groups with various care pathways, [23] models the patient process as a Markov Decision Process (MDP). Their experiments show that alternative methods to solve the model should be developed, as the MDP approach is not yet suitable for realistically sized instances.

The process of patients flowing through a network of service units can be compared to a classical job shop in industry [13], which is a network of work stations capable of producing a wide variety of jobs [14]. Hence, methods used in industry for job shop scheduling may be suitable for tactical resource and admission planning in health care. Queueing models can be used to analyze tactical production plans for a job shop [14, 20]. However, results in queueing theory are often based on steady state assumptions, and therefore, queueing models are not suitable to analyze dynamic plans with a finite planning horizon. Other methods to analyze a network of workstations in industry are in the field of project scheduling. Project scheduling is concerned with small batch production where resources are allocated to production activities over time [5]. Methods to allocate resources to activities in project scheduling are often based on mathematical programming [17, 29, 31] or MDP [3].

Summarizing, existing approaches to tactical resource and admission planning in health care are myopic, focus on developing long-term cyclical plans, or do not provide a solution for real-life sized instances. In Section 3, we propose a method to develop tactical resource and admission plans on the intermediate term, for multiple resources and multiple care pathways.

3. Model description

We aim to allocate resource capacities among the various consecutive stages of different care pathways. To this end, we propose a Mixed Integer Linear Program (MIP) to compute a patient admission plan for multiple consecutive time periods. Section 3.1 provides the constraints to model tactical resource and admission planning. We present our approach to the objective function in Section 3.2. In our objective function, a weight reflects the priority to serve patients at a particular stage in a particular care pathway. The determination of these weights is discussed in Section 3.3. Tactical resource and admission planning has multiple objectives. Health care organizations may prioritize these objectives differently, resulting in multiple possible objective functions. Hence, we discuss the performance measures that can be calculated within the MIP to form alternative objective functions in Section 3.4. In addition, we discuss other extensions of the model in Section 3.5. In the following, we introduce the problem in more detail and give some notation.

The planning horizon is discretized in consecutive time periods $\mathcal{T} = \{0, 1, 2, \ldots, T\}$. Furthermore, we consider a set of resource types $\mathcal{R} = \{1, 2, \ldots, R\}$ and a set of patient care pathways $\mathcal{G} = \{1, 2, \ldots, G\}$. The number of patients that can be served by resource $r \in \mathcal{R}$ is limited by the available resource capacity $\phi_{r,t}$ in time period $t \in \mathcal{T}$. Formally, patients that follow patient care pathway $g \in \mathcal{G}$ receive care specified by a set of stages $S_g = \{(g, 1), (g, 2), \ldots, (g, e_g)\}$, where e_g is the number of stages of the care pathway. Patients following the same care pathway have the same resource requirements in each stage (e.g., consultation times, surgery duration, number of consultations), and to serve a patient of care pathway $g \in \mathcal{G}$ in stage $j = (g, a)$ requires $s_{j,r}$ time units of resource

$r \in \mathcal{R}$. After service at a certain stage, patients move to the next stage of their care pathway or leave the system. More precisely, for two stages $i, j \in S_g$ the value q_{ij} denotes the fraction of patients that move from stage i to stage j, and the value $1 - \sum_{j \in S_g} q_{ij}$ denotes the fraction of patients that leave the system. At each stage, patients may have to queue for service. Hence for each care pathway, we obtain a set of queues $\mathcal{J}_g$ of cardinality e_g. Although patients of different care pathways share resources for service, we model the queues disjointly for different care pathways. Consequently, we have a total set of queues $\mathcal{J} = \bigcup_{g \in \mathcal{G}} \mathcal{J}_g$, where $|\mathcal{J}| = \sum_{g \in \mathcal{G}} e_g$.

In each time period $t \in \mathcal{T}$, we determine a patient admission plan, characterized by the decision variable vectors $C_{j,t} = \left(C^0_{j,t}, C^1_{j,t}, \ldots\right)$. The decision variable $C^n_{j,t}$ indicates the number of patients to serve in time period $t \in \mathcal{T}$ that have been waiting precisely n time periods at queue $j \in \mathcal{J}$. In order to calculate the decision variables $C_{j,t}$, we evaluate for each queue $j \in \mathcal{J}$ the number of patients that are waiting and the time that these patients are waiting. Therefore, we introduce waiting lists $W_{j,t} = \left(W^0_{j,t}, W^1_{j,t}, \ldots\right)$, where $W^n_{j,t}$ gives the number of patients that have been waiting precisely n time periods at queue $j \in \mathcal{J}$ at the beginning of time period $t \in \mathcal{T}$. When patients in waiting list entry $W^n_{j,t}$ are not served in time period t, they move to the entry $W^{n+1}_{j,t+1}$ in period $t+1$. Figure 1 illustrates the dynamics of the waiting list for a single queue.

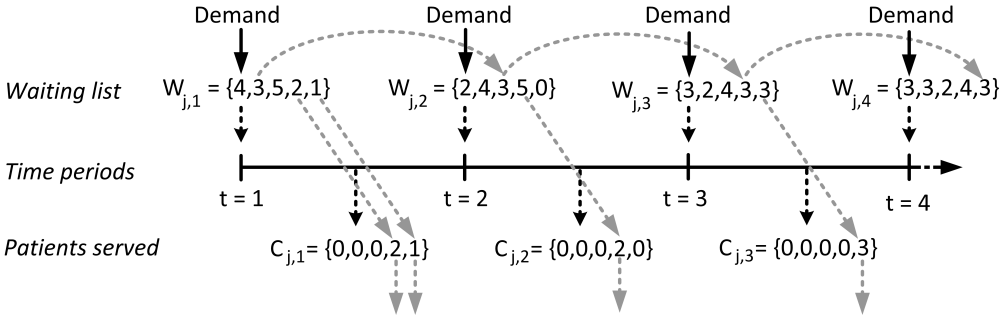

Figure 1 The dynamics of the waiting list and patient service for a system with a single queue $j \in \mathcal{J}$.

For ease of notation we summarize the transition rates between the stages/queues in a routing matrix Q of dimension $|J| \times |J|$. By using one matrix Q for the rates, we enable modeling transitions between queues of different patient care pathways (although not taken into account in this paper). Furthermore, to be able to take into account a minimum time lag before patients that have been served at one queue, can enter the following queue, we introduce a delay matrix D of dimension $|J| \times |J|$, where the entry d_{ij} denotes the minimum time lag (in time periods) between service from queue i and entrance to queue j $(i, j \in \mathcal{J})$. Finally, in addition to demand originating from serving patients from other queues, there is a deterministic demand from outside the system $\lambda_{j,t}$ $(j \in \mathcal{J}, t \in \mathcal{T})$. Together, the number of patients entering queue $j \in \mathcal{J}$ in time period $t \in \mathcal{T}$ is given by:

$$W^0_{j,t} = \lambda_{j,t} + \sum_{i \in \mathcal{J}} \sum_{n=0}^{\infty} q_{ij} \cdot C^n_{i,t-d_{ij}}. \tag{1}$$

3.1. Constraints to calculate a tactical resource and admission plan

The constraints to model the care pathways of patients in the tactical planning problem are given below. Table 1 gives the sets, indices, variables and parameters used. Possible extensions are presented in Section 3.5.

$$W^0_{j,t} = \lambda_{j,t} + \sum_{i \in \mathcal{J}} \sum_{n=0}^{\infty} q_{ij} \cdot C^n_{i,t-d_{ij}} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, \tag{2}$$

$$W^n_{j,t} = W^{n-1}_{j,t-1} - C^{n-1}_{j,t-1} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, n > 0, \tag{3}$$

$$C^n_{j,t} \leq W^n_{j,t} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, n \geq 0, \tag{4}$$

$$\sum_{j \in \mathcal{J}^r} s_{j,r} \sum_{n=0}^{\infty} C^n_{j,t} \leq \phi_{r,t} \qquad \forall r \in \mathcal{R}, t \in \mathcal{T}, \tag{5}$$

$$C_{j,t} = \sum_{n=0}^{\infty} C^n_{j,t} \in \mathbb{N} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, \tag{6}$$

$$C_{j,t} \in \mathbb{N} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}. \tag{7}$$

Sets		**Indices**	
$\mathcal{J}$	Queues	$i, j \in \mathcal{J}$	Queue
$\mathcal{T}$	Time periods	$t \in \mathcal{T}$	Time period
$\mathcal{R}$	Resource types	$r \in \mathcal{R}$	Resource type
$\mathcal{J}^r$	Queues for resource type r, $\mathcal{J}^r \subseteq \mathcal{J}$	$i, j \in \mathcal{J}^r$	Queue
		n, d	Time periods (to indicate waiting time)
Variables			
Decision variable			
$C^n_{j,t}$	The number of patients served from queue j in time period t, who have been waiting n time periods		
Auxiliary variable			
$W^n_{j,t}$	The number of patients in queue j at the start of time period t, who have been waiting n time periods		
Parameters			
β^n_j	Objective function weight of patients in queue j, who have been waiting n time periods		
$\lambda_{j,t}$	New demand in queue j in time period t		
$\phi_{r,t}$	Capacity of resource type r in time period t in time units		
q_{ij}	Probability that a patient moves from queue i to queue j		
d_{ij}	Number of time periods to move from queue i to queue j		
$s_{j,r}$	Expected capacity requirements from resource type r for a patient in queue j in time units		

Table 1 The sets, indices, variables and parameters used.

Constraints (2) and (3) stipulate that the waiting list variables are consistent. Constraint (2) determines the number of patients newly entering a queue. Constraint (3) updates the waiting list variables at each time period $t \in \mathcal{T}$. Constraint (4) stipulates that not more patients are served than the number of patients on the waiting list. Constraint (5) assures that the resource capacity of each resource type $r \in \mathcal{R}$ is sufficient to serve all patients. Constraint (6) determines the total number of patients served at a queue in a time period, and Constraint (7) is an integrality constraint for the total number of patients served at a queue in a time period.

REMARK 1. For numerical purpose, to solve our optimization problem, the number n of time periods that patients are waiting is bounded at some value $\hat{n}$. Consequently, Constraints (2) to (6) require adaptation and a constraint is added to stipulate that the number $W^{\hat{n}}_{j,t}$ of patients who are not served in time period $t-1$ and are waiting $\hat{n}$ time periods, remain on the waiting list in time period t:

$$W^{\hat{n}}_{j,t} = \textstyle\sum_{m=\hat{n}-1}^{\hat{n}} \left(W^m_{j,t-1} - C^m_{j,t-1} \right), \qquad \forall j \in \mathcal{J},\, t \in \mathcal{T}. \tag{8}$$

3.2. Objective function

From our experience with the hospitals we collaborate with, the main objectives of tactical planning are *to achieve equitable access and treatment duration for patient groups* and *to serve the strategically agreed number of patients*. Therefore, we incorporate these two objectives in our objective function (9). The other objectives of tactical planning mentioned in Section 1; *to maximize resource utilization* and *to balance the workload*, can be captured in alternative objective functions and extensions of the model. We propose starting points for these extensions in Sections 3.4 and 3.5 respectively.

We use the following objective function:

$$\min \ \sum_{j\in\mathcal{J}} \sum_{n=0}^{\infty} \beta_j^n \sum_{t\in T} W_{j,t}^n. \tag{9}$$

The objective function (9) aggregates the weighted number of patients waiting in each queue $j \in \mathcal{J}$ in each time period $t \in \mathcal{T}$. Weights β_j^n ($j \in \mathcal{J}$ and $n = 0, 1, 2, \ldots$) are incorporated in the objective function to prioritize the various queues in order to deploy resources where they are most effective. The two objectives, *to achieve equitable access and treatment duration for patient groups* and *to serve the strategically agreed number of patients*, are reflected in these weights. We propose an iterative procedure to determine these weights in Section 3.3.

3.3. Procedure to determine the weights

The effect of the resource allocation is measured in the MIP's objective. Inspired by the hospitals we collaborate with, we choose to use access time and the number of patients served as performance metrics. The procedure to determine the weights of the objective terms is an iterative one. We initialize the weights, solve the MIP and measure the metrics as we explain below, then update the weights, solve the MIP, etc. In this section we first explain how we measure the performance metrics from the MIP solution, and then explain in detail the iterative procedure of determining the weights.

Access time. Access time $A_{j,t}^{\alpha}$ may be measured from the MIP solution as follows:

$$A_{j,t}^{\alpha} = \min\{n | \sum_{m=0}^{n} W_{j,t}^m > \alpha \sum_{m=0}^{\infty} W_{j,t}^m\}, \ j \in \mathcal{J}, \ t \in \mathcal{T}, \tag{10}$$

where α is a given percentile. Hospital managers aim to control access times by imposing access time targets $\hat{a}_{j,t}^{\alpha}$. We aim to evaluate the effect of a calculated tactical resource and admission plan on the access time $A_{j,t}^{\alpha}$ in comparison with the access time target $\hat{a}_{j,t}^{\alpha}$ for each queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$. Hence, we may calculate a performance ratio $L_{\alpha,j,t}^{A}$ for the access time with:

$$L_{\alpha,j,t}^{A} = \frac{A_{j,t}^{\alpha}}{\hat{a}_{j,t}^{\alpha}}, \qquad j \in \mathcal{J}, \ t \in \mathcal{T}. \tag{11}$$

We use this ratio to evaluate how close to target the performance of the current solution is. For example, if $L_{\alpha,j,t}^{A} > 1$, then access times are above target.

The number of patients served. Health care managers aim to control the number $C_{j,t}$ of patients served by imposing a target $\hat{c}_{j,t}$ for the number of patients served. We assume that this target $\hat{c}_{j,t}$ is given for each queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$. In practice, targets may typically be set for care pathways, by setting the target for either the first or the last queue in care pathways. In our model, we assume that these care pathway targets can be converted to targets for each stage of a care pathway.

We aim to evaluate the effect of a calculated tactical resource and admission plan on the number $C_{j,t}$ of patients served in comparison with the target number $\hat{c}_{j,t}$ of patients served for each queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$. Hence, we may calculate a performance ratio $L^C_{j,t}$ for the number of patients served by:

$$L^C_{j,t} = \frac{\hat{c}_{j,t}}{C_{j,t}}, \qquad j \in \mathcal{J},\, t \in \mathcal{T}. \tag{12}$$

We use the performance ratios (11) and (12) in the procedure to calculate the weights, which we explain below. The nonnegative weights β^n_j, where $j \in \mathcal{J}$ and $n = 0, 1, 2, \ldots$ indicate the number of time periods waiting, lead to a matrix B:

$$B = \begin{pmatrix} \beta^0_1 & \beta^1_1 & \beta^2_1 & \cdots \\ \beta^0_2 & \beta^1_2 & \beta^2_2 & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ \beta^0_{|\mathcal{J}|-1} & \beta^1_{|\mathcal{J}|-1} & \beta^2_{|\mathcal{J}|-1} & \cdots \\ \beta^0_{|\mathcal{J}|} & \beta^1_{|\mathcal{J}|} & \beta^2_{|\mathcal{J}|} & \cdots \end{pmatrix}$$

Two assumptions are made regarding the structure of B.

ASSUMPTION 2. *$\beta^n_j < \beta^{n+1}_j$, for all $j \in \mathcal{J}$ and $n = 0, 1, 2 \ldots$*

ASSUMPTION 3. *If $q_{ij} > 0$, then $\max_n \beta^n_i > \min_n \beta^n_j$, for all $i, j \in \mathcal{J}$*

REMARK 4. Under Assumption 2, $\min_n \beta^n_j = \beta^0_j$, for all $i, j \in \mathcal{J}$

In the following, we justify these assumptions from a theoretical and practical point of view.

1. When patients are served first-come, first-served (FCFS) at queue $j \in \mathcal{J}$, we want the MIP to have the incentive to first serve the patient who has waited the longest in queue j. This FCFS property leads to monotonically increasing weights β^n_j for each queue $j \in \mathcal{J}$,
2. If a patient moves with positive probability from queue i to queue j $(i, j \in \mathcal{J})$, there is a local incentive to serve the patient at queue i when the maximum weight in row i is larger than the minimum weight in row j. If B is not structured in this way, then even with an infinite resource capacity at queue i, locally there is no incentive to serve a patient at queue i.

We propose the following function to determine B:

$$\beta^n_j = \begin{cases} 0 & \text{if } n = 0 \\ u_j \cdot (m_j)^n & \text{if } n > 0 \end{cases} \quad \forall j \in \mathcal{J}. \tag{13}$$

This function requires two parameters u_j and m_j per queue $j \in \mathcal{J}$ to determine B. By restricting the parameter m_j to values larger than 1, we satisfy Assumption 2. Following Remark 4, by setting $\beta^0_j = 0$, we ensure that Assumption 3 holds.

Taking into account Assumptions 2 and 3, the weights in B can be determined with various approaches. For example, one may manually decide on the weights, based on numerous performance measures and perhaps other quantifiable or subjective reasons. These performance measures can be patient oriented, such as access time, medical urgency and pain experience, and organization oriented, such as financial incentives and agreements with insurance companies about the number of patients to serve. In this paper, we propose to calculate the weights in an iterative manner as follows. First, B is initialized with starting values and the MIP is solved. After that, B is updated

based on the solution of the MIP, and the MIP is solved again with the updated B. This iterative way of updating B and solving the MIP is performed until some criterion is met. To design such an iterative procedure, three topics need to be addressed:

1. The initialization of B.
2. The adaptation of B after solving the MIP.
3. The stopping criterion.

The following iterative procedure is used to initialize and update B. In B there are at most $|J| \times \hat{n}$ elements that require initializing and updating. By using 13, we need to adjust at most $2 \times |J|$ parameters every iteration. The iterative procedure uses the performance ratios $L^A_{\alpha,j,t}$ and $L^C_{j,t}$ to update B by determining new values for u_j and m_j for each queue $j \in \mathcal{J}$. First, the parameters u_j and m_j are initialized by evaluating the performance ratios in previous planning period. Consequently, the performance prior to the planning period influences decision making in the planning period. When no historical data is available, the parameters are assumed to be $u_j = 1$ and $m_j = 1 + \epsilon$, where ϵ is a small number. The weights β^n_j corresponding to the chosen values u_j and m_j are calculated with (13) and the MIP is solved. Based on the MIP solution, the parameters u_j and m_j are updated using the performance ratios for this planning period. To avoid strong oscillations of the outcome for the performance ratios over the course of the planning period, we ensure that the number of changes of the parameters gets smaller with increasing number of iterations.

In the following, we formalize the iterative procedure to update B. The iteration number is indicated by s.

Step 1: $s := 1$. Initialize u_j and m_j, for all $j \in \mathcal{J}$, with:

$$u_j(1) = \frac{\hat{c}_{j,0}}{C_{j,0}}, \qquad m_j(1) = 1 + \frac{A^\alpha_{j,1}}{\hat{a}^\alpha_{j,1}} \qquad \forall j \in \mathcal{J}, \tag{14}$$

where $C_{j,0}$ is the number of patients served from queue $j \in \mathcal{J}$ in the data history, for example the previous planning period. $A^\alpha_{j,1}$ is the access time at the start of the planning period. If no history is available, then $u_j(1) = 1$ and $m_j(1) = 1 + \epsilon$, where ϵ is a small number.

Step 2: Determine β^n_j, for all $j \in \mathcal{J}$ and $n = 0, 1, 2, \ldots$, with (13). Solve the MIP with the obtained B.

Step 3: $s := s + 1$. Update $u_j(s)$ and $m_j(s)$, for all $j \in \mathcal{J}$, with

$$u_j(s) := \max\{0 + \epsilon, u_j(s-1) + \frac{1}{s}\left(\frac{\sum_{l=0}^{T-1} \omega_l \hat{c}_{j,l}}{\sum_{l=0}^{T-1} \omega_l C_{j,l}} - 1\right)\}, \quad j \in \mathcal{J}, \tag{15}$$

$$m_j(s) := \max\{1 + \epsilon, m_j(s-1) + \frac{1}{s}\left(\frac{\sum_{l=1}^{T} \omega_l A^\alpha_{j,l}}{\sum_{l=1}^{T} \omega_l \hat{a}^\alpha_{j,l}} - 1\right)\}, \quad j \in \mathcal{J}, \tag{16}$$

where ω_t are weights for different time periods $t \in \mathcal{T}$. In (15) and (16), we subtract 1 from the performance ratio outcome. When the subtraction results in a negative value, queue $j \in \mathcal{J}$ is *overperforming*, i.e., more resource capacities than required are allocated to this queue. This overperformance is mitigated by decreasing the parameters $u_j(s)$ and $m_j(s)$ in (15) and (16), which

cause the weights β_j^n for $n=0,1,\ldots$ and queue $j \in \mathcal{J}$ to decrease. Conversely, when a positive number is the result of subtracting 1 from the performance ratios, queue $j \in \mathcal{J}$ is *underperforming*, and the parameters $u_j(s)$ and $m_j(s)$ are increased. This results in increased weights β_j^n for $n=0,1,\ldots$ and queue $j \in \mathcal{J}$, which may increase the resource capacities that are allocated to queue $j \in \mathcal{J}$ to increase performance for queue j. By summing over all time periods in (15) and (16), we take into account performance over all time periods.

The weights ω_t can be used to emphasize results in particular time periods. For example by letting ω_t increase with t, one emphasizes the results that are obtained later in the planning period. Of course, the objective of these weights ω_t should match the application at hand. For example, a rolling horizon approach may not benefit from an emphasis on later time periods, because those later time periods are not actually implemented.

Step 4: If $\max\{|u_j(s)-u_j(s-1)|, |m_j(s)-m_j(s-1)|\} < \theta$, for all $j \in \mathcal{J}$, where θ is a small number, then stop, else repeat Steps 2-4.

The setup of the above iterative procedure is such that it leads to convergence of the weights in B. This follows from the fact that the terms between brackets in (15) and (16) are bounded. Changes in both $A_{j,t}^{\alpha}$ and $C_{j,t}$ in (15) and (16) are bounded by the limited availability of resource capacities. Since these terms are bounded, the changes in parameters $(u_j(s)-u_j(s-1)$ and $m_j(s)-m_j(s-1))$ are converging to 0 as they are multiplied by $\frac{1}{s}$ in (15) and (16). Therefore, the differences $u_j(s)-u_j(s-1)$ and $m_j(s)-m_j(s-1)$ are also converging to 0 in s. Hence, the stopping criterion is met at some s and therefore, the method converges.

In our approach, the calculation of the weights is separated from the MIP. This separation on the one hand prevents that the objective function of the MIP becomes quadratic. On the other hand, it prevents additional constraints in the MIP with regards to the weights. Another advantage of this separation is the clear distinction between calculating the weights based on explicit performance measures and calculating the patient admission plan with the MIP. This distinction provides the opportunity to determine the weights manually or with the described iterative procedure, which can be easily adapted to incorporate additional requirements.

3.4. Alternative performance metrics for tactical resource and admission planning

Recall from Section 1 that the main objectives of tactical planning are *to achieve equitable access and treatment duration for patient groups*, *to serve the strategically agreed target number of patients*, *to maximize resource utilization and to balance workload*. The priority given to different objectives of tactical planning may vary between hospitals and their particular environments. Hence, the model can be adapted and extended in various ways. In this section, we present performance measures that can be used to define alternative objective functions or to initialize and update the weights in the iterative procedure described in Section 3.3. We also show how these performance measures can be obtained from the solution of the modeled MIP.

Achieving equitable access and treatment duration for patient groups

- *Number of patients waiting longer than a norm.* The number of patients that wait longer than a certain norm $\hat{a}_{j,t}$ is measured as follows:

$$O_{j,t} = \sum_{n=\hat{a}_{j,t}+1}^{\infty} W_{j,t}^n, \qquad j \in \mathcal{J},\, t \in \mathcal{T}. \tag{17}$$

  The number of time periods that patients are waiting longer than the norm $\hat{a}_{j,t}$ may be measured as follows:

$$P_{j,t} = \sum_{n=\hat{a}_{j,t}+1}^{\infty} (n-\hat{a}_{j,t}) W_{j,t}^n, \qquad j \in \mathcal{J},\, t \in \mathcal{T}. \tag{18}$$

- *Access time.* Access time $A^{\alpha}_{j,t}$ is measured by (10) for all $j \in \mathcal{J}$, $t \in \mathcal{T}$ and α. The average access time $\bar{A}_{j,t}$ may be calculated by:

$$\bar{A}_{j,t} = \frac{\sum_{n=0}^{\infty} nW^n_{j,t}}{\sum_{n=0}^{\infty} W^n_{j,t}}, \qquad j \in \mathcal{J},\, t \in \mathcal{T}. \tag{19}$$

- *Performance ratio for access time.* The access time $A^{\alpha}_{j,t}$ in comparison with the access time target $\hat{a}^{\alpha}_{j,t}$ may be calculate by performance ratio $L^{A}_{\alpha,j,t}$ (11).
- *Total duration of a care pathway.* We may measure the duration $H^{\alpha}_{g,t}$ of a care pathway by summing over the access time in each stage as follows:

$$H^{\alpha}_{g,t} = \sum_{j \in J_g} A^{\alpha}_{j,t}, \qquad g \in \mathcal{G},\, t \in \mathcal{T}. \tag{20}$$

  The average duration $\bar{H}_{g,t}$ of a care pathway may be calculated as follows:

$$\bar{H}_{g,t} = \sum_{j \in J_g} \bar{A}_{j,t}, \qquad g \in \mathcal{G},\, t \in \mathcal{T}. \tag{21}$$

- *Performance ratio for duration of a care pathway.* We may evaluate the duration of a care pathway by aggregating the performance ratios in a care pathway's stages as follows:

$$L^{H}_{\alpha,g,t} = \frac{1}{e_g} \sum_{j \in \mathcal{J}_g} L^{A}_{\alpha,j,t}, \qquad g \in \mathcal{G},\, t \in \mathcal{T}. \tag{22}$$

Serving the strategically agreed number of patients

- *The number of patients served.* The number $C_{j,t}$ of patients served and a target $\hat{c}_{j,t}$ for the number of patients served are discussed in Section 3.3.
- *Performance ratio for the number of patients served.* The number $C_{j,t}$ of patients served in comparison with the target number $\hat{c}_{j,t}$ of patients served may be calculated by performance ratio $L^{C}_{j,t}$ (12).

Maximizing resource utilization and balancing workload

- *Fraction of resource capacities that are allocated to care pathways.* The fraction $\rho_{r,t}$ of resource capacities that are allocated to care pathways may be calculated by:

$$\rho_{r,t} = \frac{\sum_{j \in \mathcal{J}^r} s_{j,r} C_{j,t}}{\phi_{r,t}}, \qquad r \in \mathcal{R},\, t \in \mathcal{T}. \tag{23}$$

- *Resource allocation to a set $\mathcal{V}^r \subset \mathcal{J}^r$ of queues.* Hospital management may want to keep resource allocation $\gamma_{\mathcal{V}^r,t}$ to, or the number $\mu_{\mathcal{V}^r,t}$ of patients served in, a subset $\mathcal{V}^r \subset \mathcal{J}^r$ of queues consistent between time periods. These measures may be evaluated by:

$$\gamma_{\mathcal{V}^r,t} = \textstyle\sum_{j \in \mathcal{V}^r} s_{j,r} C_{j,t}, \qquad t \in \mathcal{T}, \tag{24}$$

$$\mu_{\mathcal{V}^r,t} = \textstyle\sum_{j \in \mathcal{V}^r} C_{j,t}, \qquad t \in \mathcal{T}, \tag{25}$$

  where $\mathcal{V}^r \subset \mathcal{J}^r$, for $r \in \mathcal{R}$.

3.5. Model extensions

In this section, we discuss various opportunities to extend the our method:

- Our dynamic approach makes it possible to respond appropriately to expected changes in patient demand or resource availability, but it may also result in varying patient admissions between different time periods. If necessary, this variation may be controlled by introducing additional constraints that limit the variation of the number $C_{j,t}$ of patient admissions between time periods $t \in \mathcal{T}$.
- Hospital management may want to bound the amount of resource capacities allocated to particular queues. For example, when doctors serve patients at the outpatient clinic and the operating room, a hospital manager may want to limit the capacity the doctor is allocated to the operating room based on operating room availability. To control or to balance the fraction of resource capacity that is allocated to a queue or a set of queues, constraints can be introduced.
- Already scheduled appointments may be included in the MIP. A constraint on the decision variables $C_{j,t}$ can ensure that the number of patients admitted at queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$ is larger or equal to the number of already scheduled appointments. The scheduled patients should also be incorporated in the waiting list to ensure feasibility of the MIP with regards to Constraint (4). One can also choose to disregard the already scheduled appointments in the MIP by reducing the resource capacity $\phi_{r,t}$ with the capacity required for the already scheduled appointments. Note that by excluding scheduled patients from the model, they are also omitted from the modeled waiting lists $W_{j,t}$ for $j \in \mathcal{J}$ and $t \in \mathcal{T}$.
- Hospital management can evaluate the performance of a given patient admission plan, for example a manual or a cyclical plan, by fixing the decision variables $C_{j,t}$ to the number of planned admissions in the given patient admission plan.

4. Test approach

The MIP and iterative method described in Section 3 are programmed in AIMMS 3.10, which uses ILOG CPLEX 12.1 to solve the MIP. To test our iterative method, we have implemented an instance generator that allows us to produce test instances with various parameter settings, based on examples from hospitals. Section 4.1 discusses the instance generator.

4.1. Instance generation

Table 2 lists the parameters that characterize and influence the complexity of the test instances. Some parameters influence problem size (e.g., the length of the planning horizon, the number of patient groups and the number of resource types), while other parameters influence the solution space (e.g., the initial waiting lists and the resource capacities). In our experiments, we do not take into account the delay matrix D, which has limited influence on the problem size and solution space.

The number T of time periods, the number R of resource types and the number $|\mathcal{J}|$ of queues determine the size of the MIP. The number $|\mathcal{J}|$ of queues is determined by the number of care pathways and the number of stages in each care pathway, as $|\mathcal{J}| = \sum_{g \in \mathcal{G}} e_g$.

For every instance, the values for the parameters in Table 2 are uniformly drawn from the possible values given in the third column of Table 2. We assume that new demand only arrives to the first queue in care pathways. We have three sets of values for the service time $s_{j,r}$, since these vary

Parameter	*Description*	*Used values for testing*
T	The number of time periods	$\{8\}$
R	The number of resource types	$\{2\}$
G	The number of care pathways	$\{6, 8, 10\}$
e_g	The number of stages in care pathway $g \in \mathcal{G}$	$\{3, 5, 7\}$
$s_{j,r}$	Expected service time from resource type $r \in \mathcal{R}$ for a patient in queue $j \in \mathcal{J}$ in time units (three value sets)	$\{10, 15, 20\}$, $\{100, 120, 140\}$, $\{200, 220, 240\}$
$\lambda_{j,t}$	New demand in queue $j \in \mathcal{J}$ in time period $t \in \mathcal{T}$	$\{2, 6, 10\}$
q_{ij}	The routing probabilities between queue $i, j \in \mathcal{J}$	$\{0, 0.25, 0.5, 0.75, 1\}$
$\hat{a}_{j,t}$	Target access time for queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$	$\{1, 2, \ldots, 8\}$
$\hat{c}_{j,t}$	Target number of served patients for queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$	$\{2, 3, \ldots, 10\}$
$\hat{c}_{j,0}$	Target number of served patients for queue $j \in \mathcal{J}$ in the previous planning period	$\{10, 30, 50\}$
$C_{j,0}$	The number of served patients for queue $j \in \mathcal{J}$ in the previous planning period	$\{10, 30, 50\}$
$\bar{n}_j$	The number of time periods the longest-waiting patients have been waiting on the initial waiting list for queue $j \in \mathcal{J}$	$\{1, 2, \ldots, 16\}$

Table 2 The parameters that characterize the test instances.

between different services (e.g., consultations, MRI scans and surgeries). The three sets correspond to a low, medium and high service time respectively.

We first generate $C_{j,0}$, i.e., the number of patients served in queue j in the previous planning period. We start by generating $C_{j,0}$ for the first queue in the care pathway. For all subsequent queues in the care pathway, we draw $C_{j,0}$ from $\left[0.75 \sum_{i \in \mathcal{J}} q_{ij} C_{i,0}, 1.25 \sum_{i \in \mathcal{J}} q_{ij} C_{i,0}\right]$. A similar approach is applied in generating $\hat{c}_{j,0}$ and $\hat{c}_{j,t}$, for all $t \in \mathcal{T}$. We first generate $\hat{c}_{j,0}$ and $\hat{c}_{j,t}$, for all $t \in \mathcal{T}$, for the first queue in the care pathway. For all subsequent queues in the care pathway, we choose $\hat{c}_{j,0} = \sum_{i \in \mathcal{J}} q_{ij} \hat{c}_{i,0}$ and $\hat{c}_{j,t} = \sum_{i \in \mathcal{J}} q_{ij} \hat{c}_{i,t}$, for all $t \in \mathcal{T}$.

We then generate the initial waiting list $W_{j,1} = \left(W_{j,1}^0, W_{j,1}^1, \ldots\right)$. $W_{j,1}$ represents the waiting list at the start of the planning period, because the waiting list $W_{j,1}$ is calculated before patients are served in this time period. First, we draw $\bar{n}_j$, which indicates the number of time periods the longest-waiting patients have been waiting on the initial waiting list of queue $j \in \mathcal{J}$. Then, we determine the number $W_{j,1}^n$ of patients waiting n time periods by:

$$W_{j,1}^n = \frac{b_j}{n}, \qquad j \in \mathcal{J},\ 0 < n \leq \bar{n}_j.$$

where b_j is calculated as follows. We first generate b_j for the first queue in the care pathway by:

$$b_j = \frac{\sum_{t \in \mathcal{T}} \lambda_{j,t}}{T}, \qquad j \in \mathcal{J},\ t \in \mathcal{T}.$$

For all subsequent queues in the care pathway, we draw b_j from $\left[0.75 \sum_{i \in \mathcal{J}} q_{ij} b_i, 1.25 \sum_{i \in \mathcal{J}} q_{ij} b_i\right]$. By dividing by n in (26), the number $W_{j,1}^n$ of patients waiting n time periods decreases as n grows. This structures the initial waiting list $W_{j,1}$ for each $j \in \mathcal{J}$ to resemble waiting lists observed in practice.

To determine the resource capacities $\phi_{r,t}$ for each resource type $r \in \mathcal{R}$ and time period $t \in \mathcal{T}$, we first approximate the amount $\tilde{\phi}_r$ of resources required in the current planning period by summing

the amount of resources required by arriving patients $\lambda_{j,t}$, for all $t \in \mathcal{T}$, throughout their care pathways. Using $\tilde{\phi}_r$ and a tuning parameter κ_r, we determine $\phi_{r,t}$ by:

$$\phi_{r,t} = \kappa_r \frac{\tilde{\phi}_r}{T}, \qquad r \in \mathcal{R},\, t \in \mathcal{T}. \tag{26}$$

Unless stated otherwise, we assume $\kappa_r = 1$, for all $r \in \mathcal{R}$. The method's sensitivity to varying capacity dimensions is examined by varying κ_r in the computational experiments.

We bound the computation time for the MIP by 100 seconds. For the procedure to determine the weights, we set the following entries $\alpha = 0.9$, $\epsilon = 0.01$, $\theta = 0.01$ and $\omega_t = 1$, for all $t \in \mathcal{T}$. The latter indicates that we give the same weight to each time period.

5. Results

We use the performance measures introduced in Section 3 to evaluate the proposed method for tactical resource and admission planning. We generate 300 instances following the procedure of Section 4. For each queue and time period in the 300 generated instances, we calculate the three performance ratios for access time, the number of patients served and total duration of a care pathway by (11), (12) and (22) respectively. For each type of performance ratio and each time period, we generate one list of the calculated ratios in all instances. Subsequently, these lists are sorted in ascending order. The sorted lists can be used to evaluate each type of performance ratio at a given percentile for each time period. For example, when there are 3000 ratios on a sorted list, the 300-th entry represents the 10-th percentile. When we curve these percentiles and the curve decreases (increases) for successive time periods, we know that for a given fraction of the queues in all 300 instances, the performance ratio decreases (increases). Below, we present our results for each tactical planning objective.

Achieving equitable access and treatment duration for patient groups
The curves in Figure 2 display the percentiles for the performance ratios $L^A_{0.9,j,t}$ for the access time in all queues in all instances. The curves show that resource capacities are allocated such that the performance ratios $L^A_{0.9,j,t}$ for access times become less variable, as the range between the 20-th and 80-th percentiles decreases and stabilizes over time periods. Hence, we may conclude that resources are more equitably divided over queues during the planning period, leading to less variation in performance ratios.

The performance ratios tend toward a number above 1, because the total resource capacity $\phi_{r,t}$ per resource $r \in \mathcal{R}$ in time period $t \in \mathcal{T}$ is sufficient to serve new demand, but not the already existing waiting list $W_{j,0}$. When κ_r in (26) is increased, more capacity is available to serve new demand and the existing waiting list. As a result, the performance ratios in the graph in Figure 3 tend towards a lower number than the performance ratios in the graph in Figure 2. In this case, they tend toward 1, which indicates that resource capacities are allocated such that access times for a higher fraction of queues are closer to target.

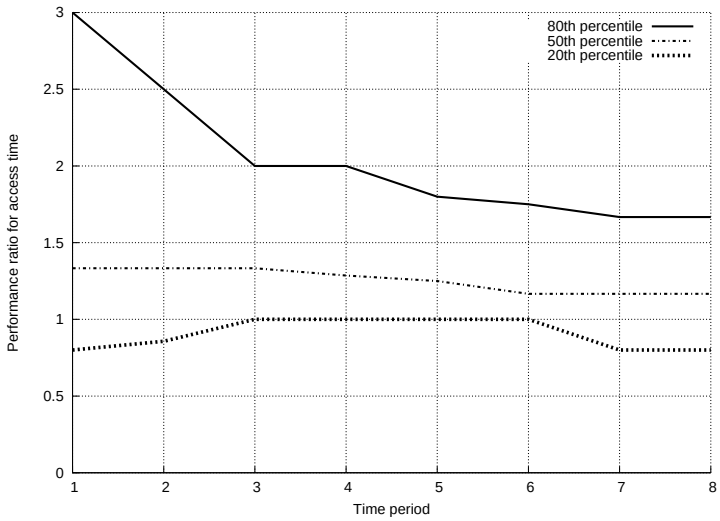

Figure 2 The 20-th, 50-th and 80-th percentiles of the performance ratios $L^A_{0.9,j,t}$ for access time for all queues in all instances.

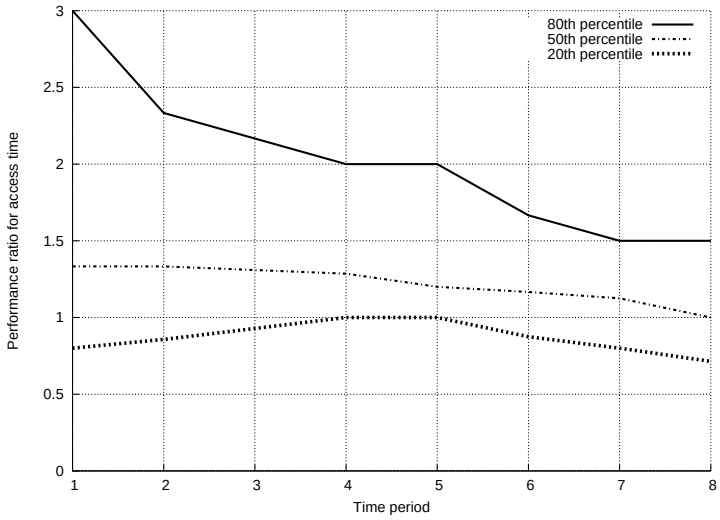

Figure 3 The 20-th, 50-th and 80-th percentiles of the performance ratios $L^A_{0.9,j,t}$ for access time for all queues in all instances, when $\kappa_r = 1.1$ for all $r \in \mathcal{R}$.

The curves in Figure 4 display the percentiles for the performance ratios $L^H_{0.9,g,t}$ for the duration of a care pathway for all care pathways in all instances. The method allocates resources such that the performance ratios $L^H_{0.9,g,t}$ for the care pathway durations tend towards 1. We may conclude that the duration of a care pathway is closer to target for a larger fraction of care pathways.

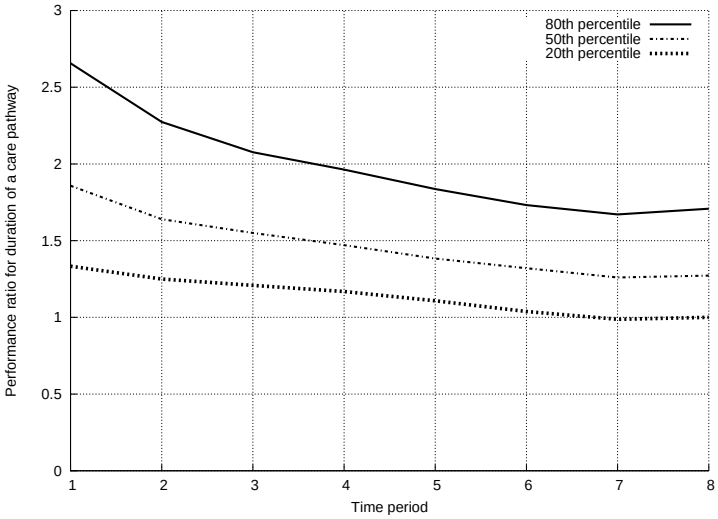

Figure 4 The 20-th, 50-th and 80-th percentiles of the performance ratios $L^H_{0.9,g,t}$ for the duration of a care pathway for all care pathways in all instances.

Serving the strategically agreed target number of patients
The curves in Figure 5 display the percentiles for the performance ratios $L^C_{j,t}$ for the number of patients served in all queues in all instances. Resources are allocated such that the performance ratios $L^C_{j,t}$ for the number of patients served are less variable and tend toward 1. This indicates that resource capacities are allocated such that the number of patients served for a higher fraction of queues are closer to target.

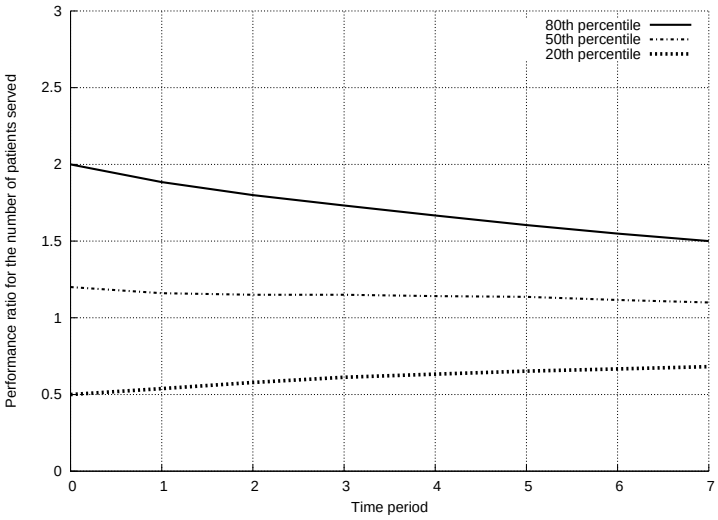

Figure 5 The 20-th, 50-th and 80-th percentiles of the performance ratios $L^C_{j,t}$ for the number of patients served for all queues in all instances.

Maximizing resource utilization and balancing workload
The fraction $\rho_{r,t}$ of resource capacities $r \in \mathcal{R}$ that are allocated to care pathways in time period $t \in \mathcal{T}$ can be used to identify bottleneck and underutilized resources. Graphing these percentages supports this identification. For example, the histogram in Figure 6 shows a decline in the percentage of

resource capacity that is allocated to care pathways in time periods $t = 3$ and $t = 4$. Hospital management can use these histograms to decide on patient admission policies, or to dimension and allocate resource capacities. In addition, the fraction $\rho_{r,t}$ of resource capacities that are allocated

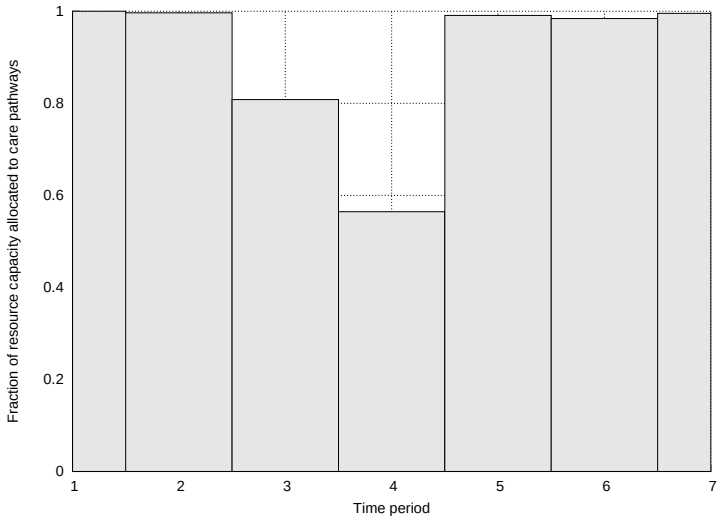

Figure 6 Example of the fraction $\rho_{r,t}$ of resource capacities allocated to care pathways for a resource type in an instance.

to care pathways can be used to evaluate the workload balance. For example, the workload is significantly lower in time periods $t = 3$ and $t = 4$ for the resource depicted in the graph of Figure 6. Constraints may be introduced in the MIP to improve the workload balance in the planning period. These constraints are discussed in Section 3.5.

The calculation time for relatively large instances ($G = 10,\ e_g = 5, \forall g \in \mathcal{G},\ T = 8,\ R = 2$) is 16 minutes, which may be assumed to be reasonable for a tactical planning method. Furthermore, the average integrality gap for all generated instances is 0.44%.

6. Conclusion and discussion

Inspired by multiple hospitals that are investigating the potential use of tactical planning, we have developed an iterative method that can be used dynamically to develop mid-term tactical resource and admission plans for real-life sized instances. These tactical resource and admission plans allocate resource capacity over care pathways and determine the number of patients to serve at a particular stage of their care pathway.

Computational results show that our method improves compliance with targets for access times, care pathway duration and the number of patients served. The method is a tool for hospital management to achieve equitable access and treatment duration for patient groups and to serve the strategically agreed target number of patients. Within this framework, the method can be adapted to maximize resource utilization and/or to balance workload. It may be used to identify bottleneck resources or underutilized resources, and for scenario analysis in anticipation of peaks in patient demand or resource (un)availability. This allows a timely response, such as temporarily increasing or decreasing resource capacities to improve access times and workload balance.

The method integrates decision making for multiple resources, multiple time periods and multiple patient groups with various uncertain care pathways. Care pathways connect multiple departments

and resources into a network and fluctuations in both patient arrivals (e.g., seasonality) and resource availability (e.g., holidays) result in bullwhip effects in the care chain. Therefore, coordinated decision making along a care chain of hospital resources offers improvement potential.

The basic elements of the tactical planning problem in health care also occur in other industries. Since our method can be extended and adapted easily, it can be used in other service and manufacturing environments. Alternative constraints and objective functions may better fit the objectives of tactical planning of a particular organization. Hence, we have mentioned that various other performance measures can be used to develop alternative objective functions and that various possible extensions of the model may be of interest, including constraints to balance the number of patient admissions and resource capacities allocated to particular care pathways over time, and the incorporation of already scheduled patients. These extensions are interesting topics for further research.

Acknowledgments

This research is inspired by multiple Dutch (academic) hospitals, a.o. 'Reinier de Graaf Groep', 'Zorg Groep Twente', 'Deventer Ziekenhuis', 'Medisch Spectrum Twente' and 'Universitair Medisch Centrum Utrecht'. We thank involved clinical staff and managers from these hospitals.

References

[1] I. Adan, J. Bekkers, N. Dellaert, J. Vissers, and X. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science*, 12(2):129–141, 2009.

[2] J. Beliën and E. Demeulemeester. Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, 176(2):1185–1204, 2007.

[3] D. Bertsimas and J. Niño-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

[4] S. Brailsford and J. Vissers. OR in healthcare: a European perspective. *European Journal of Operational Research*, 212(2):223–234, 2011.

[5] P. Brucker, A. Drexl, R. Möhring, K. Neumann, and E. Pesch. Resource-constrained project scheduling: Notation, classification, models, and methods. *European Journal of Operational Research*, 112(1):3–41, 1999.

[6] B. Cardoen and E. Demeulemeester. Capacity of clinical pathways - a strategic multi-level evaluation tool. *Journal of Medical Systems*, 32(6):443–452, 2008.

[7] T. Cayirli and E. Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

[8] E. Cerdá, L. Pablos, and M. Rodriguez. Waiting lists for surgery. *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 151–187, 2006.

[9] M. J. Cote. Patient flow and resource utilization in an outpatient clinic. *Socio-Economic Planning Sciences*, 33(3):231–245, 1999.

[10] B. T. Denton, A. J. Miller, H. J. Balasubramanian, and T. R. Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations research*, 58(4-Part-1):802–816, 2010.

[11] S. G. Elkhuizen, S. F. Das, P. J. M. Bakker, and J. A. M. Hontelez. Using computer simulation to reduce access time for outpatient departments. *British Medical Journal*, 16(5):382, 2007.

[12] L. Garg, S. McClean, B. Meenan, and P. Millard. A non-homogeneous discrete time markov model for admission scheduling and resource planning in a cost or capacity constrained healthcare system. *Health Care Management Science*, 13:155–169, 2010.

[13] P. Gemmel and R. Van Dierdonck. Admission scheduling in acute care hospitals: does the practice fit with the theory? *International Journal of Operations and Production Management*, 19:863–878, 1999.

[14] S. C. Graves. A tactical planning model for a job shop. *Operations Research*, 34(4):522–533, 1986.

[15] L. V. Green, S. Savin, and B. Wang. Managing patient service in a diagnostic medical facility. *Operations Research*, 54(1):11–25, 2006.

[16] R. W. Hall. *Patient flow: reducing delay in healthcare delivery*. Springer Verlag, 2006.

[17] E. W. Hans. *Resource loading by branch-and-price techniques*. PhD thesis, University of Twente, 2001.

[18] E. W. Hans, M. Van Houdenhoven, and P. J. H. Hulshof. A framework for health care planning and control. *Memorandum No. 19571, University of Twente, Department of Mathematical Sciences*, 2011.

[19] J. C. Hershey, E. N. Weiss, and M. A. Cohen. A stochastic service network model with application to hospital facilities. *Operations Research*, 29(1):1–22, 1981.

[20] J. R. Jackson. Jobshop-like queueing systems. *Management Science*, 50(12):1796–1802, 2004.

[21] A. S. Kapadia, S. E. Vineberg, and C. D. Rossi. Predicting course of treatment in a rehabilitation hospital: a markovian model. *Computers & Operations Research*, 12(5):459–469, 1985.

[22] H. L. Lee, V. Padmanabhan, and S. Whang. The Bullwhip Effect In Supply Chains. *Sloan Management Review*, 38(3):93–102, 1997.

[23] L. G. N. Nunes, S. V. de Carvalho, and R. C. M. Rodrigues. Markov decision process applied to the control of hospital elective admissions. *Artificial Intelligence in Medicine*, 47(2):159–171, 2009.

[24] Organisation of Economic Co-operation and Development (OECD). *Data retrieved October 10, 2010, from: http://www.oecd.org/health*, 2010.

[25] M. E. Porter and E. O. Teisberg. How physicians can change the future of health care. *Journal of the American Medical Association*, 297(10):1103, 2007.

[26] J. M. van Oostrum, M. Van Houdenhoven, J. L. Hurink, E. W. Hans, G. Wullink, and G. Kazemier. A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR spectrum*, 30(2):355–374, 2008.

[27] I. B. Vermeulen, S. M. Bohte, S. G. Elkhuizen, H. Lameris, P. J. M. Bakker, and H. L. Poutré. Adaptive resource allocation for efficient patient scheduling. *Arificial Intelligence in Medicine*, 46(1):67–80, 2009.

[28] E. N. Weiss, M. A. Cohen, and J. C. Hershey. An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital patient flow. *Operations Research*, 30(6):1082–1104, 1982.

[29] G. Wullink. *Resource loading under uncertainty*. PhD thesis, University of Twente, 2005.

[30] R. Y. T. Yeung, G. M. Leung, S. M. McGhee, and J. M. Johnston. Waiting time and doctor shopping in a mixed medical economy. *Health economics*, 13(11):1137–1144, 2004.

[31] W. H. M. Zijm. Towards intelligent manufacturing planning and control systems. *OR Spectrum*, 22(3):313–345, 2000.