Health Care Manag Sci (2013) 16:152–166
DOI 10.1007/s10729-012-9219-6

Tactical resource allocation and elective patient admission planning in care processes

Peter J. H. Hulshof · Richard J. Boucherie · Erwin W. Hans · Johann L. Hurink

Received: 25 July 2012 / Accepted: 4 November 2012 / Published online: 4 January 2013

Abstract 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 processes and determine the selection of patients to be served that are at a particular stage of their care process. Our method is developed in a Mixed Integer Linear Programming (MILP) framework and copes with multiple resources, multiple time periods and multiple patient groups with various uncertain treatment paths through the hospital, 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 fraction of allocated resource capacity. Our approach is generic, as the base MILP 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.

Keywords Health care · Tactical planning · Resource capacity planning · Patient admission planning · Mixed Integer Linear Programming (MILP)

1 Introduction

Tactical planning is a key element of hospital planning and control that concerns the intermediate term allocation of resource capacities and elective patient admission planning. 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.

This research was inspired by many hospitals in the Netherlands. The hospitals we cooperate with have the aim to provide equitable access and treatment duration for patient groups by controlling access times. Access time is the time a patient spends 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 [37]. Access time is incurred at each care stage in a patient's treatment at the hospital, for example before an outpatient clinic visit and before surgery. Also, in some reimbursement systems, hospitals receive payments only after patients have completed their health care process. Hence, it can be costly for hospitals when patients have to wait, as resources and materials have already been invested, but revenues are still to come.

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.

P. J. H. Hulshof (✉) · R. J. Boucherie · E. W. Hans (✉) · J. L. Hurink
Center for Healthcare Operations Improvement and Research (CHOIR), University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: p.j.h.hulshof@utwente.nl
e-mail: e.w.hans@utwente.nl

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 break the clinician's time down over separate activities (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 process (e.g, consultation or surgery). We use the term care process in this article to identify a chain of care stages for a patient. These care stages constitute a patient's logistical treatment path. For example, a care process may comprise a visit to an outpatient clinic, a surgery and a revisit to the outpatient clinic. Because care processes 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 in the care chain [32]. From a patient's perspective, this means access times for each separate stage in a care process strongly fluctuate. 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 [10, 21, 31], 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.

Typically, tactical planning is done for a subset of care processes in a hospital (e.g., one specialty, a subset of specialties), and not for the entire hospital. Tactical planning problems observed at the hospitals we cooperate with typically comprise 6–10 care processes, 4–8 weeks as a planning horizon, and 1–3 resource types. In these hospitals, tactical planning is organized around a biweekly meeting with decision makers involved in developing the tactical plans. These meetings are used to develop and adjust the tactical resource and admission plans for future time periods, based on information subtracted from the hospital information system about waiting lists, resource availability, expected demand and the number of patients served in prior periods. In this way, tactical resource and admission plans are developed in response to anticipated changes in demand or supply on the mid-term, which leads to improved utilization, shorter access times and improved control of the number of patients served. Currently, the decision makers are using spreadsheet solutions to base their tactical resource and admission plans on. Our model provides an optimization step that supports rational decision making in tactical planning. The model can be used to propose a tactical resource and admission plan to the decision makers, and to evaluate particular scenarios with regards to foreseen resource (un)availability, a proposed change in access time targets, expected demand surges, etc.

The available approaches on the development of tactical resource and admission plans in the Operations Research and Management Science 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. Our aim is to provide a theoretical contribution to the development of tactical resource and admission plans in health care. 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 processes, thereby integrating decision making for a chain of hospital resources. This paper is not about developing clinical care pathways, as described in [16], but about methods for the logistical coordination (i.e., allocation and planning) of resource capacities in patient care processes. Clinical care pathways may be used to identify the patient care processes, but the identification of patient care processes is not the main focus of this paper.

The resource capacity and admission plans are provided for each stage in the care process, this includes for example the outpatient clinic and the operating theater. The method incorporates available knowledge about the state of the waiting lists and the available resource capacities. To test our method, we use it to develop tactical resource and admission plans for generated instances that are inspired by practical problem instances provided by the hospitals we cooperate with. 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 process 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 discusses 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 discusses the managerial and practical implications of developing tactical resource and admission plans. Section 7 concludes this paper.

2 Background

Due to increasing demand for health care and increasing expenditures [29], health care organizations are trying to organize processes more efficiently and effectively. Planning and control in health care has received an increased amount of attention over the last ten years [6], 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 [2, 5, 9, 23]. While strategic planning addresses the dimensioning of resource capacities, tactical planning subdivides the settled resource capacities among patient groups (e.g., identified by specialty) 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 [34], 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 process for patient groups. For example, they focus on the outpatient clinic [11, 15], diagnostic services [20, 34] or operating rooms [1, 3, 12, 14, 33]. Although the benefits of an integrated approach are often recognized [10, 21, 31], relatively few articles integrate decision making for a chain of resources or departments along the patient's care process. To support integrated tactical resource and admission planning, [27] models care processes as Markov chains to derive resource requirements for each stage of a patient's care process. Similar approaches for evaluation of resource requirements are taken in [13, 17, 24, 35]. In order to calculate optimal static, elective patient admission plans for multiple resources and multiple patient groups with various care processes, [28] 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 [18], which is a network of work stations capable of producing a wide variety of jobs [19]. Hence, methods used in industry for job shop scheduling may be suitable for tactical resource and admission planning in health care. Job shop scheduling is applied to surgical care services in [25, 30]. In these papers, mathematical programming is used to allocate surgical resources and to schedule surgical patients. Queueing models can also be used to analyze tactical production plans for a job shop [19, 26]. 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 [7, 8]. Methods to allocate resources to activities in project scheduling are often based on mathematical programming [22, 36, 38] or MDP [4].

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 processes.

3 Model description

We aim to allocate resource capacities among the various consecutive stages of different care processes. To this end, we propose a Mixed Integer Linear Program (MILP) 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 process. 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 MILP to form alternative objective functions in Section 3.4. In addition, we also discuss other extensions of the model in Section 3.4. In the following, we introduce notation and discuss the problem in more detail.

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 processes $\mathcal{G} = \{1, 2, \ldots, G\}$. 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 process $g \in \mathcal{G}$ 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 process. Patients following the same care process have the same resource requirements in each stage (e.g., consultation times, surgery duration, number of consultations), and to serve a patient of care process $g \in \mathcal{G}$ 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 process 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 process, we obtain a set of queues $\mathcal{J}_g$ of cardinality e_g. Although patients of different care processes share resources for service, we model the queues disjointly for different care processes. Consequently, we have a total set of queues $\mathcal{J} = \bigcup_{g \in \mathcal{G}} \mathcal{J}_g$, where $|\mathcal{J}| = \sum_{g \in \mathcal{G}} e_g$.

For each time period $t \in \mathcal{T}$, we determine a patient admission plan, characterized by the decision variable vectors $C_{j,t} = (C^0_{j,t}, C^1_{j,t}, \ldots)$. 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 $W_{j,t} = (W^0_{j,t}, W^1_{j,t}, \ldots)$, 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.

For ease of notation we summarize the transition rates between the stages/queues in a routing matrix Q of dimension $|\mathcal{J}| \times |\mathcal{J}|$. Furthermore, to be able to take into account a minimum (required) time lag before patients that have been served at one queue, can enter the following queue, we introduce a delay matrix D of dimension $|\mathcal{J}| \times |\mathcal{J}|$, where the entry d_{ij} denotes the minimum time lag (in time periods) between service from queue i and entrance to queue j ($i, j \in \mathcal{J}$). Such deterministic delay d_{ij} may for example be specified by a doctor, when a given time lag between two stages is medically required in the care process (e.g., such that patients can recover from a procedure). 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:

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

Assumptions 1–5 summarize the problem assumptions that underly our modeling approach, which is presented in Section 3.1.

Assumption 1 *Patient arrivals, delay times, resource requirements and resource capacities are considered to be deterministic and known.*

Assumption 2 *All patients arriving to a queue remain in the queue until service completion.*

Assumption 3 *All patients in queue j require resources $s_{j,r}$, $r \in \mathcal{R}$.*

Assumption 4 *Resource capacity $\phi_{r,t}$ for resource type r and time period t is not transferable from one time period to another time period $s \neq t$; $s, t \in \mathcal{T}$, i.e., when (part of) the resource capacity $\phi_{r,t}$ is unused in time period t, it is 'lost'.*

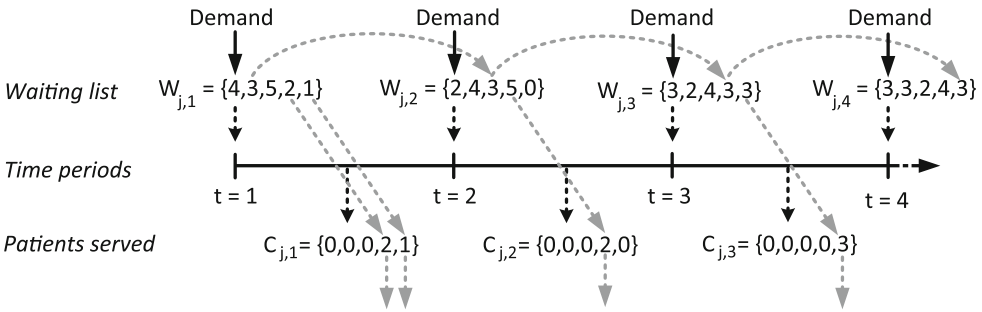

Fig. 1 The dynamics of the waiting list and patient service for a system with a single queue $j \in \mathcal{J}$

Assumption 5 *Every patient planned according to the decision $C_{j,t}$ will be served in queue j in period t, i.e., there is no deferral to other time periods.*

3.1 Constraints to calculate a tactical resource and admission plan

The constraints to model the care processes 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 3.4.

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

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

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

$$\sum_{j \in \mathcal{J}^r} s_{j,r} C_{j,t} \leq \phi_{r,t} \qquad \forall r \in \mathcal{R}, t \in \mathcal{T}, \tag{5}$$

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

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

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 In Constraint (6) the number $C_{j,t}$ of patients that are served from queue j at time period t is calculated. We only require $C_{j,t}$ to be integer and not the entry $C^n_{j,t}$, which expresses the number of patients that are waiting already n time periods. The reason is that the entries $C^n_{j,t}$ are related to $W^n_{j,t}$ by Constraints (2)–(4), and that the $W^n_{j,t}$ may have noninteger values. Based on the integer constraint on $C_{j,t}$, only 'full' patients are served in the model.

Remark 2 For numerical purpose, to solve our optimization problem, the number n of time periods that patients are waiting is bounded at some value N. Consequently,

Table 1 The sets, indices, variables and parameters used

Sets	
$\mathcal{J}$	Queues
$\mathcal{T}$	Time periods
$\mathcal{R}$	Resource types
$\mathcal{J}^r$	Queues for resource type r, $\mathcal{J}^r \subseteq \mathcal{J}$
Indices	
$i, j \in \mathcal{J}$	Queue
$t \in \mathcal{T}$	Time period
$r \in \mathcal{R}$	Resource type
$i, j \in \mathcal{J}^r$	Queue
n, d	Time periods (to indicate waiting time)
Variables	
Decision variables	
$C^n_{j,t}$	The number of patients served from queue j in time period t, who have been waiting n time periods
$C_{j,t}$	The total number of patients served from queue j in time period t
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

Constraints (2)–(6) require adaptation and a constraint is added to stipulate that the number $W^N_{j,t}$ of patients who are not served in time period $t-1$ and are waiting N time periods, remain on the waiting list in time period t:

$$W^N_{j,t} = \sum_{m=N-1}^{N} \left(W^m_{j,t-1} - C^m_{j,t-1} \right), \quad \forall j \in \mathcal{J}, t \in \mathcal{T}. \tag{8}$$

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 Section 3.4.

We use the following objective function:

$$\min \quad \sum_{j \in \mathcal{J}} \sum_{n=0}^{\infty} \sum_{t \in \mathcal{T}} \beta^n_j W^n_{j,t}. \tag{9}$$

The objective function (9) aggregates the weighted number of patients waiting in each queue $j \in \mathcal{J}$ in each time period $t \in \mathcal{T}$. Note that patients appear multiple times in the summation over $W^n_{j,t}$. To illustrate this, consider the following two cases: (1) a served patient may move from $W^n_{j,t}$ to $W^0_{i,t+1}$ (if $q_{ji} > 0$), and (2) a patient that is not served moves from $W^n_{j,t}$ to $W^{n+1}_{j,t+1}$. If $t, t+1 \in \mathcal{T}$, then the four mentioned waiting lists are in the objective function's summation. Weights β^n_j ($j \in \mathcal{J}$ 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 MILP'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 MILP and measure the metrics as we explain below, then update the weights, solve the MILP, etc. In this section we first explain how we measure the performance metrics from the MILP solution, and then explain in detail the iterative procedure of determining the weights.

Access time As mentioned in Section 1, access time is the time a patient spends on the waiting list before being served. The elements $W^n_{j,t}$ in our MILP provide information about the structure of the waiting list for each queue j in each time period t. We get insight into access time by measuring $A^\alpha_{j,t}$ from the MILP solution as follows:

$$A^\alpha_{j,t} = \min \left\{ n | \sum_{m=0}^{n} W^m_{j,t} > \alpha \sum_{m=0}^{\infty} W^m_{j,t} \right\}, \quad j \in \mathcal{J}, t \in \mathcal{T}, \tag{10}$$

where α is a given percentile. $A^\alpha_{j,t}$ in Eq. (10) gives the number of periods that the α-th percentile of all patients in a queue j are waiting. In other words, a fraction of $(1-\alpha)$ of all patients in queue j at time period t have been waiting already for at least $A^\alpha_{j,t}$ time periods.

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

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

We use this ratio to evaluate how close to target the performance of the current solution is. For example, if $L^A_{\alpha,j,t} > 1$, then $A^\alpha_{j,t}$ is 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 processes, by setting the target for either the first or the last queue in care processes. In our model, we assume that these care process targets can be converted to targets for each stage of a care process.

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 $n = 0, 1, 2, \ldots$ indicates the number of time periods waiting, lead to a matrix B. Two assumptions are made regarding the structure of B.

Assumption 6 *$\beta^n_j < \beta^{n+1}_j$, for all $j \in \mathcal{J}$ and $n = 0, 1, 2 \ldots$*

Assumption 7 *If $q_{ij} > 0$, then $\max_n \beta^n_i > \min_n \beta^n_j$, for all $i, j \in \mathcal{J}$.*

Remark 3 Under Assumption 6, $\min_n \beta^n_j = \beta^0_j$, 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 MILP 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:

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

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 6. Following Remark 3, by setting $\beta^0_j = 0$, we ensure that Assumption 7 holds.

Taking into account Assumptions 6 and 7, 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 MILP is solved. After that, B is updated based on the solution of the MILP, and the MILP is solved again with the updated B. This iterative way of updating B and solving the MILP 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 MILP.
3. The stopping criterion.

The following iterative procedure is used to initialize and update B. In B there are at most $|\mathcal{J}| \times (N + 1)$ elements that require initializing and updating. By using Eq. (13), we need to adjust at most $2 \times |\mathcal{J}|$ 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 Eq. (13) and the MILP is solved. Based on the MILP 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:

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

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. $A^\alpha_{j,1}$ gives the evaluation of Eq. (10) 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_j^n, for all $j \in \mathcal{J}$ and $n = 0, 1, 2, \ldots$, with Eq. (13). Solve the MILP with the obtained B.

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

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

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

where ω_t are weights for different time periods $t \in \mathcal{T}$.

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.

In Eqs. (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 Eqs. (15) and (16), which causes the weights β_j^n for $n = 0, 1, \ldots$ and queue $j \in \mathcal{J}$ to decrease. This may decrease the allocated resource capacity to this queue, for example when the involved resource capacity can be used to improve performance in other queues. 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 Eqs. (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.

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 Eqs. (15) and (16) are bounded. Changes in both $A^{\alpha}_{j,t}$ and $C_{j,t}$ in Eqs. (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 Eqs. (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 MILP. This separation on the one hand prevents that the objective function of the MILP becomes quadratic. On the other hand, it prevents additional constraints in the MILP 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 MILP. 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 and model extensions

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 MILP.

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.* With Eq. (10), the measure $A^{\alpha}_{j,t}$ can be calculated for all α. The average $\bar{A}_{j,t}$ for this measure $A^{\alpha}_{j,t}$ may be calculated by:

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

– *Access time performance ratio.* $A^{\alpha}_{j,t}$ may be compared to its target $\hat{a}^{\alpha}_{j,t}$ by calculating the access time performance ratio $L^{A}_{\alpha,j,t}$ with Eq. (11).
– *Total access time of a complete care process.* We get insight in the total access time of a complete care process (i.e., all queues/stages $\mathcal{J}_g$ in the care process) by summing over $A^{\alpha}_{j,t}$ in each stage as follows:

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

The average $\bar{H}_{g,t}$ of this measure may be calculated as follows:

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

– *Access time performance ratio for a care process.* We may get insight in the access time performance ratio for a care process by aggregating the access time performance ratios in a care process'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 $L^{C}_{j,t}$ with Eq. (12).

Maximizing resource utilization and balancing workload

– *Fraction of resource capacities that are allocated to care processes.* The fraction $\rho_{r,t}$ of resource capacities that are allocated to care processes 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:

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

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

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

In addition to using alternative metrics in the model, there are also various opportunities for extending the model. Four examples of those opportunities are discussed below.

Constraints to limit variation of patient admissions 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 $t \in \mathcal{T}$.

Constraints to limit resource allocation to particular queues 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.

Previously scheduled appointments Previously scheduled appointments may be included in the MILP. 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 MILP with regards to Constraint (4). One can also choose to disregard the already scheduled appointments in the MILP 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 $t \in \mathcal{T}$.

Evaluation of given admission plans 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 MILP and iterative method described in Section 3 are programmed in AIMMS 3.10, which uses ILOG CPLEX 12.1 to solve the MILP. 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

This section describes the instance generation procedure. Various parameter settings can be used to influence the test instances that are generated, in order to align these test instances with the examples from practice. In the hospitals we cooperate with, tactical planning is typically done for a subset of care processes in a hospital (e.g., one specialty, a subset of specialties), and not for the entire hospital. Tactical planning problems at the hospitals we cooperate with typically comprise 6–10 care processes (G), 4–8 weeks (T) and 1–3 resource types (R). For example, the care processes of an orthopedic surgery group may comprise hip surgery, shoulder surgery, knee surgery, etc. Care stages in each care process may for example be described by the initial outpatient clinic visit, preanesthesia visit, surgery, and a follow-up outpatient clinic visit. Typical resources that are involved in each care process are for example a clinician, a nurse, and the allocated operating room time.

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 and the number $|\mathcal{J}|$ of queues principally determine the size of the MILP. The number $|\mathcal{J}|$ of queues is determined by the number of care processes and the number of stages in each care process, 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 processes. We have three sets of values for the service time $s_{j,r}$, since these vary 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 process. For all subsequent queues in the care process, 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 process. For all subsequent queues in the care process, 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 $W_{j,1} = (W^0_{j,1}, W^1_{j,1}, \ldots)$. $W_{j,1}$ 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 $W^n_{j,1}$ of patients waiting n time periods by:

$$W^n_{j,1} = \frac{b_j}{n}, \qquad j \in \mathcal{J}, 0 < n \leq \bar{n}_j. \tag{26}$$

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

$$b_j = \frac{\sum_{t \in \mathcal{T}} \lambda_{j,t}}{T}, \qquad j \in \mathcal{J}, t \in \mathcal{T}. \tag{27}$$

For all subsequent queues in the care process, 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 Eq. (26), the number $W^n_{j,1}$ 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 processes. Using $\tilde{\phi}_r$ and a tuning parameter κ_r, we determine $\phi_{r,t}$ by:

$$\phi_{r,t} = \kappa_r \frac{\tilde{\phi}_r}{T}, \qquad r \in \mathcal{R}, t \in \mathcal{T}. \tag{28}$$

Table 2 The parameters that characterize the test instances

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 processes	{6, 8, 10}
e_g	The number of stages in care process $g \in \mathcal{G}$	{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 for $A^{\alpha}_{j,t}$ for queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$	{1, 2, . . . , 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, . . . , 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, . . . , 16}

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 MILP by 100 seconds. This setting results in an average integrality gap 0.01 % for instances with 6 time periods and 50 queues (see Tables 3 and 4 for more information). 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 process by Eqs. (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 3,000 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 Fig. 2 display the percentiles for the access time performance ratios $L^{A}_{0.9,j,t}$ in all queues in all instances. The curves show that resource capacities are allocated such that the performance ratios $L^{A}_{0.9,j,t}$ 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.

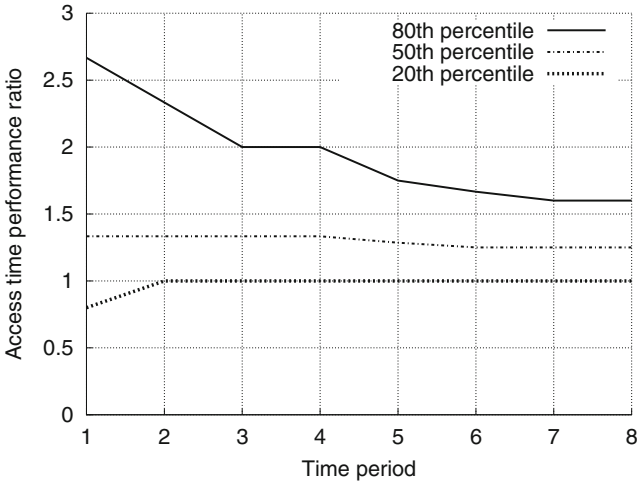

Fig. 2 The 20-th, 50-th and 80-th percentiles of the access time performance ratios $L^{A}_{0.9,j,t}$ for all queues in all instances

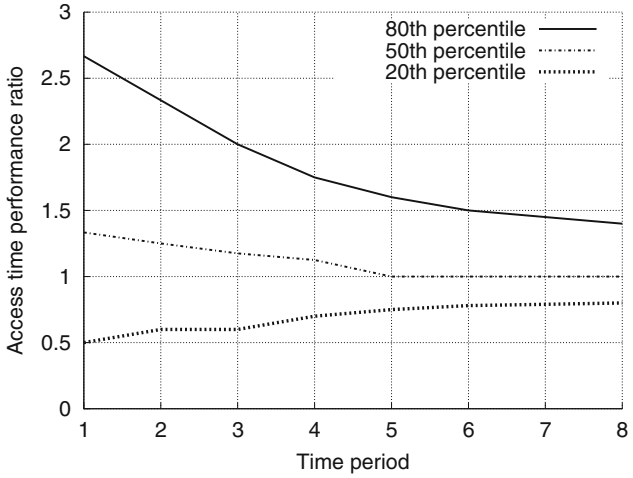

Fig. 3 The 20-th, 50-th and 80-th percentiles of the access time performance ratios $L^{A}_{0.9,j,t}$ for all queues in all instances, when $\kappa_r = 1.1$ for all $r \in \mathcal{R}$

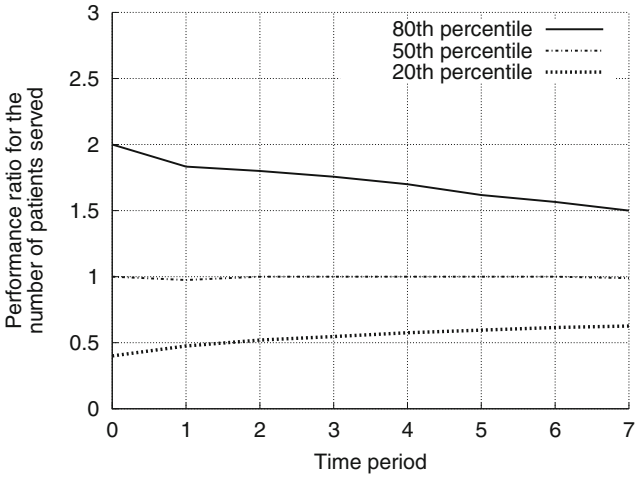

Fig. 5 The 20-th, 50-th and 80-th percentiles of the performance ratios $L^{C}_{j,t}$ for the number of patients served for all queues in all instances

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 Eq. (28) 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 Fig. 3 tend towards a lower number than the performance ratios in the graph in Fig. 2. In this case, they tend toward 1, which indicates that resource capacities are allocated such that our measures $A^{\alpha}_{j,t}$ for a higher fraction of queues are closer to target. The curves in Fig. 4 display the percentiles for the access time performance ratios $L^{H}_{0.9,g,t}$ for a complete care process, for all care processes in all instances. The method allocates resources such that the performance ratios $L^{H}_{0.9,g,t}$ tend towards 1. We may conclude that the measure $H^{\alpha}_{g,t}$ is closer to target for a larger fraction of care processes.

Serving the strategically agreed target number of patients The curves in Fig. 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.

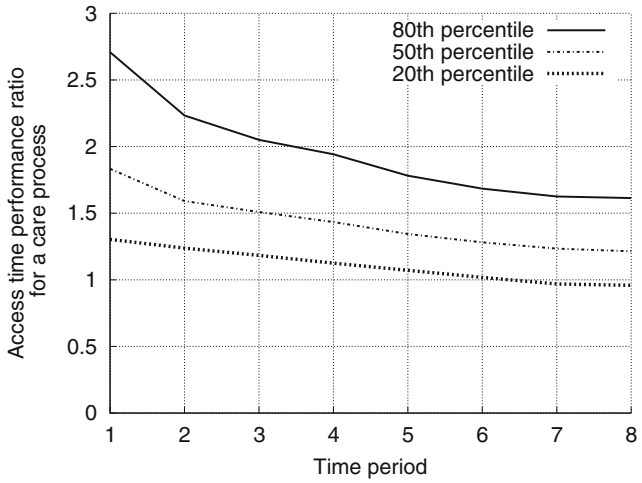

Fig. 4 The 20-th, 50-th and 80-th percentiles of the access time performance ratios $L^{H}_{0.9,g,t}$ for all care processes in all instances

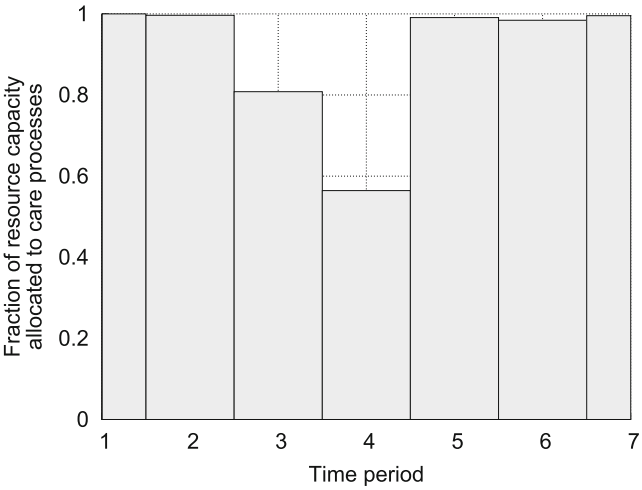

Fig. 6 Example of the fraction $\rho_{r,t}$ of resource capacities allocated to care processes for a resource type in an instance

Maximizing resource utilization and balancing workload The fraction $\rho_{r,t}$ of resource capacities $r \in \mathcal{R}$ that are allocated to care processes 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 Fig. 6 shows a decline in the percentage of resource capacity that is allocated to care processes 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 to care processes 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 Fig. 6. This can for example be caused by varying demand in different time periods (particular demand patterns), or by allocation decisions for resources in preceding stages and previous time periods. To improve workload balance for specific resources in the planning period, resource allocation constraints may be introduced in the MILP, as discussed in Section 3.4.

The average calculation time for relatively large instances ($G = 10$, $e_g = 5, \forall g \in \mathcal{G}$, $T = 6$, $R = 2$) is 4 minutes, which may be assumed to be reasonable for a tactical planning method. Furthermore, the average integrality gap for these instances is 0.01%. The calculation time is principally influenced by the number T of time periods and the number $|\mathcal{J}|$ of queues. Tables 3 and 4 give more details on average calculation time and average integrality gap for various instances.

Table 3 The average calculation time in seconds for various instances

Queues	Time periods		
	4	6	8
30	43	69	111
50	82	224	1,482
70	134	1,075	3,741

Table 4 The average integrality gap for various instances

Queues	Time periods		
	4	6	8
30	0.00 %	0.03 %	0.03 %
50	0.00 %	0.01 %	0.05 %
70	0.01 %	0.02 %	0.07 %

6 Managerial implications

We collaborate with various hospitals which increasingly implement procedures for tactical resource capacity planning *to achieve equitable access for patient groups*, *to serve the strategically agreed target number of patients*, *to maximize resource utilization* and *to balance the workload*. In their tactical planning approaches, some of these hospitals have spreadsheet solutions in place to evaluate for example waiting lists, access time and resource utilization. They use this information for resource allocation decision making, for example to allocate operating time and consultation time. Our method provides an optimization procedure for this step.

The tactical resource and admission plan proposed by our model is implemented using a rolling horizon approach: only near-term decisions are implemented. Every time period, the model is used to evaluate the tactical plan and to set the near-term decisions. The weight determination procedure is performed each time the tactical plan is reevaluated or redeveloped, as the weights are dependent on for example the expected patient arrivals and the selected tactical resource and admission plan. It is out of the scope of this paper to develop the operational decision rules that address unanticipated events during the execution of a tactical plan, such as a lower or higher demand than forecasted.

Implementation of our method at a particular hospital requires insight in the hospital's performance. Waiting list data, access times, the number of served patients, and expected resource availability should be made available every time period, to be able to propose a tactical plan. Also, the care processes in scope need to be defined (patient groups, the various stages, resource requirements and transition probabilities). The care stages in each care process are defined in close cooperation with medical staff, and by analyzing patient data obtained from the hospital information system. With the information about the care processes and individual patient procedures, methods described in for example [27] can be used to develop the transition probabilities for each stage in the care process. Correct administration of for example patient procedures, the sequence of these procedures, access times, the number of served patients is key in developing the patient care processes and providing the information to develop credible tactical plans.

Introduction of a dynamic tactical planning concept requires flexibility from all involved resources. It requires tactical rules (e.g., how many time periods before implementation is a tactical plan 'cast in stone'?), operational rules (e.g., when are resource capacities reallocated to other care processes?), and organizational changes in the various medical departments to be able to respond to changes

in the tactical plan effectively. One particular tactical rule was a prerequisite for participation of the involved medical departments and the successful implementation of dynamic tactical planning in one of the hospitals. The involved decision makers agreed that a decided reduction of allocated resource capacity (in this case operating time) can always be revoked when the resource capacity is required again in the future. Under this agreement, the involved decision makers can be more open for adjustments of the tactical plan, as they are certain that they can always go back to the prior tactical plan. Also, to support the process of tactical planning, agreements are required between the involved decision makers on what should be done (e.g., data analysis, calculating scenarios, discussing proposed plans) and who is involved (e.g., hospital managers, doctors, nurses) in each step of developing a tactical plan.

7 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 processes and determine the number of patients to serve at a particular stage of their care process.

Computational results show that our method improves compliance with access time targets, care process 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 processes. Care processes 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. For example, the model can be useful for tactical planning in a production environment. In such an environment, various products (care processes) are typically produced by multiple resource types. The product goes through different production stages (care stages) and at each stage there is 'work in progress' waiting to be processed (waiting list). The objectives in production may be to use resources effectively, to meet production targets and to have a certain amount of work in progress. These aspects are reflected in the objective function and constraints of our model. Clearly, 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 processes over time, and the incorporation of already scheduled patients. These extensions are interesting topics for further research.

Acknowledgements 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. Adan I, Bekkers J, Dellaert N, Vissers J, Yu X (2009) Patient mix optimisation and stochastic resource requirements: a case study in cardiothoracic surgery planning. Health Care Manag Sci 12(2):129–141
2. Anthony RN (1965) Planning and control systems: a framework for analysis. Harvard Business School Division of Research, Boston
3. Beliën J, Demeulemeester E (2007) Building cyclic master surgery schedules with leveled resulting bed occupancy. Eur J Oper Res 176(2):1185–1204
4. Bertsimas D, Niño-Mora J (2000) Restless bandits, linear programming relaxations, and a primal-dual index heuristic. Oper Res 48(1):80–90
5. Blake JT, Carter MW (1997) Surgical process scheduling: a structured review. J Soc Health Syst 5(3):17–30
6. Brailsford S, Vissers J (2011) OR in healthcare: a European perspective. Eur J Oper Res 212(2):223–234
7. Brucker P, Knust S (2012) Resource-constrained project scheduling In: Complex scheduling. GOR-Publications. Springer, Berlin
8. Brucker P, Drexl A, Möhring R, Neumann K, Pesch E (1999) Resource-constrained project scheduling: notation, classification, models, and methods. Eur J Oper Res 112(1):3–41
9. Butler TW, Karwan KR, Sweigart JR (1992) Multi-level strategic evaluation of hospital plans and decisions. J Oper Res Soc 43(7):665–675
10. Cardoen B, Demeulemeester E (2008) Capacity of clinical pathways—a strategic multi-level evaluation tool. J Med Syst 32(6):443–452

11. Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. Prod Oper Manag 12(4):519–549
12. Cerdá E, Pablos L, Rodriguez M (2006) Waiting lists for surgery. In: Hall RW (ed) Patient flow: reducing delay in healthcare delivery. International series in operations research & management science, vol 91. Springer, Berlin
13. Côté MJ (1999) Patient flow and resource utilization in an outpatient clinic. Socio-Econ Plann Sci 33(3):231–245
14. Denton BT, Miller AJ, Balasubramanian HJ, Huschka TR (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. Oper Res 58(4-Part-1):802–816
15. Elkhuizen SG, Das SF, Bakker PJM, Hontelez JAM (2007) Using computer simulation to reduce access time for outpatient departments. Br Med J 16(5):382
16. Every NR, Hochman J, Becker R, Kopecky S, Cannon CP (2000) Critical pathways: a review. Am Heart Assoc 101(4):461–465
17. Garg L, McClean S, Meenan B, Millard P (2010) A non-homogeneous discrete time Markov model for admission scheduling and resource planning in a cost or capacity constrained healthcare system. Health Care Manag Sci 13(2):155–169
18. Gemmel P, Van Dierdonck R (1999) Admission scheduling in acute care hospitals: does the practice fit with the theory. Int J Oper Prod Manag 19:863–878
19. Graves SC (1986) A tactical planning model for a job shop. Oper Res 34(4):522–533
20. Green LV, Savin S, Wang B (2006) Managing patient service in a diagnostic medical facility. Oper Res 54(1):11–25
21. Hall RW (2006) Patient flow: reducing delay in healthcare delivery. Springer, Berlin
22. Hans EW (2001) Resource loading by branch-and-price techniques. PhD thesis, University of Twente, The Netherlands
23. Hans EW, Van Houdenhoven M, Hulshof PJH (2012) A framework for healthcare planning and control. In: Hall RW (ed) Handbook of healthcare system scheduling. International series in operations research & management science, vol 168. Springer, Berlin
24. Hershey JC, Weiss EN, Cohen MA (1981) A stochastic service network model with application to hospital facilities. Oper Res 29(1):1–22
25. Hsu VN, de Matta R, Lee CY (2003) Scheduling patients in an ambulatory surgical center. Nav Res Logist 50(3):218–238
26. Jackson JR (2004)) Jobshop-like queueing systems. Manage Sci 50(12):1796–1802
27. Kapadia AS, Vineberg SE, Rossi CD (1985) Predicting course of treatment in a rehabilitation hospital: a Markovian model. Comput Oper Res 12(5):459–469
28. Nunes LGN, de Carvalho SV, Rodrigues RCM (2009) Markov decision process applied to the control of hospital elective admissions. Artif Intell Med 47(2):159–171
29. Organisation of Economic Co-operation and Development (OECD) (2012) Data retrieved May 10 2012, from: http://www.oecd.org/health
30. Pham DN, Klinkert A (2008) Surgical case scheduling as a generalized job shop scheduling problem. Eur J Oper Res 185(3):1011–1025
31. Porter ME, Teisberg EO (2007) How physicians can change the future of health care. J Am Med Assoc 297(10):1103
32. Samuel C, Gonapa K, Chaudhary PK, Mishra A (2010) Supply chain dynamics in healthcare services. Int J Health Care Qual Assur 23(7):631–642
33. van Oostrum JM, Van Houdenhoven M, Hurink JL, Hans EW, Wullink G, Kazemier G (2008) A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectrum 30(2):355–374
34. Vermeulen IB, Bohte SM, Elkhuizen SG, Lameris H, Bakker PJM, Poutré HL (2009) Adaptive resource allocation for efficient patient scheduling. Artif Intell Med 46(1):67–80
35. Weiss EN, Cohen MA, Hershey JC (1982) An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital patient flow. Oper Res 30(6):1082–1104
36. Wullink G (2005) Resource loading under uncertainty. PhD thesis, University of Twente, The Netherlands
37. Yeung RYT, Leung GM, McGhee SM, Johnston JM (2004) Waiting time and doctor shopping in a mixed medical economy. Health Econ 13(11):1137–1144
38. Zijm WHM (2000) Towards intelligent manufacturing planning and control systems. OR Spectrum 22(3):313–345