Optimizing the Strategic Patient Mix

Peter T. Vanberkel[1,2] Richard J. Boucherie[1], Erwin W. Hans[1], Johann L. Hurink[1]

[1]Centre for Healthcare Operations Improvement and Research,
University of Twente, Enschede, The Netherlands

[2]Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital,
Amsterdam, The Netherlands

Abstract

In this paper we address the decision of choosing a patient mix for a hospital that leads to the most beneficial treatment case mix. We illustrate how capacity, case mix and patient mix decisions are interrelated and how understanding this complex relationship is crucial for achieving the maximum benefit from the fee-for-service financing system. Although studies to determine the case mix that is of maximum benefit exist in the literature, the hospital actions necessary to realize this case mix has seen less attention. We model the hospital as an $M/G/\infty$ queueing system to evaluate the impact of accepting certain patient types. Using this queueing model to generate the parameters, an optimization problem is formulated. We propose two methods for solving the optimization problem. The first is exact but requires an integer linear programming solver whereas the second is an approximation relying only on dynamic programming. The model is applied in the department of surgery at a Dutch hospital. The model determines which patient types result in the desired growth in the preferred surgical treatment areas. The case study highlights the impact of striving for a certain case mix without providing a sufficiently balanced supply of resources. In the case study we show how the desired case mix can be better achieved by investing in certain capacity.

Keywords: Diagnosis Related Groups, Queueing Theory, Combinatorial Optimization

1 Introduction

In recent years hospital financing has changed from a budget oriented (lump sum) system to a fee-for-service system [4]. This transformation is intended to enhance accountability and to motivate hospitals to become more efficient. Diagnosis Related Groups (DRG), a concept which makes health care services a commodity, is facilitating this change. A DRG describes the whole

✉ Peter T. Vanberkel, University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands, p.t.vanberkel@utwente.nl, Tel: +31 53 489 3603, Fax: +31 53 489 2159

spectrum of activities involved in treating a certain disease or condition. The reimbursement to the hospital for each DRG treatment is fixed, meaning hospitals that provide the treatment for lower costs can realize greater profits, hence motivating efficiency. Variants of DRGs were introduced to achieve the same hospital financing transformation in many countries [19].

As a consequence of fee-for-service financing, hospitals are looking at their DRG case mix and are evaluating which services should be expanded and which should be discontinued [9,21]. This is evident in papers [10, 18] that focus on determining which DRGs are of the most benefit to the hospital. How to realize this DRG case mix has so far not been addressed in the literature. Furthermore, empirical research indicates that hospitals struggle to make choices that lead to desired DRG case mixes [1, 2].

To achieve a desired DRG case mix, hospitals must entice certain patients to the hospital. Patients are usually referred to the hospital by a general practitioner (GP) who evaluates a patient's symptoms and decides whether the patient should see a specialist. A referral from a GP does not specify which DRG treatment is required but rather the symptoms and the most appropriate modality. The patient then meets a specialist who decides on a treatment plan and assigns DRGs accordingly. The assigned DRGs may or may not be the ones of greatest benefit, however it is atypical to turn patients away at this point.

Through advertising and promotion to GPs, hospitals can encourage patients with certain diseases or symptoms to come to their hospital for treatment. However, knowing which symptoms will lead to the desired DRG case mix is not immediately obvious. Arrivals of patients (characterized by their symptoms) follow a stochastic process, and the required treatment can not be predicted with certainty. Determining, on the basis of symptoms, which types of patients (patient mix) to entice to the hospital in order to achieve the desired DRG case mix is the focus on this paper.

An as example, consider the treatment of Colorectal Cancer. A patient suspected of having Colorectal Cancer is referred to a hospital for further testing. The results from the testing could lead to surgery, chemotherapy, radiotherapy or palliative care for malignant cases, further screening for benign cases, a referral to a different oncology modality or even no further treatment. Within each of the treatment scenarios, there are several treatment options (i.e. DRGs) of which some are more desirable than others. Patient types in this example can be defined in many ways, but common factors indicating the prevalence of Colorectal Cancer include: personal or family history of polyps, Colorectal Cancer and/or Bowel Disease, ethnic background, diet, weight, alcohol use or smoking. Patient types can be further defined by symptoms such as constipation, diarrhea, blood in stool or jaundice. Patient types have uncertain arrival rates and with some probability require specific treatments. Thus choosing the best patient types to achieve the hospital's desired DRG case mix is not immediately obvious.

Hospitals are also constrained by their capacity levels which presumably relate to their desired DRG case mix. When capacity is overwhelmed the number of patients in the system increases, resources become more highly utilized and patient access times become worse. In this paper, to account for quality degradation due to demand exceeding capacity, we limit the fraction of time

when demand is allowed to exceed capacity.

In this paper, we determine which patient types to include, such that the maximum benefit from the DRG case mix is achieved. Once added, patient types cannot be removed in future periods, as allowing such an "on-again, off-again policy" would create undesired confusion about the offerings of the hospital. In this way, our problem has similar properties as the Project Sequencing Problem (PSP). As discussed in the following sections, the PSP determines which capacity expansion projects to implement in order to fulfill a growing demand for capacity. In this paper we model the hospital as an $M/G/\infty$ queueing system and formulate an ILP to exactly solve our problem. Using results from PSP literature we also formulate an approximate solution.

Statement of contribution: We develop a mathematical model to determine the policy for accepting new patient types that best matches the desired DRG case mix. To our knowledge it is the first paper where capacity, DRG case mix and patient mix decisions are accommodated in a single model facilitating joint decision making.

The paper is organized as follows, Section 2 formally defines and specifies the optimization problem and the queueing model. Section 3 introduces the PSP and illustrates how it can be used to approximately solve our problem. In Section 4 a case study is solved and the approximation evaluated. Throughout the paper the terms DRG and treatment are used interchangeably.

2 Model Description

The problem addressed in this paper is as follows: given that a hospital desires a certain DRG case mix, which patient types should be accepted (and when) to achieve this case mix while ensuring capacity restrictions are accounted for. We assume the relative importance of the DRGs are known and the capacity of the hospital to provide treatments is known for a finite time into the future. After a patient type is accepted, the number of arrivals of that patient type is modeled as a stochastic process. Upon arrival, a patient of a given patient type receives treatments according to some given probability distribution. Our model treats time as continuous and considers a finite planning horizon.

The formal problem description and a formalization as a combinatorial optimization problem are presented in Subsection 2.1. The calculation of some of the parameters of this combinatorial optimization problem is done using a queueing model which is described in Subsection 2.2.

2.1 Combinatorial Optimization Problem

Consider a set of patient types $P = \{1, 2, \ldots, N\}$ and a set of possible treatments $T = \{1, 2, \ldots, M\}$. A patient of type $n \in P$ has a probability $p_{n,m}$ of requiring treatment $m \in T$. The duration of treatment m has cumulative distribution $B_m(\cdot)$ with mean $\mathbb{E}[B_m]$. Let the number

of arrivals of patient type n in period $[0, t)$ be specified by a given random variable $\Lambda_n(t)$ and let $G_m(t)$ be a given model input which describes the volume of ongoing treatments m for which the hospital has capacity for at time t. For modelling the problem, we introduce variables $S_m(t)$ and $D_m(t)$ where $S_m(t)$ is the distribution for the number of patients receiving treatment m at time t and $D_m(t)$ is the distribution for the number of completed treatments m at time t. Note that $S_m(t)$ and $D_m(t)$ result from the choice of patient types to be accepted. The desired DRG case mix is reflected by values w_m, which specify the relative importance (or value) of treatment m.

The problem now is to indicate for each patient type the first moment in time t_n that patient type n is accepted. Note that for all times after t_n patient type n must also be accepted. Then the goal is to determine t_n such that the weighted number of treatments (weighted according to w_m) is maximized while ensuring that the number of treatments does not exceed $G_m(t)$ for more than a fraction φ_m of time. In other words, a hospital with capacity $G_m(t)$ wishes to maximize the weighted number of treatments they perform, whereby it is acceptable to exceed their capacity for a certain fraction $(1-\varphi_m)$ of the time. The value $\varphi_m \in (0, 1)$ is an input parameter reflecting the hospital's risk aversion for operating over capacity. A high φ_m value means demand will exceed capacity frequently (causing, for example, backlogged demand) whereas a low φ_m value means demand will exceed capacity less frequently (causing, for example, under utilized resources).

Let $\gamma = (t_1, t_2, \ldots, t_N)$ be a vector of chosen times to accept patient types n and let $C_t(\gamma)$ be a reward function measuring the discounted weighted sum of completed treatments for decision γ. Discounting future costs by $e^{-\alpha t}$ (where $\alpha \in (0, 1)$ is the discount factor) to time 0 ensures that later costs are adequately taken into account. Finding the optimal γ leads to the following optimization problem,

$$\begin{aligned} &\text{maximize } \int_0^T C_t(\gamma) e^{-\alpha t} dt \\ &\text{subject to } \mathbb{P}(S_m(t) \geq G_m(t)) \leq \varphi_m \qquad \forall m,\ t \end{aligned} \tag{P1}$$

where

$$C_t(\gamma) \quad = \quad \sum_{m=1}^{M} \mathbb{E}[D_m(t)] w_m. \tag{2}$$

Reward function (2) rewards according to the number of treatments completed, and is motivated by the financing structure at the hospital under study. Other choices are possible and the choice can be determined by the underlying decision process. Obvious choices include 1) $C_t(\gamma) = \sum_{m=1}^{M} \mathbb{P}(S_m(t) \geq G_m(t))$ that rewards according to the fraction of patients that exceed capacity, 2) $C_t(\gamma) = \sum_{m=1}^{M} \max\{\mathbb{E}[S_m(t) - G_m(t)], 0\}$ that rewards according to the expected amount by which capacity is exceeded and 3) $C_t(\gamma) = \sum_{m=1}^{M} \mathbb{E}[S_m(t)]$ which rewards by the expected number of patients receiving treatment m (assuming $S_m(t)$ is appropriately constrained in relation to $G_m(t)$).

Capacity in our model ($G_m(t)$) is specified as the volume of ongoing treatments m that the hospital can accommodate at time t. This implies that the decision of how to allocate available resource *time* to treatment types, has already been made (again this follows from the hospital under study). For example, consider an MRI machine that is available for 2000 minutes per week for treatments A and B. Assume treatments A and B require 10 and 20 minutes respectively. One way to divide the 2000 MRI minutes is to allocate 1200 minutes to A and 800 minutes to B, leading to a volume of $G_A(t) = 1200/10 = 120$ A treatments and $G_B(t) = 800/20 = 40$ B treatments per week. Because there are numerous ways to distribute the 2000 minutes between the two treatment types, there are numerous possibilities for $G_A(t)$ and $G_B(t)$.

Whereas in our model, the decision of how much time to allocate to each treatment type is already made i.e. we have only a single value of $G_A(t)$ (and a single value of $G_B(t)$), in other settings, it may be desirable to have this resource allocation decision be part of the model. The presented model can be generalized to accomplish this by defining capacity by the available treatment time of a resource. In this case the demand for resources can be determined by multiplying the volume of ongoing treatments $S_m(t)$ by the resource time needed for treatment m.

To solve (P1), we have to specify how random variables $S_m(t)$ and $D_m(t)$ can be determined for a given vector of γ. The queueing model defined in Subsection 2.2 is used for this purpose.

2.2 Queueing Model

To calculate the number of patients in treatment ($S_m(t)$) and the number of patients completing treatment ($D_m(t)$), we model the hospital as an $M/G/\infty$ queueing system. Since the population is a large, and patients get ill independently of each other, it is natural to assume that they get ill according to a non stationary Poisson process $\tilde{\Lambda}_n(t)$. It follows that patients of type n that arrived to the hospital have a Poisson distribution $\Lambda_n(t) = \tilde{\Lambda}_n(t)F_n(t)q_n$ with mean $\lambda_n(t)$ where q_n is the fraction of the population choosing the considered hospital and where $F_n(t) = 1$ when patient type n is accepted at time t (i.e. $t_n \geq t$) and $F_n(t) = 0$ otherwise (without loss of generality we assume $q_n = 1$).

With probably $p_{n,m}$ [or $p_{n,m}(t)$] a patient of type n requires treatment m and, thus, the arrival process of patient type (n, m) is Poisson at rate $\lambda_{n,m}(t) = \lambda_n(t)p_{n,m}$. It can readily be seen that the number of patients of type n requiring treatment m at time t, $S_{n,m,t_n}(t)$, is distributed as the number of customers in a non stationary $M/G/\infty$ queue with arrival rate $\lambda_{n,m}(t)$ and cumulative service time distribution $B_m(\cdot)$ and is given by (see [12]),

$$\mathbb{P}(S_{n,m,t_n}(t) = s) = \frac{e^{-\mathbb{E}[S_{n,m,t_n}(t)]}(\mathbb{E}[S_{n,m,t_n}(t)])^s}{s!} \tag{3}$$

where

$$\mathbb{E}[S_{n,m,t_n}(t)] = \mathbb{E}[\lambda_{n,m}(t - B_{e,m})]\mathbb{E}[B_m] \tag{4}$$

and $B_{e,m}$ is the excess service time with the following cumulative distribution function,

$$\mathbb{P}(B_{e,m} \leq t) = \frac{1}{\mathbb{E}[B_m]} \int_0^t (1 - B_m(w))\, du. \tag{5}$$

Observe that $S_m(t) = \sum_{n=1}^{N} S_{n,m,t_n}(t)$ is the sum of Poisson random variables and therefore is Poisson distributed with rate parameter $\mathbb{E}[S_m(t)] = \sum_{n=1}^{N} \mathbb{E}[S_{n,m,t_n}(t)]$. Like $S_m(t)$, $D_m(t)$ is also Poisson distributed in a non stationary $M/G/\infty$ queue. In particular, the number of completed treatments m of patient type n at time t is Poisson with mean $\mathbb{E}[D_{n,m,t_n}(t)] = \mathbb{E}[\lambda_{n,m}(t - B_m)]$ and $\mathbb{E}[D_m(t)] = \sum_{n=1}^{N} \mathbb{E}[D_{n,m,t_n}(t)]$.

2.3 Model Characteristics

In this subsection, some observations about infinite server queues and (P1) are explained. These observation are exploited later when solving the problem. First, entities (patients in our case) are independent of each other in an infinite server queue. This means, for example, that the amount of time in the system is not impacted by other patients or the order in which they arrive. It follows then that the total reward for accepting a new patient type can already be computed at its acceptance time t_n, since these patients are not influenced by other patients, and therefore not influenced by future decisions. It follows that the reward for adding patient type n at time t_n is,

$$c_{n,t_n} = \sum_{m=1}^{M} \left(\int_{t_n}^{T} \left(\mathbb{E}[D_{n,m,t_n}(t)] e^{-\alpha t} \right) dt \right) w_m \tag{6}$$

and it follows that the objective function from (P1) can be rewritten as,

$$\int_0^T C_t(\gamma) e^{-\alpha t} dt = \sum_{n=1}^{N} c_{n,t_n} e^{-\alpha(t_n)}. \tag{7}$$

The second observation about (P1) is that the constraints are expressions for the φ_m percentiles of distributions $S_m(t)$ which, when $S_m(t)$ is known, can be evaluated exactly. Let $\mathbb{Q}[x]$ represent the φ_m percentile of distribution x, then

$$\mathbb{Q}[S_m(t)] = \sup\{s : \mathbb{P}(S_m(t) \geq s) \leq \varphi_m\} \tag{8}$$

$$= \sup\left\{ s : \sum_{k=1}^{s} \frac{\mathbb{E}[S_m(t)]^k e^{-\mathbb{E}[S_m(t)]}}{k!} \leq \varphi_m \right\} \tag{9}$$

Using this observation the constraints from (P1) can be rewritten as,

$$\text{subject to } \mathbb{Q}[S_m(t)] = \mathbb{Q}\left[\sum_{n=1}^{N} S_{n,m,t_n}(t) \right] \leq G_m(t) \qquad \forall m,\ t. \tag{10}$$

2.4 Discrete Time Formulation

In this subsection we formulate a discrete time version of (P1) which eliminates the need to evaluate the integral in its objective function. In the discrete time version of our problem the continuous time discount equation $e^{-\alpha t}$ is replaced by the equivalent discrete time discount equation $(1/(1+\alpha))^t$. Let $x_{n,t}$ be binary decision variables reflecting the first moment in time when patient type n is accepted. For example, when $x_{3,5} = 1$ this means $t_3 = 5$. This leads to the following integer linear programming (ILP) problem,

$$\text{maximize} \sum_{n=1}^{N} \sum_{t=0}^{T} c_{n,t_n} x_{n,t} \left(\frac{1}{1+\alpha} \right)^t \tag{ILP}$$

subject to

$$\sum_{t=1}^{T} x_{n,t} \leq 1 \quad \forall n \tag{12}$$

$$\sum_{\tau=0}^{t} \mathbb{Q} \left[\sum_{n=1}^{N} S_{n,m,t-\tau}(t) \right] x_{n,t-\tau} \leq G_m(t) \quad \forall t,\ m \tag{13}$$

$$x_{n,t} = 0 \text{ or } 1 \quad \forall n,\ t.$$

Constraints (12) ensure that each patient type is accepted only once, while constraints (13) ensure that the φ_m percentile of demand for treatment m does not exceed $G_m(t)$. For moderate sized instances the resulting ILP can be solved by commercial ILP solvers like e.g. ILOG CPLEX Solver.

3 Approximate Solution Approach

In this section we introduce an approximate solution approach (ASA) to our problem which does not rely on an ILP solver. The approximation exploits the structure of our problem which is similar to the well studied Project Sequencing Problem (PSP). The PSP can be solved directly with dynamic programming [3]. As explained in this section, although our problem is similar to a PSP, it is too general to be solved directly with dynamic programming. To overcome this, we relax one of the elements of our problem. The relaxation amounts to ignoring the time required for the newly accepted patient type to populate the system i.e. we ignore the "startup" time. Formally this is achieved by changing the definition of $S_m(t)$ (See Subsection 3.2). Furthermore we ignore one of the conditions of the dynamic program necessary to obtain an optimal solution. We argue that although the solution is not guaranteed to be optimal, within the range of typical problems, ignoring this assumption has little impact on the solution quality (as illustrated with numeric examples in Section 4).

The steps in the approximation are as follows: We first relax our problem (by defining $S_m(t)$ differently) so that it is a PSP. We then apply the dynamic program of [3] to determine the

sequence in which the patient types should be accepted, i.e. the order not the time. Once this sequence is known we use the original definition of $S_m(t)$ to iteratively determine the best time to accept each patient type for the given sequence.

Using the ASA to solve (P1) has three distinct advantages over the ILP approach. The first being that the ASA relies on dynamic programming and does not require a costly ILP solver. The second is that the ASA separates the sequencing decision from the timing decision and as such reduces the complexity. Hence large problem instances that are intractable for the ILP can be approximated using the ASA. The third advantage is that the PSP is a commonly studied problem and as such, there are extensions to the dynamic programming approach to solve even larger problem instances [8]. By formally defining the relationship between our problem and a PSP, we are able to leverage existing PSP literature to approximately solve very large problem instances.

This section is organized as follows; in Subsection 3.1 we introduce the PSP and discuss the dynamic programming solution approach. In Subsection 3.2 we formally introduce the relaxation of our problem needed for this approach. In Subsection 3.3 we formally define the ASA to our problem.

3.1 Project Sequencing Problem Review

There is a rich literature on Capacity Expansion problems, as indicated by a number of reviews [11,15] and books [5,7] spanning multiple decades. Broadly put, the literature concerns choosing the timing, size and type of capacity expansion needed to fulfill growing demand. Typically, the objective is to minimize the discounted cost of meeting demand and in doing so, finds the optimal balance between expanding in large increments to achieve economies of scale, verses small increments which reduce opportunity (or excess capacity) costs. The literature addresses many variants of the problem, including finite / infinite planning horizons, linearly /non-linearly growing demand, deterministic / stochastic demand, and continuous / discrete expansion sizes. When the choice of expansion projects is limited to a finite set of projects, each with a specified expansion size, the problem is called a PSP. In this subsection we introduce and review the PSP.

The PSP problem [3,11,17], assumes a finite set of projects $R = \{1, 2, ..., Y\}$ where each project $y \in R$ has an implementation cost v_y and capacity z_y. The objective is to determine the implementation times (t_y) for each project which minimizes the discounted cost while fulfilling a pattern of deterministic demand $s(t)$. Formally the problem can be stated as follows,

$$\begin{aligned} &\text{minimize } \sum_{y \in R} v_y e^{-\alpha t_y} && (14) \\ &\text{subject to } \sum_{y \in R} z_y \delta_y(t) \geq s(t) \qquad \forall\ t \end{aligned}$$

where $\delta_y(t) = 1$ if $t \geq t_y$ (and $\delta_y(t) = 0$ otherwise).

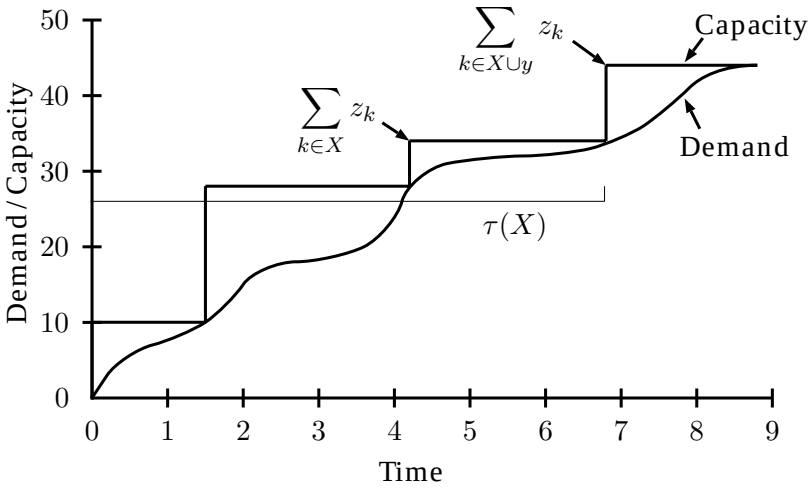

Figure 1: The capacity expansion process

To solve (14), Erlenkotter [3] proposes a backward dynamic programming model to determine the optimal implementation sequence. From this sequence, the optimal implementation time for each project can be derived. Let X be the state variable where X describes a subset of R representing the set of implemented projects. Let $\tau(X)$ be the latest time when the capacity available from the implemented projects is greater than the demand, i.e. $\tau(X) = \max\{t | \sum_{y\in X} z_y \geq s(t)\}$.

When minimizing excess capacity costs (and when α is positive), it is optimal to add additional capacity only when existing capacity is exhausted [3, 13, 14, 17]. Figure 1 displays the capacity expansion process for a one dimensional problem. The following backward dynamic program determines the optimal sequence for implementing the Y projects,

$$\begin{aligned} f(R) &= 0 \\ f(X) &= \min_{y\notin X} \left(v_y e^{-\alpha\tau(X)} + f(X \cup y) \right), \quad \forall X \subset R. \end{aligned} \tag{15}$$

A network representation of (15) can be constructed. Figure 2 displays an example with $Y = 3$ projects. Finding the shortest path from node $\emptyset$ to $\{1, 2, 3\}$ also gives the optimal sequence in which the projects should be implemented. Once this sequence is known the time to implement each project can be determined iteratively using $\tau(X)$. When demand is linear, the optimal policy can be determined without dynamic programming as described in [3]. Other solution approaches involve integer programming techniques which assume a discrete time scale [16, 20] and "select jointly the choice of expansion and its timing" [3].

Under the following five assumptions, the Erlenkotter dynamic programming approach finds the optimal solution:

1) *The full capacity is available for use instantaneously upon completion of a project.*

2) *Project capacity once created has infinite life and does not change over time.*

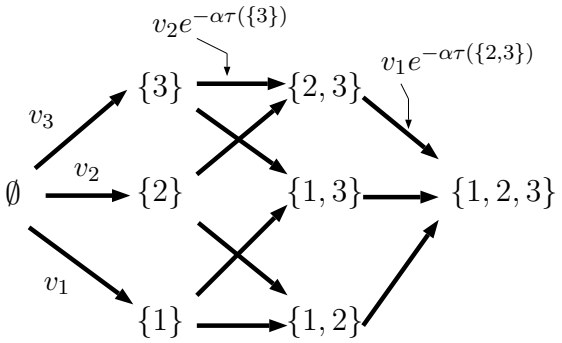

Figure 2: A network representation of the PSP

3) a) *The investment cost is incurred at the time the project is completed.* b) *The investment cost does not vary with time.*

4) *Variable operating and distribution costs are proportional to the amount actually produced and identical for all projects.*

5) *Demand must be supplied from current production*

3.2 Analogy between our problem and the PSP

By considering each patient type n as a project, our problem is similar to a PSP. Note however, that we are maximizing the reward, not minimizing capacity investments and that we restrict this reward by available capacity $G_m(t)$ which is independent of the decision variables. In our problem $S_m(t)$ depends on the decision variables, thus, when comparing our problem to the PSP, $\mathbb{Q}[S_m(t)]$ is akin to "capacity" and $G_m(t)$ is akin to "demand".

As discussed below, we can not use Erlenkotter's dynamic program directly to solve our problem because it violates assumption 1. To overcome this we modify (relax) our problem. Assumption 2 is also violated, however as discussed below, the impacted of violating this assumption is minimal. The remaining 3 assumptions are satisfied in our problem.

1) *The full capacity is available for use instantaneously upon completion of a project.*

This assumption is needed so that the total capacity of a given state can be determined from the state description. It follows that since the state indicates only the implemented projects (and not *when* each project was implemented) that the capacity of a project be independent on how much time has passed since it was implemented. In other words the capacity must be determined independently of the time of implementation. Using the definition of $S_{n,m,t_n}(t)$ from Subsection 2.2, this is not possible because $\mathbb{E}[S_{n,m,t_n}(t)]$ depends on t_n, see (4).

To overcome this and allow us to use dynamic programming we relax our definition of $S_{n,m,t_n}(t)$. In our relaxation, we model the number of patients in the system with a steady state M/G/∞

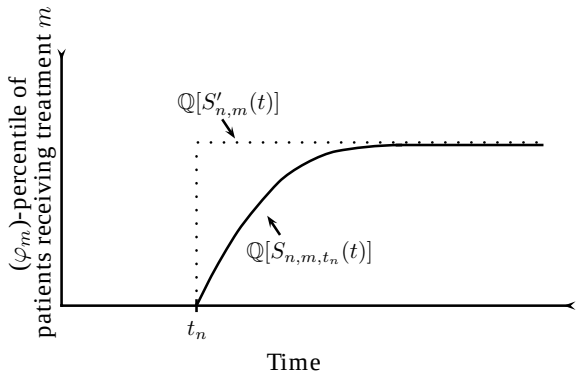

Figure 3: Illustration of $\mathbb{Q}[S_{n,m,t_n}(t)]$ and $\mathbb{Q}[S'_{n,m}(t)]$

queue [6]. In such a queue, the number of patients in the system ($S'_{n,m}(t)$) is Poisson distributed with mean $\mathbb{E}[S'_{n,m}(t)] \approx \lambda_n(t)p_{n,m}\mathbb{E}[B_m]$. Notice that the definition of $\mathbb{E}[S'_{n,m}(t)]$ is independent of t_n. Formally the relaxation that we use in the ASA is as follows,

$$\mathbb{E}[S_{n,m,t_n}(t)] \approx \mathbb{E}[S'_{n,m}(t)] \approx \lambda_n(t)p_{n,m}\mathbb{E}[B_m]. \tag{16}$$

Obviously, in a similar manner $S'_m(t)$ denotes the volume of patient receiving treatment m at time t in this relaxed setting. $S'_m(t)$ is Poisson distributed with mean $\mathbb{E}[S'_m(t)] = \sum_{n=1}^{N} \mathbb{E}[S'_{n,m}(t)]$.

This modification is consistent with other application of the PSP where a project's capacity needs to be brought online gradually. Furthermore, when the time between accepting new patient types is much longer than the mean service time, one would expect that $\mathbb{E}[S_{n,m,t_n}(t)]$ will have converged before the next patient type is accepted and thus this assumption does not impact the solution. In Figure 3 a sketch of the difference between $\mathbb{Q}[S_{n,m,t_n}(t)]$ and $\mathbb{Q}[S'_{n,m}(t)]$ is given.

2) *Project capacity once created has infinite life and does not change over time.*

Since $\mathbb{Q}[S'_{n,m}(t)]$ depends on $\lambda_n(t)$ which is non stationary in time, this assumption is violated. However, given that the prevalence of a disease (i.e. the arrival rate for a population of patients) changes gradually over time, one expects that the impact of violating this assumption is minimal.

3a) *The investment cost is incurred at the time the project is completed.*

Since the patients are assumed to be independent, the reward can be accounted for at the moment in time when the patient type is accepted. Although the rewards will take place in the future, they are known at time t_n and can be discounted to time t_n.

3b) *The investment cost does not vary with time.*

This is true in our case since the reward per treatment (w_m) does not change over time.

4) *Variable operating and distribution cost are proportional to the amount actually produced and*

identical for all projects.

In our problem there are not any variable operating or distribution cost.

5) *Demand must be supplied from current production*

Our problem is equivalently constrained.

3.3 Dynamic Programming Formulation

In this subsection we first describe how the Erlenkotter approach is used to determine the optimal sequence to add patient types, given $S'_{n,m}(t)$. We then use the original definition of $S_{n,m,t_n}(t)$ to determine the times t_n and the reward associated with the sequence. These steps together constitute the ASA.

Sequence in which to add patient types

Let I be a subset of P indicating the patient types n that have already been accepted at a given moment in time. Our objective is essentially to maximize treatments, and since our discount rate is positive, it is optimal to accept additional patient types as soon as possible, i.e. at the first moment in time when a patient type can be added such that $\mathbb{Q}[S'_m(t)] \leq G_m(t)$ for all m. When the set of patient types I is already accepted and when patient type $r \notin I$ is to be added, then let this point in time be $\tau(I \cup \{r\})$ which is computed as follows,

$$\tau(I \cup \{r\}) = \inf\{t | \mathbb{Q}[S'_m(t)] \leq G_m(t),\ \forall m\}. \tag{17}$$

As explained in Subsection 2.1, the reward for accepting a patient type is the discounted weight sum of the completed treatments. For the dynamic program to work, we must compute this reward at the moment in time when the patient type is first accepted. When the set of patient types I are already accepted and when patient type r is to be added, then the total reward (discounted to time $\tau(I \cup \{r\})$) for adding patient type r is,

$$c_r = \sum_{m=1}^{M} \left(\int_{\tau(I \cup \{r\})}^{T} \mathbb{E}[D_{r,m,\tau(I \cup \{r\})}(t)] e^{-\alpha t} dt \right) w_m. \tag{18}$$

Using the following backward dynamic program, the sequence in which to add patient types can be determined,

$$\begin{aligned} f(P) &= 0 \\ f(I) &= \max_{r \notin I} \left(c_r e^{-\alpha \tau(I \cup \{r\})} + f(I \cup \{r\}) \right) \quad \forall i \subset P. \end{aligned} \tag{19}$$

Let $\vartheta = \{x_1, x_2, \ldots\}$ be this sequence, where x_i is the i^{th} patient type accepted. For example when $\vartheta = \{6, 2, 4, \ldots\}$ then patient type 6 should be accepted first, followed by type 2 and type 4, etc.

Implementation Times and Reward

Once ϑ is known it is possible to determine the implementation times and the overall reward associated with this sequence. One option to determine the implementation times and the overall reward, is to add the following constraints to the ILP formulation of Subsection 2.4,

$$t_{x_k} \leq t_{x_k+1}, \quad k = 1, 2, \ldots, N.$$

Of course this means the ASA requires an ILP solver and thus, this is not a desirable option. Alternatively, it is possible for a given sequence to compute the implementation times iteratively as follows: Start initially with $t_{x_1} = t_{x_2} = t_{x_3} = \ldots = \infty$ and then use (20) to sequentially calculate $t_{x_1}, t_{x_2}, \ldots$.

$$t_{x_z} = \inf\{t | \mathbb{Q}[S_m(t)] \leq G_m(t), \ \forall m\} \tag{20}$$

Note that in (20) we use $S_m(t)$ as defined in Subsection 2.2 and not the relaxation $S'_m(t)$.

If γ' denotes the set of implementation times associated with sequence ϑ, the reward for the sequence is,

$$r(\gamma') = \int_0^T C_t(\gamma') e^{-\alpha t} dt. \tag{21}$$

In summary we approximate the solution to our problem by relaxing the definition of $S_m(t)$ according to (16). In this way our problem has a similar structure as a PSP. Then we use the dynamic programming formulation of Erlenkotter (19) to determine the optimal sequence to add patient types in the relaxed problem. From this optimal sequence we use the original definition of $S_m(t)$ and equations (20) and (21) to determine the implementation times (γ') and the reward $r(\gamma')$. We call this approach the Approximate Solution Approach (ASA).

4 Application and Evaluation of ASA

We have introduced two approaches to solve (P1). The ILP approach (described in Subsection 2.4) is a discrete time approximation of (P1), although for simplicity and clarity in the text, we call the ILP solution the optimal solution. The second approach is the ASA which approximates the solution to (P1). In this section we solve a case study problem and use numeric problem instances to evaluate and compare the two solutions approaches.

To evaluate the ASA we introduce the problem addressed at the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital (NKI-AVL). Using this problem instance as an initial case we complete a sensitivity analysis on the parameters that influence the ASA in an effort to characterize the problem instances where the ASA gives similar results as the ILP approach.

4.1 Application

NKI-AVL is a comprehensive cancer center in Amsterdam, The Netherlands, which provides hospital care and research. The hospital has 150 inpatient beds and sees about 24,000 new patients every year. As with many Dutch hospitals, NKI-AVL is eager to expand and provide state-of-the-art treatments with state-of-the-art equipment. To finance such expansions, managers have identified which DRGs are of the most value and hence should be the focus of the expansion. To achieve growth in the desired DRGs, the hospital must decide which patient types are the best to attract to the hospital and when to do so. In this section we apply our model to support this decision for the surgical speciality.

Since typically each patient type has a limited number of treatment options, a hospital can determine which additional patient types to accept by solving a number of subproblems. For example, historical records for the hospital in this study indicate that there are 37 different patient types (defined by their cancer diagnosis) and 109 different treatments. However some patient types have 0% probability of needing certain treatments and as such, it is possible to divide this 37 by 109 dimension problem into a number of subproblems of much smaller dimensions. The subproblem solved in this section is called the "surgical specialty" subproblem as it encompasses all surgical treatments and their corresponding patient types resulting in six patient types and 34 treatment types. Seven of the 34 treatment types represent over 90% of the total volume of treatments, thus in the interest of brevity, we only include these seven in the analysis.

Management of the surgical specialty do not expect new patient types to be added to the hospital, rather that the volume of existing patient types will increase if/when they try to attract them. This is essentially the same problem whereby the increased arrival rate of a patient type resulting from some hospital action, represents a new patient type. Our goal is to determine the time when this "new patient type" should be accepted. Management also have forecasted estimates of the amount by which the volume of each patient type will increase, which corresponds to the arrival rates of the "new patient types". All data, other than these estimates, comes from the hospital's historic records. All the patient types discussed in the remainder of this section represent "new patient types".

The ILP was solved using ILOG CPLEX Solver 12.2 and the parameters (generated via the queueing model) were computed using Microsoft Visual Basic. The dynamic program used in the ASA was solved with Dijkstra's algorithm programmed in MatLAB. The parameters for the ASA were also computed using Microsoft Visual Basic.

4.1.1 Application Data

The fraction of patients of type n that will receive treatment type m is available in Table 1. Note that since it is possible for a patient to have multiple treatments, the sum of the probabilities across each row can be greater than 100%. In a similar manner, the sum can be less than 100%

Table 1: The fraction of patients of type n that receive treatment type m ($p_{n,m}$)

		Treatments (m)						
	Patient Types (n)	1	2	3	4	5	6	7
1	Lung Cancer	24.8%	0%	57.1%	0%	3.7%	0%	6.7%
2	Lung Cancer (Pleura)	21.2%	0%	53%	0%	7.7%	0%	4.9%
3	Breast Cancer	54.8%	35.9%	0%	7.4%	3.3%	2.8%	0%
4	Colon Cancer	23.6%	0%	46.2%	0%	9.2%	0%	11.5%
5	Colon Cancer (Sigmoid)	32.1%	0%	38.4%	0%	9%	0%	23%
6	Colon Cancer (Rectum)	4%	0%	56%	0%	0%	0%	16%

Table 2: The mean arrival rate (patients / month) for patient type n

	Years											
n	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
1	2.17	2.51	2.91	3.36	3.89	4.49	5.2	6.01	6.95	8.03	9.28	10.73
2	0.59	0.68	0.79	0.91	1.05	1.22	1.41	1.63	1.88	2.18	2.52	2.91
3	34.3	37.73	41.51	45.66	50.23	55.25	60.77	66.85	73.54	80.89	88.98	97.88
4	4.79	5.54	6.4	7.4	8.55	9.89	11.43	13.22	15.28	17.66	20.42	23.61
5	2.09	2.38	2.72	3.1	3.54	4.03	4.6	5.24	5.97	6.81	7.77	8.85
6	2.39	2.73	3.11	3.55	4.05	4.61	5.26	6	6.84	7.8	8.89	10.14
7	0.74	0.86	0.99	1.15	1.33	1.53	1.77	2.05	2.37	2.74	3.17	3.67

since only 7 of the 34 treatment types are included.

The relative importance of each DRG (w_m) relates to the renumeration amount that the hospital receives for each. The actual values are excluded for proprietary reasons, although they rank as follows, $w_7 > w_6 > w_4 > w_3 > w_2 > w_5 > w_1$. The forecasted arrival rate (patients / month) for each patient type n is shown in Table 2.

The mean service time ($\mathbb{E}[B_m]$) for each treatment type m is shown in Table 3. The distributions have been left out of the text for brevity. The capacity ($G_m(t)$) to provide treatment type m over the 11 year planning horizon is shown in Table 4.

For φ_m, several values ranging from 0.9 to 0.1 were considered in order to illustrate the sensitivity of this parameter. We discretized the problem into monthly periods, meaning a patient type can only be added at the beginning of the month.

Table 3: The mean service time ($\mathbb{E}[B_m]$) for each treatment type m

	Treatments (m)						
	1	2	3	4	5	6	7
$\mathbb{E}[B_m]$ (months)	16.3	11.8	16.7	3.6	7.7	3.0	5.0

Table 4: The capacity ($G_m(t)$) to treat patients of type m over the 11 year planning horizon

	Years											
m	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
1	2004	2204	2424	2546	2673	2807	2947	3094	3249	3411	3582	3761
2	1573	1730	1903	1998	2098	2203	2313	2429	2550	2678	2812	2952
3	409	450	495	519	545	573	601	631	663	696	731	767
4	266	292	322	338	355	372	391	411	431	453	475	499
5	210	231	254	267	280	294	309	325	341	358	376	395
6	89	98	108	113	119	125	131	138	145	152	159	167
7	79	87	96	101	106	111	117	122	129	135	142	149

Table 5: Initial solution to case study (given in months)

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP Solution	0	0	53	143	100	112	24,121

4.1.2 Results

Solving (P1) with the above data as input, the optimal reward and timing for accepting the patient types was determined and is displayed in Table 5. However after comparing the volume of patients receiving treatment with the available capacity, it was observed that for all treatment types (other than $m = 7$), the volume receiving treatment was much less that the capacity. It followed that treatment type $m = 7$ was the bottleneck resource and that it was significantly debilitating to the system. For example, $\mathbb{Q}[S_7(t)] > G_7(t)$ (for all t) whereas the opposite was true for the 6 other treatment types. To improve on this, we used the model to investigate how the reward changes as the capacity to provide treatment $m = 7$ is increased.

We increased the capacity of treatment type $m = 7$ by 10, 20, 30 and 40% and observed the corresponding reward. As illustrated in Figure 4, increasing the capacity in this way allows the reward to increase by more than a factor of 2. Also observable in Figure 4, is that increasing the capacity by more than 30% does not increase the reward, as a new bottleneck emerges.

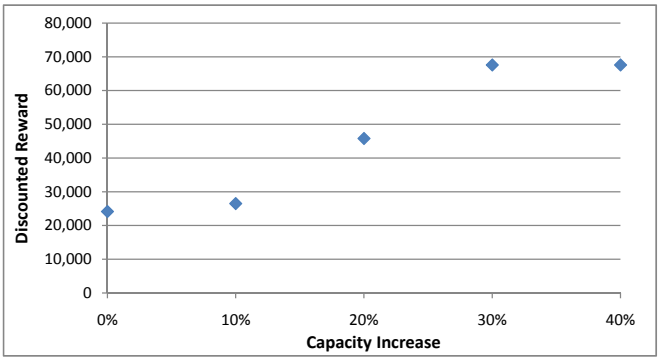

Figure 4: Reward associated with increasing capacity for treatment type $m = 7$

Table 6: Comparison of the ILP and ASA solutions to the Case Study problem instance

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP Solution	25	0	53	14	0	3	67,220
ASA Solution	0	0	53	53	0	11	56,529

Figure 4 demonstrates that if additional resources were made available for treatment type $m = 7$ then the hospital's reward would increase significantly. The decision is left with management whether the increase in the reward justifies the additional investment in capacity. Repeating this process of identifying the bottleneck and deciding whether additional expansions are warranted can help balance the supply and demand in the hospital. In the same vein the model can be used to determine the extent to which capacity can be decreased for certain treatment types.

After several iterations of the model, appropriate resource levels were determined and a final solution was found. The "ILP Solution" row of Table 6 lists the reward and time when each patient type should be accepted.

Using this information the hospital can develop a strategy for enticing patients to the hospital. In general we conclude that in the short term (0 to 15 months) the hospital should focus on attracting patient types $2, 5$ and 6. In the medium term (15 to 30 months) the hospital should focus on patient types 1 and 4. In the long term, patient type 3 should be the focus. That said, the model can be used to reevaluate this policy in later years, after the level of success in attracting new patients is known. This of course could result in a new strategy to replace the current long term strategy.

4.2 Evaluation of Approximate Solution Approach

The analysis of the preceding subsection were completed according to the ILP approach. In this subsection we use the same case study and compare solutions of the ASA with those from the ILP approach. Furthermore, we perform a sensitivity analysis on the main parameters that impact the approximation.

To evaluate the ASA, consider that all model parameters are the same and the only difference between the ASA and the ILP approach is how we model the volume of patient receiving treatment ($S'_m(t)$ for the ASA and $S_m(t)$ for the ILP approach). Thus to evaluate the ASA, we complete a sensitivity analysis on the parameters that impact this volume ($\lambda_n(t)$ and $\mathbb{E}[B_m]$) and also the capacity parameter ($G_m(t)$). For parameters $\lambda_n(t)$ and $G_m(t)$ we vary the rate of change from t to $t+1$ and for $\mathbb{E}[B_m]$ we simply vary its value. More than 20 problem instances were evaluated to draw conclusions on the validity of the ASA. We use the following four instances to illustrate these conclusions.

Problem Instance 1: This is the case study described above. Table 6 compare the solutions to the case study found by both approaches.

Table 7: Comparison of the ILP and ASA solutions to problem instance 2

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP Solution	12	0	126	30	0	5	52,616
ASA Solution	30	0	126	30	0	5	50,306

Table 8: Comparison of the ILP and ASA solutions to problem instance 3

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP Solution	37	0	79	34	16	26	20,792
ASA Solution	22	16	79	33	16	37	19,954

Problem Instance 2: In this problem instance we increase the rate by which function $\lambda_n(t)$ changes from one period to the next. We make the rate the same for all patient types such that $\lambda_n(t)/\lambda_n(t+12) = 15\%$ for all n, meaning each patient population is increasing by 15% per year. This problem reflects a situation where the prevalence of certain diseases is increasing rapidly. Results for this problem instance are given in Table 7.

Problem Instance 3: In this problem instance we decrease the rate by which function $G_m(t)$ changes from one period to the next. We make the rate the same for all treatment types such that $G_m(t)/G_m(t+12) = 5\%$ for all m. As such, this problem instance represents a hospital with modest growth ambitions. Results for this problem instance are given in Table 8.

Problem Instance 4: The mean time to complete a treatment ($\mathbb{E}[B_m]$) is set to 3.5 months in this example. Results for this problem instance are given in Table 9.

When comparing the performance of the ASA for the 4 problem instances above, it performed the worst in the case study. In all other instances, the ASA reward was within 6% of optimal (in the case study the ASA reward was within 16% of optimal). This leads to the conclusion that good solutions can be found with the ASA. In particular, this is true when $\mathbb{E}[B_m]$ is small and when the rates of change of $\lambda_n(t)$ and $G_m(t)$ are homogeneous with respect to patient and treatment types.

However, not only is the objective function value important, but also the structure of the solution indicating when patient types should be accepted. For our instances this structure is similar in both approaches. This is particularly true in the short term (i.e. the first two or three patient types which should be accepted). In this way the ASA can be used to solve the short term plan, and then at (or near) the end of the term, the problem can be reevaluated with up-to-date parameters. As such, in the absence of an advanced ILP solver, the ASA approach can be used as the kernel of a control policy.

Table 9: Comparison of the ILP and ASA solutions to problem instance 4

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP Solution	18	0	55	26	0	16	62,704
ASA Solution	55	0	55	24	0	16	60,225

Furthermore, like the ILP approach, the ASA can be used to identify the bottleneck resource. For example, analyzing the case study with the ASA leads to the same conclusions about the bottleneck resource. The increase in the reward that results from adding capacity to treatment type $m = 7$ is significant (and much greater than the increase in reward which results from the using the optimal versus approximate solution). Hence improvements with respect to balancing capacity further justify the ASA.

5 Discussion

In this paper we illustrate how capacity, case mix and patient mix decisions are interrelated. Understanding the complex relationships existing between these factors is crucial for achieving the maximum benefit from the DRG fee-for-service financing system. The case study highlights the impact of striving for a certain case mix without providing a balanced supply of resources.

The presented model can be extended to situations where the hospital has different motivators but still wishes to strike a balance between capacity, case mix and patient mix. In the remainder of this section we discuss three such situations.

Consider a hospital which wishes to maximize utilization of its resources instead of maximizing the number of completed treatments. By changing the reward function to account for the difference between $S_m(t)$ and $G_m(t)$ such a motivator can be accommodated.

Although this work has been applied in a hospital with fee-for-service financing, the model also has merit for health care system operating with a lump-sum budget. In such systems, hospitals typically work together with more complicated cases being treated in larger general hospitals and less complicated cases being treated in smaller regional hospitals. The decision on whether or not to treat a certain patient type at a certain hospital is similar to the problem addressed in this paper.

The third situation considers ongoing changes in the Dutch health care system where the renumeration of some DRGs can be negotiated. In this case hospitals may strive for a certain quota of treatments, as exceeding it (or not meeting it) may result in penalties. To incorporate this, the described model can be used by making the reward depend on completed treatments per year. Furthermore, because overproduction will be penalized, the capacity constraints can be removed and the capacity levels required to meet treatment quotas become a model output.

References

[1] Cutler, D.: Empirical evidence on hospital delivery under prospective payment (1990). MIT

[2] Dafny, L.: How Do Hospitals Respond to Price Changes? The American Economic Review **95**(5), 1525–1547 (2005)

[3] Erlenkotter, D.: Sequencing expansion projects. Operations Research **21**(2), 542–553 (1973)

[4] Folmer, K., Mot, E.: Diagnosis and Treatment Combinations in Dutch Hospitals. Tech. rep. (2003)

[5] Freidenfelds, J.: Capacity expansion-analysis of simple models with applications. North Holland (1981)

[6] Gross, D.: Fundamentals of queueing theory. Wiley-India (2008)

[7] Hochang, L.: Project selection problems for production-inventory-distribution scheduling in manufacturing plants. In: Computer-Aided Design, Engineering, and Manufacturing, pp. –. CRC Press (2000)

[8] Hreinsson, E.: Hydroelectric Project Sequencing using Heuristic Techniques and Dynamic Programming. In: Power Systems Computation Conference, Cascais, Portugal (1987)

[9] van de Ketterij, J., Schaepkens, F., de Vries, P.: Health information developments in the Netherlands, chap. DBC 2003 what's in it for me (2002)

[10] Leonard, K.J., Rauner, M.S., Schaffhauser-Linzatti, M.M., Yap, R.: The effect of funding policy on day of week admissions and discharges in hospitals: the cases of austria and canada. Health Policy **63**(3), 239 – 257 (2003)

[11] Luss, H.: Operations research and capacity expansion problems: A survey. Operations Research **30**(5), 907–947 (1982)

[12] Massey, W., Whitt, W.: Networks of infinite-server queues with nonstationary Poisson input. Queueing Systems **13**(1), 183–250 (1993)

[13] Morin, T.: Optimal sequencing of capacity expansion projects. Journal of the Hydraulics Division **99**(9), 1605–1622 (1973)

[14] Morin, T.: Multidimensional sequencing rule. Operations Research **23**(3), 576–580 (1975)

[15] Morin, T.L., Shin, Y.S.: Optimal expansion of flood control systems, volume 1. Northwestern University (1977)

[16] Neebe, A., Rao, M.: The Discrete-Time Sequencing Expansion Problem. Operations Research **31**(3), 546–558 (1983)

[17] Neebe, A., Rao, M.: Sequencing capacity expansion projects in continuous time. Management Science **32**(11), 1467–1479 (1986)

[18] Rauner, M.S., Schaffhauser-Linzatti, M.M.: Impact of the new austrian inpatient payment strategy on hospital behavior: a system-dynamics model. Socio-Economic Planning Sciences **36**(3), 161 – 182 (2002)

[19] Sutherland, J., Hamm, J., Hatcher, J.: Adjusting case mix payment amounts for inaccurately reported comorbidity data. Health Care Management Science **13**, 65–73 (2010)

[20] Taylor, R.G.: A general form for the capital projects sequencing problem. In: Proceedings of the 21st international conference on Computers and industrial engineering, pp. 47–50. Elsevier Science Publishers Ltd., Essex, UK (1997)

[21] Westert, G.P., Burgers, J.S., Verkleij, H.: The Netherlands: regulated competition behind the dykes? BMJ **339**, 839–842 (2009)