Memorandum 1991 (September 2012). ISSN 1874-4850. Available from: http://www.math.utwente.nl/publications
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Organizing Multidisciplinary Care for Children with Neuromuscular Diseases

M.F. van der Velde[12], N. Kortbeek[1234], N. Litvak[12]

[1]Center for Healthcare Operations Improvement and Research (CHOIR),
University of Twente, Drienerlolaan 5, 7500 AE Enschede, The Netherlands

[2]Stochastic Operations Research, Department of Applied Mathematics,
University of Twente, Drienerlolaan 5, 7500 AE Enschede, The Netherlands

[3]Department of Quality and Process Innovation,
Academic Medical Center Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

[4]n.kortbeek@utwente.nl

The Academic Medical Center (AMC) in Amsterdam, The Netherlands, recently opened the 'Children's Muscle Center Amsterdam' (CMCA). The CMCA diagnoses and treats children with neuromuscular diseases. These patients require care from a variety of clinicians. Through the establishment of the CMCA, children and their parents will generally visit the hospital only once a year, while previously they visited on average six times a year. This is a major improvement, because the hospital visits are both physically and psychologically demanding for the patients. This article describes how quantitative modelling supports the design and operations of the CMCA. First, an integer linear program is presented that selects which patients to invite for a treatment day and schedules the required combination of consultations, examinations and treatments on one day. Second, the integer linear program is used as input to a simulation to study to estimate the capacity of the CMCA, expressed in the distribution of the number patients that can be seen on one diagnosis day. Finally, a queueing model is formulated to predict the access time distributions based upon the simulation outcomes under various demand scenarios.

Key words: Health care management; Patient flow; Appointment scheduling; Queueing systems; Integer linear programming.

1. Introduction

The Academic Medical Center (AMC) Amsterdam, The Netherlands, recently opened a center for children with neuromuscular diseases. Neuromuscular diseases is the generic term for a broad set of disorders which impair the functioning of the muscles via muscle or nerve pathology. Most of the diseases are progressive in time, sometimes leading to an early death of the patient (10). Most neuromuscular diseases have no cure, so the goal of the treatment is to reduce symptoms, and increase both mobility and life expectancy (19). Examples of neuromuscular diseases are the diseases of Duchenne, Becker and Charcot Marie Tooth (26). Children with neuromuscular diseases typically need care from various physicians and therapists. To drastically reduce the required number of hospital visits, and to improve the coordination of care, the AMC decided to cluster the expertise of these care providers each several weeks in the 'Children's Muscle Center Amsterdam' (CMCA). This article describes how quantitative modelling supports the AMC in the design and operations of the center.

Due to their disease, children may suffer from a variety of symptoms. Therefore, children are usually seen by a rehabilitation physician, a neurologist, a clinical geneticist, a cardiologist, and a pneumonologist. In addition, psychologists, dietitians and even cardiac surgeons may be required. Accurate coordination of such multidisciplinary treatment is crucial to achieve high quality of care. If such coordination is insufficient, under- or overtreatment may take place, treatments may be performed in a non-optimal order, or certain aspects of the disease may be overlooked. However,

since the different disciplines are accommodated at different locations within the hospital, this coordination is a challenging task.

With the opening of the CMCA, care coordination for children will be significantly improved. Regularly, a treatment day is organized on which the required disciplines come together to see multiple patients. Physicians discuss the condition of the different patients, so that diagnoses are settled earlier and treatments are better customized. Through the establishment of the CMCA, children and their parents will generally visit the hospital only once a year, while previously they visited on average six times a year. This is a major improvement, because the hospital visits are both physically and psychologically demanding for the patients. Also, the great responsibility and burden is taken away from the parents: to gather all relevant information from the different hospital visits and to schedule the right appointments at the right point in time. Consequently, the CMCA will simultaneously increase quality of care and patient-centeredness.

However, realizing centralized care is not an easy task. First of all, all physicians should cooperate and reserve time for the treatment days in their already busy schedules. Second, for each treatment day patients have to be selected and scheduled in an optimal way. Since the schedules are heavily constrained, construction by hand is very time consuming and does not guarantee the best solution. Third, due to the small size of the patient group, the treatment days are not often organized, and due to the many constraints, only a few patients can be scheduled in one treatment day. Consequently long access times may arise. This article will show that Operations Research techniques can be very helpful in these design and control issues. It contributes to the logistic questions on two levels:

- ***Treatment day scheduling.*** First, we developed a day scheduling algorithm to tackle the challenging task of scheduling the required combinations of consultations, diagnostics and treatment in combination on one day. Based upon the analysis of historical data and interviews with physicians and therapists, we collected all relevant restrictions and preferences. Next, we developed an Integer Linear Program (ILP) in close cooperation with the neuromuscular disease care experts. This ILP simultaneously selects which patients to invite for a particular treatment day, and generates an optimal day schedule, in compliance with all restrictions and preferences.
- ***Access time evaluation.*** Second, we developed a model to derive the distribution of the access times of newly diagnosed patients. Simulations are performed in which the scheduling algorithm is iteratively applied. As such, probability distributions of the number of patients that can be invited to one treatment day is obtained. These distributions are input for a Markov model by which access time distributions are determined. Since the CMCA has started very recently demand predictions are very uncertain, estimations run from 20 to 50 new patients per year. Therefore, various demand scenarios are considered. The influence of the several constraints on the day schedules on access times are analysed and improvements are proposed.

The outcomes of this study are used to advise the AMC on how often treatment days should be organized, for which care providers the availability should be reconsidered, and which day-schedule preferences could better be dropped. For example, initially the physicians were to have at least two appointments per day. Our research showed that such constraint resulted in very long access times because only patients with the same needs could be scheduled together. Basing on our results, this constraint and several other highly restricting constraints have been adjusted or removed.

This paper is organized as follows. Section 2 describes the characteristics of the case study setting. Section 3 gives an overview of related literature. Section 4 presents the ILP model for planning of a treatment day. The planning algorithm is applied to data of the aimed patient group, and the results for these patients are presented. Based on the results of this planning algorithm, an access time model is derived in Section 5, and numerical results are given for the AMC case. The paper ends with a discussion in Section 6.

2. Background

The CMCA aims at children up to eighteen years old who have a neuromuscular disease (follow-up patients), or are suspected of having one (new patients). The CMCA does not invite follow-up and new patients on the same day, because a different team of physicians is required. Therefore, 'diagnosis days' are organized for patients suspected of having a neuromuscular disease, and 'follow-up days' for patients who have already been diagnosed. A summary of the patient flow is displayed in Figure 2.

Preconsultation. When a physician suspects a neuromuscular disease, the patient and the parents are first asked to fill out a questionnaire at home, which is then assessed by the CMCA. If the questionnaire does not support the suspicion of a neuromuscular disease, the patient will not be invited to the CMCA.

Diagnosis day. If a patient is eligible for a diagnosis day, a set of required consultations and examinations is determined during a meeting between the 'core members' of the multidisciplinary treatment team. The core members are the physicians who are together responsible for the patient's treatment. They will all see the patient during the diagnosis day. The team is completed by a nurse practitioner who provides administrative support to both physicians and patients. For diagnosis days the core of the team is formed by a paediatric neurologist, a clinical geneticist, and the nurse practitioner.

Next, the patient is scheduled to come to the AMC for a diagnosis day. On this day, there will first be an intake meeting between the patient and the nurse practitioner. Then, the prescribed

Figure 1 Patient flow diagram.

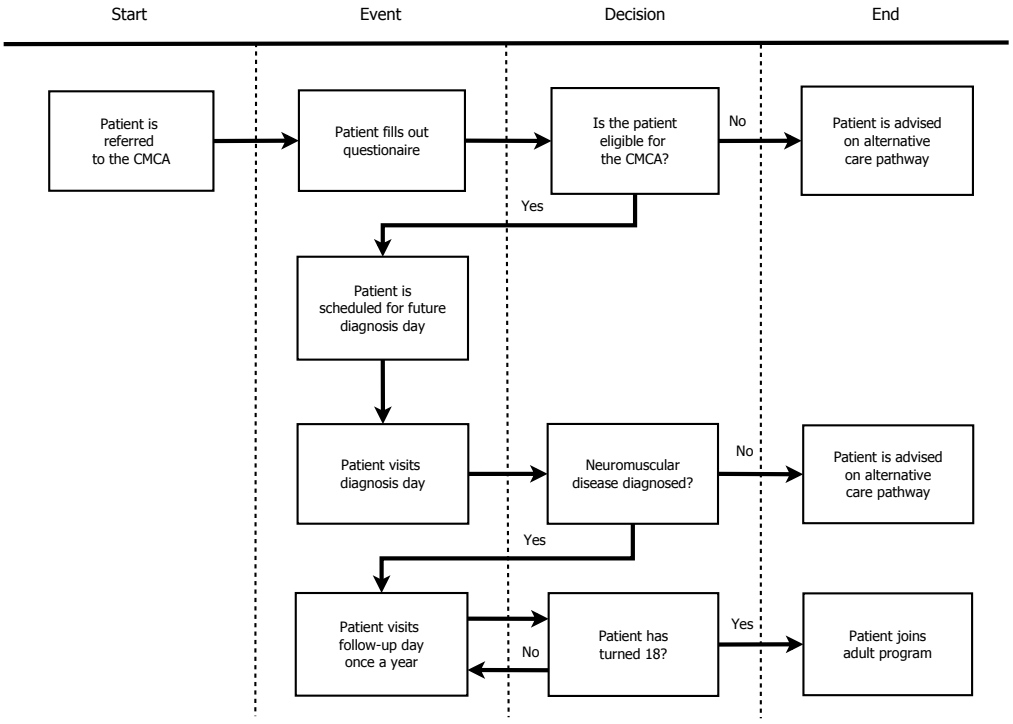


Table 1 The required appointments per patient type type.

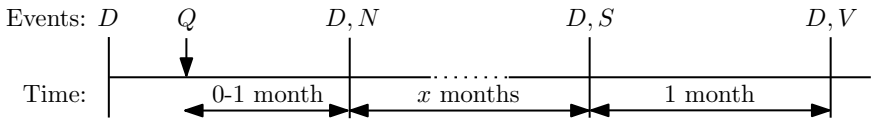
		Appointment (minutes)														
		Necessary				*Desirable*										
		Intake and evaluation (3 x 15)	Clinical geneticist (45)	Neurologist (45)	MTM (15)	Paediatrician (30)	Rehabilitation physician (30)	Physiotherapist (45)	Blood examination (15)	Clinical photograph (15)	Cardiac ultrasound / ECG (60)	EMG (75)	Muscle ultrasound (45)	MRI (60)	X-ray (30)	Needle biopsy muscle (120)
Patient type	*%*	*Required by x% of the patients*														
Myopathy (MP)	39	100	100	100	100	50	50	50	100	100	50	15	100	15	20	15
Neuropathy (NP)	39	100	100	100	100	50	50	-	100	100	-	50	-	-	-	-
Spinal muscular atrophy (SMA)	20	100	100	100	100	50	50	50	100	100	-	15	50	-	10	-
Neuromotor disease (NMD)	2	100	100	100	100	50	50	-	50	100	-	-	-	-	-	-

consultations and examinations will take place. Halfway the afternoon, a Multidisciplinary Team Meeting (MTM) is scheduled in which the outcomes of the consultations and examinations are discussed. If possible a diagnosis is settled, and a care plan is designed. Otherwise, additional examinations are scheduled (not on the same day). After the MTM, the neurologist shares the conclusions with the patient. During the day the nurse practitioner is present to act as a host for the patient and to guide the patient through the different examinations. The day finalizes with a meeting between the patient, parents, the nurse practitioner, to answer remaining questions and to explain the further care pathway.

Follow-up programme. If the conclusion of the diagnosis day is that the patient has a neuromuscular disease, he will continue to the follow-up programme. During this programme, the physicians monitor the health of the patient and give advice on how to reduce and handle symptoms. Most of the check-ups have to be performed annually, therefore the patient will visit a follow-up day once a year. Here, the core of the team consists of a paediatric neurologist, a paediatrician, a rehabilitation physician and the nurse practitioner. Also, the set of required appointments is different than for the diagnosis day and depends on the type and severity of the disease a patient suffers from. The set-up of a follow-up day is as follows: an intake with the nurse practitioner, examinations, a first MTM, consultations, a second MTM, a feedback consultation by a rehabilitation physician, and a final consultation with the nurse practitioner. When the patient turns 18, he will proceed to the adults track.

Day schedule. A month prior to a diagnosis day, the patients are selected who are preferably invited for the diagnosis day of the next month. If there are it at least two candidate patients, a diagnosis day is scheduled. Otherwise, the hospital considers it to be inefficient. A feasible day schedule has to be composed to asses how many patients can actually be invited. The schedule of a treatment day is highly constrained: some physicians are only available on specific times of the day, for some consultations several physicians have to be present, appointment precedence constraints have to be satisfied, all results of diagnostic tests have to be available before the MTM can start, etcetera. Based upon the predominant suspected disease, patients are assigned to one out of four patient types. For each type, the percentage of patients that require a certain appointment

Figure 2 Time line patient access time to diagnosis day (Legend: D = diagnosis day, Q = questionnaire received, N = not yet scheduled because of the waiting list, S = scheduled for next treatment day, V = visit takes place).



is listed in Table 1. These numbers are based on estimations of the involved physicians and data from the patients who have been treated before, outside the CMCA. The core team members subdivide for each patient the list of required consultations and examinations in 'necessary' and 'desirable' appointments. A patient is invited to a diagnosis day if all necessary appointment can be scheduled. Further, the trade-off has to be addressed between skipping some of the desirable (but not necessary) appointments in order to invite more patients, or keeping all appointments and inviting less patients. Clearly, the latter option will result in longer access times. In Section 4, an integer linear program is presented that simultaneously addresses: (1) rational patient selection in conjunction with the appointments to be executed, and (2) the creation of a day schedule. The resulting access times are analysed in Section 5.

Access times. The next concern in designing and operating the CMCA are the access times that are to be expected for diagnostic patients. For follow-up days access times are no major issue, since the candidate patient are known well in advance, and patients are required to revisit between 12 to 15 months after their last treatment day. Therefore, the access times for follow-up patients are well-predictable. This, in contrast to the access times of new patients. For these patients the access time is counted as the number of days between the reception of the completed questionnaire and the patient's visit. The time line for a patient to get access to a diagnosis day is illustrated in Figure 2. The AMC strives for a maximum access time of seven weeks for diagnosis days, which is quite ambitious in the initial setting, as diagnosis days are initially intended to be planned once a month. The access times will grow rapidly if the number of patients treated in a diagnosis day is insufficient. However, due to the variety in patient types, the complexity of the set of scheduling constraints, and diverging availability of the different care providers, the number of patients that can be invited for each diagnosis day cannot be easily predicted. Section 5 addresses this issue by presenting a Markov model by which the access time distribution for diagnosis patients is derived based upon the probability distribution of the number of patients that can be seen on a particular diagnosis day.

3. Literature

Designing appointment systems in health care is a topic that has received considerable attention in the literature. Two comprehensive surveys are provided in (4, 15). Appointment systems can be regarded as a combination of two distinct queueing systems. The first queueing system concerns customers making an appointment and waiting until the day the appointment takes place. The second queueing system concerns the process of a service session during a particular day. Gupta and Denton (15) identify several open challenges in appointment scheduling, of which prominent ones are planning coordinated packages of care for patients needing treatment from several health services, scheduling in highly constrained situations, and linking the access process and the day process. These challenges are addressed in the current paper.

Appointment scheduling concerns designing blueprints for day-appointment schedules with typical objectives as minimizing customer waiting time, and maximizing resource utilization or minimizing resource idle time (16). The literature has mostly focused on scheduling a given number

of single appointments on a particular day for an individual service provider (4). Scheduling multiple appointments at once for a single discipline for a planning horizon of one day or one week is done in (6, 7, 23), without considering access times. In these references, given sets of physical therapy treatments of given sets of patients on a particular day are scheduled. Chien et al. (6, 7) formulate this problem as a hybrid shop scheduling problem, and solve it by developing a genetic algorithm (7), combined with data mining techniques in later work (6). Podgorelec and Kokol (23) develop a scheduling algorithm based on genetic algorithms and machine learning.

ILP approaches for highly constrained *mono*disciplinary treatment planning can be found in (8, 9, 22, 25) for radiotherapy and chemotherapy treatment planning. For these patients treatments have to be scheduled during a given number of weeks, strictly taking into account the required rest periods. When access times have to be minimized, it is important to have good rules according to which patients are selected to be admitted from the waiting lists (17, 27). Conforti et al. (8, 9) develop an ILP for radiotherapy treatment planning, so that a maximum number of patients is planned from the waiting list, thereby minimizing the access time of patients while maximizing device utilization. However, they do not explicitly evaluate access times. Turkcan et al. (25) develop a two stage ILP approach for solving a similar problem, but without the patient selection decision. In the first stage patients are assigned to days, and in the second stage appointment times are given to all patients on their assigned days. The objectives are minimizing access times, treatment delays and staff overtime. A time horizon of a week is considered in (22), in which one appointment per patient should be planned for a single discipline. Ogulata et al. (22) develop an integer linear programming (ILP) model, that is broken down into three manageable hierarchical stages to resolve computational difficulty. In the first stage patients are selected, in the second stage patients are assigned to therapists, and in the third stage patients are scheduled throughout a day. A different three-stage hierarchical model is introduced in (24), where mixed ILP is used to design schedules for multidisciplinary rehabilitation care, on a larger planning horison. The applicability of the model in (24) is enhanced by allowing exogeneously planned appointments. An ILP approach to schedule *multi*disciplinary treatments one-by-one for rehabilitation patients is presented in (2). In this reference, access times are evaluated by simulation.

Models that do evaluate access times without considering day schedules are available for homogeneous patient populations. Examples of isolated access time models based on continuous-time queueing models are the $M(t)|M|s(t)$ queue as studied in (13) and the adapted $M|M|s$ queue that models time-dependent demand as studied in (14). Models to evaluate the length of hospital waiting lists are introduced in (29), and further studied in for example (11). To preserve the discrete-time nature, (18) presents a slotted queueing model in discrete time that is solved by a generating function approach based upon (3). We model access times as waiting times of customers in a queue with batch service, where batch size is at least two, and the maximum batch size is derived from the solution of the scheduling problem. Queueing systems with batch service were first considered by Bailey (1), motivated, as in this paper, by evaluation of access times for out-patients in hospitals. Other applications mentioned in the literature are in transport, control of traffic flows, and manufacturing. There is a vast literature on the analysis and numerical evaluation of queues with batch service, see e.g. (5, 20, 21, 12). In this paper we approximate the queueing process with a finite Markov chain and use the renewal theory to derive stationary waiting times.

4. Appointment scheduling

In this section, a mathematical model is formulated and implemented to decide which patients are invited to visit the center for the next treatment day (both diagnosis and follow-up) and to compute an optimal schedule for this day. We first give an overview of the properties of the model. Next, we make some remarks on computation of the solution, and finally, present the results for diagnosis days. For clarity of presentation, the detailed mathematical formulation of the model

is displayed in the appendix. The model was formulated in close cooperation with the CMCA healthcare professionals. Several versions of the model were designed and tested. Each time, the formulation and the inclusion or exclusion of specific constraints and objectives were discussed based upon the outcomes on various test problems. Here, we restrict ourselves to presenting the end result.

4.1. Model formulation

We model the construction of a day schedule for a treatment day as an Integer Linear Program (ILP). To do so, we divide a treatment day in time slots of equal length. The decisions to be made are: which patient gets which appointment at what time slot with which resource(s), which can be staff members and/or equipment. Thus, the decision variables are:

$$z_{j,p,s,t} = \begin{cases} 1 & \text{if patient } j \text{ has an appointment with resource } s \\ & \text{for procedure } p \text{ starting at time slot } t, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

We say that a patient has a *complete visit* if all his/her appointments are scheduled (both the necessary and desirable appointments, recall Section 2). If some of the desirable (but not necessary) appointments are omitted we say that a patient has a *partial visit*. We are interested in which patients have a complete visit, which have a partial visit, and at what time they have which appointment. The formal description of these variables can be found in Table 5, see Appendix A. The constraints and objectives will be described in the next subsections, the mathematical formulation is given in Appendix A.

Constraints. We distinguish several types of constraints:

Patient selection. A patient has most of his appointments, all of them, or none. The visit of such a patient is thus a complete visit, a partial visit, or the patient is not scheduled. The amount of appointments that are allowed to be omitted is patient specific. At least two patients should be scheduled on a treatment day for it to take place. Patients are scheduled according to the first-come-first-serve discipline.

Basic constraints. A patient gets each treatment at most once, a treatment is carried out by a resource that has the necessary qualifications, a resource can only be scheduled at one place at a time and should be available.

Precedence constraints. Some treatments have to be performed before others, there is a minimum amount of time between the starting times of some combinations of treatments.

CMCA specific appointment constraints. Some of the appointments of a patient can take place simultaneously. For example, an orthopedist and physiotherapist can treat one patient at the same time. A patient needs time to rest, therefore, at least half an hour per three hours must be free from appointments. Some appointments are obligatory. If these are not scheduled, the patient cannot be scheduled.

Patient specific constraints. A patient cannot get more appointments than he or she can physically handle in a day. Therefore, sometimes appointments of a patient have to be spread over several days. This number of days is given per patient. However, an appropriate fraction of total appointment time should be scheduled on the first treatment day in order to avoid scheduling problems in future.

Multidisciplinairy Team Meeting (MTM) constraints. The MTM is scheduled as a series of appointments, one per patient, in consecutive time slots, to make it just one meeting. All core team members should be present at the MTM. In some cases, the MTM must start at a fixed time. In other cases, this time may be flexible.

Defining constraints. The final constraints are required for the mathematical formulation. They determine the starting and end times of the patients and staff members, and determine whether a patient has an appointment at a certain time or not.

Objectives. The following objectives have been formulated, in descending order of priority:

1. Maximize the number of patients that have a complete visit
2. Maximize the number of patients that have a partial visit
3. Maximize the treatment time of all scheduled patients
4. Minimize the idle time in the schedules of the staff
5. Minimize the idle time in the schedules of the patients

The objective of the ILP is to maximize the sum of the weighted rewards on these objectives. In the objective function (see (34)), objectives 1, 2, 3, 4, and 5 are respectively rewarded by weight factors $\alpha, \beta, \gamma, \delta$, and ϵ. One may observe that the objective function contains multiple goals that are possibly in conflict. By varying the weight factors in the objective function, the relative importance of the various goals can be specified.

4.2. Computed schedules

The ILP was implemented using the program AIMMS. The solver employed is Cplex 12.2, using the branch and bound technique. A run is stopped as soon as the gap between the LP-bound and the best solution so far is less than 1%, thus, when a nearly optimal solution has been found. The first four objectives have always found their optimal value by then, the fifth not necessarily.

All input parameters of the ILP are set according to the real AMC data. Table 2 displays the availability of all resources during a diagnosis day. For brevity, we do not include other data such as appointments precedence and qualifications of the staff to carry out the procedures. Table 3 lists the values used for weight factors. To determine these values, the CMCA clinicians scored the relative importance of objective on a 0-10 scale. As the objectives are not measures on the same scale, we applied a normalization factor to each factor in order to get comparable measures. These normalization factors, multiplied by the relative importance, resulted after several calibration runs in the listed weight factor values. An example of the result for the schedule of a diagnosis day is displayed in Figure 4.2.

Table 2 Resource availabilities on a diagnosis day.

Resource	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00
Neurologist										
Clinical geneticist										
Nurse practitioner										
Rehabilitation physician										
Physiotherapist										
Blood examination										
Clinical photograph										
Cardiac ultrasound / ECG										
EMG										
Muscle ultrasound										
MRI										
X-ray										

Table 3 Weight factor values.

Objective	Weight factor	Importance	Value
1	α	10	100
2	β	8	50
3	γ	10	2
4	δ	6	20
5	ϵ	5	2

Several bottlenecks have been identified in the scheduling of diagnosis days. The following issues restrict the capacity of the CMCA, and need consideration when the CMCA desires to expand:

- Each patient has to visit the clinical geneticist for 45 minutes. However, this physician is available only from 10:30h. These consultations have to take place before the examinations. Since a fifth patient can visit the clinical geneticist at the earliest at 13:30h, just a little time is left for the examinations.
- The result of the blood examination has to be known before the MTM. However, obtaining these results takes two hours, and the blood examination cannot be done before the consultation with the clinical geneticist. Thus, at most three patients that can have a blood examination, regardless of their other appointments.
- Each patient has two appointments after the MTM, one with the neurologist and one with the nurse practitioner. In combination with the growing length of the MTM as there are more patients, this results in less time for consultations and examinations before the MTM.
- Half of the patients with a neurological disease need to have an EMG examination. The examination takes more than an hour, and the outpatient clinic is closed during lunch time (12:00-13:00). Therefore, at most two patients can have this examination on one day. When there are five patients on one diagnosis day, just one patient can take the EMG examination, regardless of the other appointments needed.

5. Access time analysis

Recall that an access time of a patient is defined as the time span from returning the questionnaire until being present at a diagnosis day. In this section, we derive the distribution of access times, assuming that the patients are scheduled in the first-come-first-serve (FCFS) fashion. First, the capacity of a diagnosis day of the CMCA is analysed. Then, a Markov model is formulated to evaluate access time distribution. Finally, we present numerical results on various demand scenarios.

5.1. Number of scheduled patients per diagnosis day

The number of patients that can be scheduled in one diagnosis day is defining for access times. However, due to the complexity of the scheduling problem, this number cannot be directly modelled or predicted, therefore, a simulation study has been performed to determine its distribution. A list of 5900 patients and their arrival times was constructed at random, based on the data given in Table 1. Next, for consecutive diagnosis days, the first five patients on the list were selected, and an optimal schedule was constructed. If less than five patients were on the waiting list, then the optimal schedule was constructed for all patients on the list. When a patient was scheduled, he was deleted from the list. Table 4 shows the distribution of the number of patient visits scheduled, given the waiting list size. We schedule the patients in groups of five because, as observed from Table 4, there is a high probability that a group of five patients can be scheduled, while it is never possible to schedule six patients. The latter statement can easily be proven by combining the information

Figure 3 An exemplary day schedule for a diagnosis day.

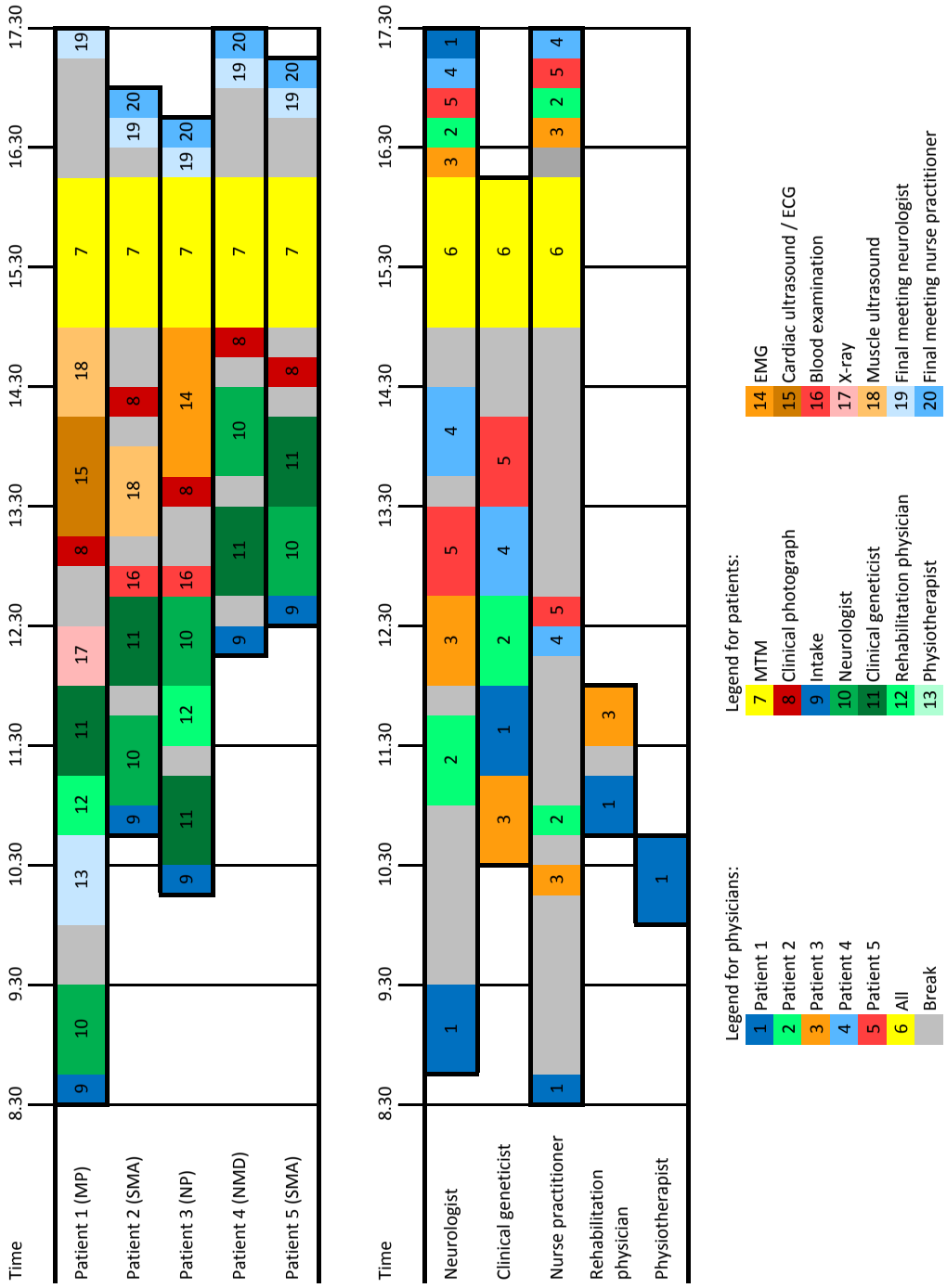


Table 4 The distribution of the capacity of diagnosis days.

	Number of patients scheduled					
Length waiting list	0	1	2	3	4	5
1	100%	-	-	-	-	-
2	-	-	100%	-	-	-
3	-	-	-	100%	-	-
4	-	-	-	0.3%	99.7%	-
5	-	-	-	0.3%	13.7%	86.0%
6	-	-	-	0.3%	13.7%	86.0%

on the availability of the clinical geneticist, the blood examination, the length of the MTM and the priority of complete visits over partial visits. If two patients are scheduled both have a complete visit. If three or more patients are scheduled, then in almost all cases (99%) three patients have a complete visit and the others have a partial visit, otherwise two patients have a complete visit and the others have a partial visit.

As soon as one patient cannot be scheduled while his predecessors have been scheduled, it is obligatory to schedule this patient on the next diagnosis day. Simulations have shown that mainly the needs of four scheduled patients, and not the needs of a fifth patient, determine whether the fifth patient can be scheduled or not. Thus, we can assume that the number of patients scheduled on a diagnosis day depends only on the size of the waiting list, and is independent of how many and which patients were scheduled for other diagnosis days.

5.2. Model description

We model the arrivals of new patients as a Poisson process, of which the arrival rate λ (patients per year) is known. The service discipline is FCFS. A year is split into m time periods of equal length, with one diagnosis day per time period. At the opening of the CMCA, the value for m proposed by the AMC is twelve.

Recall the procedure given in Figure 2. An access time of a patient consists of three parts. (1) The time until the end of the time period. This time is stochastic and has a uniform distribution. (2) The number of full time periods the patient has to wait until being scheduled. This is stochastic, and has a discrete distribution W which has to be determined. (3) The time between being scheduled and the actual visit to the hospital. This time is deterministic.

Let A_n be the random number of arrivals in time period n. Denote by Q_n the number of waiting patients at the *end* of time period n. Out of Q_n waiting patients, a random number B_n of patients are scheduled for diagnosis day n. The distribution of B_n depends on the value of Q_n. Table 4 contains conditional probabilities $P(B_n = b|Q_n = q)$, $q \geq 0$, $0 \leq b \leq q$, for the CMCA case study, obtained from simulations as discussed in Section 5.1.

The main performance characteristic of interest is the access time of the patients. Denote by W_n the access time of a patient that arrived at time period n. The distribution of W_n depends on the queue length at the *beginning* of time period n, denoted by Y_n. The following equations hold:

$$Q_n = Y_n + A_n, \tag{2}$$

$$Y_{n+1} = Q_n - B_n, \quad n = 1, 2, \ldots. \tag{3}$$

Here A_n is independent of the other random variables, and B_n depends on Q_n. Assuming that the arrival rate is not too high, it will often happen that all waiting patients in the queue are scheduled. In that case, the stochastic process W_n will soon reach stationarity, therefore, we choose to obtain its stationary distribution W. To this end, we first determine the stationary distribution Y of Y_n and then obtain the distribution of W using the renewal theory argument. The details of the derivation are provided in the Appendix B.

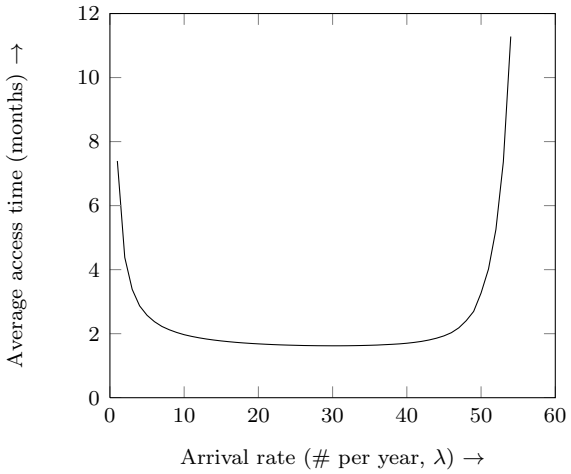


Figure 4 Average access times in months (m=12).

5.3. Numerical results

The distribution of W is evaluated numerically, by approximating Y_n with a finite Markov chain. This approach is justified by the fact that the queue lengths are typically short. We will present the results computed for our case study. The initial frequency of diagnosis days that the CMCA will apply is (at most) twelve such days per year, therefore, $m = 12$. Next, as stated in Section 2, the arrival rate is estimated to be between twenty and fifty patients per year. The time between being scheduled for a diagnosis day and the actual visit to CMCA is precisely one month. For these input parameters we obtain the total average access times, from the time the questionary is received till the hospital visit. The results are presented in Figure 4. We see that excessively large access times are observed in two extreme cases. When the arrival rate is small, less than ten patients per year, large access times arise because at least two patients have to be scheduled on one day, and thus arriving patients often have to wait for another arrival. When there are more than fifty patients a year, large average access times arise since the maximal capacity is almost met. In between, the average access time is stable at a value just below two months.

The shape of the distribution of the access times heavily depends on the arrival rate λ. This can be seen in Figure 5. When the arrival rate is low, the moment in a month when a patient arrives does not have any influence on the distribution of the access times. However, as the value of λ increases, a heavier dependence shows, because the patients that arrive at the beginning of a month have a considerably higher chance to be scheduled earlier.

We emphasize that the access times are heavily dependent on the constraints of the scheduling problem. When the bottleneck constraints, as mentioned in Section 4.2, are relaxed, then large groups of patients can be scheduled together. For example, if the clinical geneticist would be available all day, it will in some cases be possible to schedule one more patient per day. This will increase the maximum capacity and result in smaller access times.

6. Discussion

We showed how combining integer linear programming, simulation, and queuing theory helps the AMC in organizing care for children with neuromuscular diseases. The treatment center embodies a transformation from supply-driven to demand-driven patient care. Customized diagnostics and treatment can be offered in a combined visit. To realize this, all practical constraints and preferences were collected and incorporated in an ILP by which feasible day schedules for multiple patient visits can be constructed. Simulations give insight in the capacity of the CMCA, given the availability

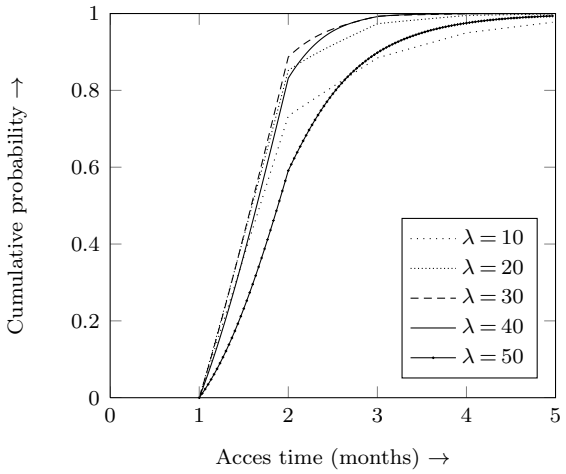


Figure 5 Distributions of access times (m=12).

of staff and equipment and estimates on patient demand in the number of arrivals and required appointments. Finally, a Markov model predicts the access time distributions for diagnostic patients based upon the simulation outcomes.

Formulating the day scheduling model was an iterative process, intensively involving the clinicians. Results on initial formulations predicted very long patient access times. Presenting these results to the clinicians, yielded that some highly restricting constraints were loosened or deleted. For example, there was a constraint that the physicians were to have at least two appointments per day. Thus, only patients with the same needs could be scheduled together, resulting in long access times. Another example, where the AMC still struggles with, is the choice whether the MTM should start at a fixed time or not. Although a fixed time is preferred by the clinicians, from a patient's point of view, based on the experimentation outcomes, we strongly recommend it to be flexible. We believe that the benefit of quantitative analysis in such a 'negotiation' process is that it rationalizes the process of realizing a good trade-off between interests of clinicians and patients.

The main limitation of our study is the availability of accurate data. Since the center just opened its doors, no historical data was available besides data on realizations of how the treatment was previously delivered by different outpatient clinics. Having a focused care center may increase the attractiveness for patients to come to the AMC, which makes predictions on the number of patient arrivals highly uncertain. In addition, patient type mix and required appointments could only be estimated from physician's expert opinions and data on the former patient population. Therefore, we recommend the center to constantly monitor its operations, and to regularly repeat the analysis so to reconsider both the frequency of carrying out treatment days and the staff/equipment availability during a treatment day.

Several aspects have not been included in the study. For example, in a hospital there is often a problem of a no-show and appointments running late. A patient has to know several days in advance whether he has to come to the hospital. So in case a patient does not show up, it will not be possible to select a patient that could not be scheduled primarily. However, our methods allow to tighten the schedule by running the planning algorithm again with the reduced patient set.

In this paper, we have considered the first-come-first-serve discipline for patients to be admitted from the waiting list. There might be a discipline which gives a better performance, if such a discipline increases the number of patients that can be seen on one day. Investigating the existence of such a policy might be an interesting direction for future research. However, when changing the service discipline, the justification for the assumption of independence between the batches of patients scheduled on consecutive treatment days has to be reconsidered. Also, the issue of practical

acceptance needs to be addressed, since it questionable how clinically acceptable it would be to not admit the patient that has been on the waiting list the longest.

The first patients visited the CMCA in January 2011 in a pilot phase of the entire treatment concept. During this pilot phase, the nurse practitioner enters the needs of the patients in an Excel sheet. Given a set of patients with prescribed consultations and examinations, and the availabilities of the staff and equipment, the optimal schedule is determined using AIMMS. This is not the desired ultimate state, since it still requires copying the resulting appointments in the electronic agenda system by hand. Also, the AMC strongly opposes the implementation of different software tools in different parts of the hospital, to prevent the maintenance and support task of ICT department to become inefficient if not impossible. Therefore, the scheduling algorithm is intended to be incorporated in the new hospital-wide electronic agenda system that is currently under construction. For such a system, it will be required to be able to communicate with an ILP solver, which will be a main challenge for the ICT design. Modern ICT systems for hospital organization increasingly often embrace OR/MS solutions, in particular, in capacity evaluation and appointment scheduling. In the near future, advanced qualitative schemes, as the one described in this article, should become a standard part of hospitals' integral ICT support, for transparent and efficient planning of high quality care.

Acknowledgments

1.First, we would like to thank Ellen Dibbits and Bert Kiewiet for joining in the initial phase of this study. Next, we thank Gerhard Post for his input in formulating and implementing the ILP. Finally, we thank Liesbeth Flippo for being the interface in the communication with the involved clinicians.
2. 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.

Appendix A: Mathematical formulation

This appendix contains the mathematical formulation of the ILP described in Section 4.1.

Variables and Parameters

Recall from (1) that the decision variables are denoted by $z_{j,p,s,t} \in \{0,1\}$, that equals one if patient j has an appointment with resource s for procedure p starting at time slot t. Besides that, we use several other variables and parameters. As soon as a variable or parameter is used for the first time, it will be introduced briefly. A complete list of sets, indices, variables, parameters and their properties can be found in Tables 5 and 6.

Constraints

In the ILP, several types of constraints are considered. We distinguish: constraints on the selection of patients, basic planning constraints, precedence constraints, appointment constraints, MTM constraints, and defining constraints. Below each constraint is presented in detail.

Table 5 Sets and indices ILP.

Set	Description	Index
J	patients	j
P	procedures	p, p'
S	resources	s
T	time slots	t

Table 6 Parameters and variables ILP.

Notation	Description
Binary parameters	
$Q_{s,p}$	1 if resource s is qualified to perform procedure p
$A_{s,t}$	1 if resource s is available in time slot t
$C_{p,p'}$	1 if procedures p and p' can be performed simulteneously
$H_{p,p'}$	1 if procedure p has to be performed before procedure p'
$N_{j,p}$	1 if patient j needs to undergo procedure p
$E_{j,p}$	1 if appointment p is necessary for patient j
Integer parameters	
$F_{p,p'}$	minimal number of time slots before start of procedure p' after start of p
t_{MTM}	starting time slot of the MTM
p_{MTM}	procedure number of the MTM
$L_{j,p}$	number of time slots that procedure p takes for patient j. $L_{j,p}=0$ indicates that procedure p is not required for patient j.
m_j	maximum number of appointment time slots patient j can handle on a day
d_j	number of treatment days over which the appointments of patient j may be spread
o_j	maximum number of desired appointments that patient j is allowed to skip in a partial visit
Real parameters	
u_s	relative weight of idle time of staff member s
k_j	relative weight of patient j
Binary variables	
$z_{j,p,s,t}$	1 if patient j has an appointment with resource s for procedure p starting at time slot t
$x_{j,p,t}$	1 if patient j has an appointment for procedure p starting at time slot t
g_j	1 if patient j has a complete visit
e_j	1 if patient j has a partial visit
$b_{j,t}$	1 if patient j has an appointment at time slot t
$c_{j,p,p'}$	1 if patient j has both appointments p and p' scheduled
General integer variables	
y_j^{min}	first time slot at which patient j has an appointment
y_j^{max}	last time slot at which patient j has an appointment
y_s^{min}	first time slot at which staff member s has an appointment
y_s^{max}	last time slot at which staff member s has an appointment

Selection of patients. A patients has most of his appointments, all of them, or none. He or she thus has a complete visit, a partial visit, or is not scheduled. We denote by g_j and e_j the binary variables that indicates whether a patient has, respectively, a complete visit ($g_j=1$) or a partial visit ($e_j=1$). If $d_j>1$ then only complete visit is allowed, which will be guaranteed by (17). A patient may only have a partial visit if $d_j=1$. Thus, we only define the variable e_j for patients that have $d_j=1$. A patient cannot have a partial and complete visit at the same time, therefore we have:

$$g_j+e_j\leq 1, \quad \text{for all } j. \tag{4}$$

The number of desired appointments that can be skipped is patient specific. This is given by o_j. The binary parameter $N_{j,p}$ denotes whether patient j needs procedure p or not. Denote by the binary $x_{j,p,t}$ whether patient j has an appointment for procedure p starting at time slot t or not.

The next constraint only needs to be satisfied if patient j has a partial visit, thus we formulate it as a big-M constraint (28) where $M_1 = |P|$ satisfies:

$$M_1 \cdot (1 - e_j) + o_j + \sum_{p,t} x_{j,p,t} \geq \sum_{p} N_{j,p}, \quad \text{for all } j. \tag{5}$$

At least two patients should be scheduled on a treatment day, otherwise it is cancelled:

$$\sum_{j} (g_j + e_j) \geq 2. \tag{6}$$

Patients are scheduled according to the first-come-first-served discipline:

$$g_j + e_j \geq g_{j'} + e_{j'}, \quad \text{for all } j, j' \in J \text{ such that } j < j'. \tag{7}$$

Basic planning constraints. A patient gets each treatment at most once. Denote by $L_{j,p}$ the number of time slots patient j needs to undergo procedure p. If $L_{j,p} = 0$, the patient does not need the procedure. Thus, we get the following constraint:

$$\sum_{t} x_{j,p,t} \leq 1, \quad \text{for all } j, p \text{ such that } L_{j,p} > 0. \tag{8}$$

A treatment is carried out by a resource that has the necessary qualifications. Denote by the binary $Q_{s,p}$ whether resource s is qualified to perform procedure p. Each scheduled procedure should have a qualified resource performing it at the intended time. This constraint, in (9), holds for all procedures except the MTM, for which we have a separate set of constraints. Furthermore, (9) in combination with (8) ensures that no dummy appointments are scheduled to reduce the idle time of staff members:

$$\sum_{s} z_{j,p,s,t} \cdot Q_{s,p} = x_{j,p,t}, \quad \text{for all } i, j, p \text{ such that } p \neq p_{MTM} \text{ and } L_{j,p} > 0. \tag{9}$$

A resource can only be scheduled at one place at the time and only if the resource is available for the CMCA. Denote by the binary $A_{s,t}$ whether resource s is available for the CMCA at time t. Recall that we have defined the starting time of an appointment with $z_{j,p,s,t}$ and the length with $L_{j,p}$, so for each time slot we have to look in the past whether an appointment has started which is not yet finished at the moment:

$$\sum_{j,p} \sum_{t'=t-L_{j,p}+1}^{t} z_{j,p,s,t'} \leq A_{s,t}, \quad \text{for all } s, t. \tag{10}$$

Precedence constraints. Some treatments have to be performed before others. For example, the intake appointment of the patient should be before all other appointments of the day. Denote with the binary $H_{p,p'}$ whether procedure p should be performed before procedure p' in case a patient needs both procedures. Since the constraint only holds for appointments p, p' that are both scheduled, we introduce the binary variable $c_{j,p,p'}$ which is one if both p and p' are scheduled for patient j. This variable is only relevant when both procedures are needed by the patient and there is a precedence constraint. The following constraint ensures $c_{j,p,p'} = 1$ when both procedures are needed:

$$\sum_{t} (x_{j,p,t} + x_{j,p',t}) - 1 \leq c_{j,p,p'}, \quad \text{for all } j, p, p' \text{ such that } H_{p,p'} = 1,\ L_{j,p} > 0,\ L_{j,p'} > 0. \tag{11}$$

Now we can formulate the precedence constraint, in combination with the constraint on a minimum amount of time between the starting times of some combinations of treatments. This minimum amount of time slots is denoted by $F_{p,p'}$. The big-M formulation ensures the constraint is always satisfied when $c_{j,p,p'} = 0$. In this constraint, a value of $M_2 = 3 \cdot |T|$ suffices:

$$\sum_t t \cdot (x_{j,p',t} - x_{j,p,t}) - L_{j,p} - F_{p,p'} + (1 - c_{j,p,p'}) \cdot M_2 \geq 0, \quad \text{for all } j, p, p' \text{ such that } H_{p,p'} = 1,\ L_{j,p} > 0,\ L_{j,p'} > 0. \tag{12}$$

CMCA appointment constraints. A patient can get some of his treatments simultaneously. Denote by the binary $C_{p,p'}$ whether the procedures p and p' can be carried out for the same patient at the same time. The next constraint checks, for each time slot and each combination of appointments, whether they are being performed or not. This is only relevant if the two procedures cannot be performed concurrently, and a qualified resource should be available. This is not relevant for the MTM, since patients are not present there.

$$\sum_{t'=t-L_{j,p}+1}^{t} x_{j,p,t} + \sum_{t'=t-L_{j,p'}+1}^{t} x_{j,p',t}, \leq 1, \quad \text{for all } j, p, p', t \text{ such that } C_{p,p'} = 0,\ p > p',\ \textstyle\sum_s A_{t,k} \cdot (Q_{p,s} + Q_{p',s}) > 0,\ p \neq p_{MTM},\ p' \neq p_{MTM}. \tag{13}$$

A patient needs a time to rest. Therefore, in the span of three hours, there is at least half an hour free from appointments. These breaks should have the length of at least one quarter of an hour. Denote by the binary variable $b_{j,t}$ whether patient j has an appointment at time t or not. In this variable, the MTM is not considered as an appointment because the patient is not present at the MTM. With time slots of fifteen minutes, this gives the following constraint:

$$\sum_{t'=t}^{t+11} b_{j,t} \leq 10, \quad \text{for all } j, t. \tag{14}$$

Note that when time slots have a different length, constraint (14) is easily adjusted to ensure the patient has enough time to rest. However, an additional constraint will be neccesary to ensure each break is at least fifteen minutes long.

Some appointments are obligatory: the 'necessary' appointments. If these are not scheduled, the patient cannot visit the CMCA. Denote by the binary $E_{j,p}$ whether an appointment is necessary or not. Thus, for all appointments that are necessary, we require:

$$g_j + e_j \leq \sum_t x_{j,p,t}, \quad \text{for all } j, t \text{ such that } E_{j,p} = 1. \tag{15}$$

Patient specific constraints. A patient gets no more treatment time than he can handle on a day. Denote by m_j the maximum number of time slots of appointments that patient j can have on one treatment day. Then we have the following constraint:

$$\sum_t b_{j,t} \leq m_j, \quad \text{for all } j. \tag{16}$$

Sometimes appointments of a patients have to be spread over several days because of the requirements of the patient. The number of treatment days a patient j has left is given by d_j. However, an appropriate fraction of appointment time should be scheduled on the first treatment day in order to avoid scheduling problems on later treatment days. Thus we have:

$$g_j \cdot \sum_{p \neq p_{MTM}} L_{j,p} \leq d_j \sum_t \sum_{p \neq p_{MTM}} x_{j,p,t} \cdot L_{j,p}, \quad \text{for all } j. \tag{17}$$

Multidisciplinary Team Meeting constraints. We schedule the MTM as one appointment for all patients. Using precedence constraints below, we will ensure the length of the MTM is correct. The core team members are formally assigned to the first patient:

$$\sum_s z_{j,p,s,t} = x_{j,p,t} \cdot \sum_s Q_{p,s} \quad \text{for all } t,\ p = p_{MTM},\ j = 1. \tag{18}$$

Now we define the precedence constraints for the MTM. Some of the appointments must be finished before the MTM. Thus, we define a constraint similar to (12):

$$\sum_t t \cdot (x_{j',p',t} - x_{j,p,t}) - L_{j,p} - F_{p,p'} + (1 - c_{j,p,p'}) \cdot M_2 \geq 0, \tag{19}$$
$$\text{for all } j,\ j',\ p,\ p' \text{ such that } H_{p,p'} = 1,\ L_{j,p} > 0,\ p' = p_{MTM}.$$

Some appointments can start only after the MTM. The length of the MTM for scheduled patient j is $(g_j + e_j) \cdot L_{j,p_{MTM}}$. This yields a constraint similar to (19):

$$\sum_t t \cdot (x_{j,p,t} - x_{j',p',t}) - \sum_{j''} L_{j'',p'} \cdot (g_{j''} + e_{j''}) - F_{p,p'}$$
$$+ (1 - \sum_t x_{j',p',t}) \cdot M_2 + (1 - \sum_t x_{j,p,t}) \cdot M_2 \geq 0, \tag{20}$$
$$\text{for all } j,\ j',\ p,\ p' \text{ such that } p' = p_{MTM},\ L_{j,p'} > 0,\ H_{p',p} = 1. \tag{21}$$

Sometimes it is desirable to always start the MTM at a fixed time. Denote by t_{MTM} the time slot in which the MTM should start. Then we obtain a constraint for the starting time of each MTM appointment:

$$\sum_t x_{j,p,t} = t_{MTM}, \text{for all } j,\ p = p_{MTM}. \tag{22}$$

Defining constraints. This group of constraints determines the starting and the end times of the patients and staff members. Denote by y_j^{min} the first time slot when patient j has an appointment. We have to take into account only those appointments that are actually planned, so we construct a big-M constraint. Here $M_3 = |I|$ is sufficient:

$$y_j^{min} \leq M_3 + (t - M_3) \cdot x_{j,p,t}, \quad \text{for all } j,\ p,\ t \text{ such that } p \neq p_{MTM}. \tag{23}$$

The last time slot when patient j has an appointment, y_j^{max}, is determined by the following constraint:

$$y_j^{max} \geq (t + L_{j,p}) \cdot x_{j,p,t}, \quad \text{for all } j,\ p,\ t \text{ such that } p \neq p_{MTM}. \tag{24}$$

Note that when patient j is not planned, y_j^{min} and y_j^{max} can take any integer value in the interval $[0, \ldots, |T|]$.

In a similar fashion we can derive the minimum and maximum values for staff members:

$$y_s^{min} \leq M_3 + (t - M_3) \cdot z_{j,p,s,t}, \quad \text{for all } j,\ p,\ s,\ t; \tag{25}$$

$$y_j^{max} \geq (t + L_{j,p}) \cdot z_{j,p,s,t}, \quad \text{for all } j,\ p,\ s,\ t. \tag{26}$$

The next constraint determines whether a patient has an appointment at a certain time or not, recall that this is denoted by the binary variable $b_{j,t}$. The following constraint forces $b_{j,t} = 1$ when

a patient has an appointment. Since a patient can have multiple appointments at one time slot, $M_4 = |P|$ satisfies the following inequality:

$$\sum_{p \neq p_{MTM}} \sum_{t'=t-L_{j,p}+1}^{t} x_{j,p,t'} \leq M_4 \cdot b_{j,t}, \quad \text{for all } j, t. \tag{27}$$

The following constraint ensures $b_{j,t} = 0$ whenever patient j has no appointment at time t:

$$b_{j,t} \leq \sum_{p \neq p_{MTM}} \sum_{t'=t-L_{j,p}+1}^{t} x_{j,p,t'}, \quad \text{for all } j, t. \tag{28}$$

Objective function

The objective function consists of several parts.

Maximize the number of patients that have a complete visit. Denote by k_j the relative weight of patient j. Then we want to maximize the following expression:

$$\sum_j g_j \cdot k_j. \tag{29}$$

Maximize the number of patients that have a partial visit. This expression is similar to (29):

$$\sum_j e_j \cdot k_j. \tag{30}$$

Maximize the treatment time of all scheduled patients. Note that constraint (8) ensures that no dummy appointments are being scheduled. We wish to maximize the total length of all scheduled appointments, so if e.g. two appointments with lengths $L_{j,p}$ and $L_{j,p'}$ are scheduled at the same time then we need to add $L_{j,p} + L_{j,p'}$ to the total treatment time. Thus, we want to maximize:

$$\sum_{j,p,t} x_{j,p,t} \cdot L_{j,p}. \tag{31}$$

Minimize the idle time in the schedules of the staff. We have already defined the starting and end times of a staff member. Since the idle time of some staff members (or resources) might be more important than that of others, we assign a relative weight u_s to the idle time of staff member (or resource) s. Thus, we wish to minimize the following expression:

$$\sum_s u_s \cdot (y_s^{max} - y_s^{min} - \sum_{j,p,t} z_{j,p,s,t} \cdot L_{j,p}). \tag{32}$$

Minimize the idle time in the schedules of the patients. It is assumed that this is equally important for all patients. Note that constraint (14) ensures that each patient has enough time to rest. Then the total idle time of the patients equals to

$$\sum_j \left(y_j^{max} - y_j^{min} - \sum_t b_{j,t} \right). \tag{33}$$

The expressions (29)–(33) contribute to the objective function, each having its own relative importance. The coefficients determining the relative importance are given by α, β, γ, δ and ϵ. Thus, we obtain the following objective function:

$$\max \qquad \alpha \cdot (29) + \beta \cdot (30) + \gamma \cdot (31) - \delta \cdot (32) - \epsilon \cdot (33). \tag{34}$$

Appendix B: Derivation of the waiting time distribution W

In this appendix we derive the stationary distribution of the stationary waiting time W defined in Section 5.2.

First, we write the transition probabilities for Y_n. From (3), by conditioning on $[Q_n = q]$ and noting that $P(B_n = q - i | Q_n = q) = 0$ whenever $q - i < 0$, we obtain:

$$P(Y_{n+1} = i | Y_n = j) = \sum_{q=\max(i,j)}^{\infty} P(B_n = q - i | Q_n = q) \cdot P(Q_n = q).$$

Next, using (2) we get:

$$P(Y_{n+1} = i | Y_n = j) = \sum_{q=\max(i,j)}^{\infty} P(B_n = q - i | Q_n = q) \cdot P(A_n = q - j)$$

From the transition probabilities above we determine the stationary distribution Y of Y_n. In the case study, we obtain an approximation for the stationary distribution. To this end, we bound the maximal value of Y_n with some large number N so that $P[Y \geq N]$ is sufficiently close to zero. Then the stationary distribution for the bounded chain is computed by numerically solving the balance equations. Finally, we approximate Y with the stationary distribution of the bounded Markov chain, and use $P[Y_n = k] = 0$ when $k \geq N$.

Now our goal is to derive the stationary waiting times. To this end, denote by $P(W \in \mathcal{A} | Y = i)$ the stationary probability that waiting time of an arriving patient is the number in a set $\mathcal{A} \subset \{0, 1, \ldots\}$, provided that there were $i = 0, 1, \ldots$ waiting patients at the beginning of the time slot of the arrival. Consider the sequence of time periods n such that $[Y_n = i]$. The distribution of the waiting times of the A_n patients arriving in such time period is completely defined by Y_n. Thus, given Y_n, these waiting times are independent of the waiting times of the patients arriving in the other time periods. Denote by $A_n^{(k)}$ the number of patients that have arrived in time period n and have to wait k time slots before being scheduled. Then using the renewal reward theory we write:

$$P(W = k | Y = i) = \frac{\mathbb{E}[A_n^{(k)} | Y_n = i]}{\mathbb{E}[A_n | Y = i]} = \frac{\mathbb{E}[A_n^{(k)} | Y_n = i]}{\mathbb{E}[A_n]} = \frac{\mathbb{E}[A_n^{(k)} | Y_n = i]}{\lambda / m}. \tag{35}$$

Let us now define the probability $P(W \geq 1 | Y = i)$. From (35) it follows that

$$P(W \geq 1 | Y = i) = \frac{\mathbb{E}[A_n - A_n^{(0)} | Y_n = i]}{\lambda / m}, \tag{36}$$

where for the numerator we write

$$\begin{aligned} \mathbb{E}[A_n - A_n^{(0)} | Y_n = i, A_n = k_n] = \sum_{b_n=0}^{i+k_n} & P(B_n = b_n | Q_n = i + k_n) \\ & \times \mathbb{E}[A_n - A_n^{(0)} | Y_n = i, A_n = k_n, B_n = b_n], \end{aligned} \tag{37}$$

and for the last component above holds:

$$\mathbb{E}[A_n - A_n^{(0)} | Y = i, A_n = k_n, B_n = b_n] = \min\{k_n, i + k_n - b_n\}. \tag{38}$$

Using (36) – (38) the probability $P(W \geq 1 | Y = i)$ can be directly computed.

Similarly, we can write the expression for $P(W \geq 2 | Y = i)$. Note that sometimes a patient has to wait longer because there are not enough patients on the list to form a batch of minimal size.

Thus, the waiting times of patients arriving in time slot n depend also on the arrivals in time slot $n+1$. Specifically, we derive the following:

$$P(W \geq 2|Y=i) = \frac{m}{\lambda} \sum_{k_n=0}^{\infty} P(A_n = k_n) \sum_{b_n=0}^{i+k_n} P(B_n = b_1|Q_n = i+k_n) \times \mathbb{E}[A_n - A_n^{(0)} - A_n^{(1)}|Y=i, A_n=k_n, B_n=b_n],$$

where

$$\mathbb{E}[A_n - A_n^{(0)} - A_n^{(1)}|Y=i, A_n=k_n, B_n=b_n] = \sum_{k_{n+1}=0}^{\infty} P(A_{n+1}=k_{n+1}) \times \sum_{b_{n+1}=0}^{i+k_n+k_{n+1}+b_n} P(B_{n+1}=b_{n+1}|Q_{n+1}=i+k_n+k_{n+1}+b_n) \times \mathbb{E}[A - A^{(0)} - A^{(1)}|Y=i, A_n=k_n, B_n=b_n, A_{n+1}=k_{n+1}, B_{n+1}=b_{n+1}],$$

and

$$\mathbb{E}[A_n - A_n^{(0)} - A_n^{(1)}|Y=i, A_n=k_n, B_n=b_n, A_{n+1}=k_{n+1}, B_{n+1}=b_{n+1}] = \max\{0, \min\{k_n, i+k_n-b_n-b_{n+1}\}\}.$$

In a similar fashion, we derive $P(W \geq k|Y=i)$ for $k=3,4,\ldots$. Finally, $P(W=0|Y=i) = 1 - P(W \geq 1|Y=i)$.

To complete the derivation, we assume that the system functions in a stationary regime, and we use the full probability formula, where the exchange of the limit and the summation is justified by the dominated convergence theorem:

$$P(W \geq k) = \sum_{i=0}^{\infty} P(Y=i)P(W \geq k|Y=i), \quad k=0,1,\ldots.$$

References

[1] N. Bailey. On queueing processes with bulk service. *Journal of the Royal Statistical Society Series B (Methodological)*, 189(1):80–87, 1954.

[2] A. Braaksma and N. Kortbeek. Integral multidisciplinary rehabilitation planning. *Technical report, University of Twente, submitted to HCMS, available on request*, 2012.

[3] H. Bruneel and I. Wuyts. Analysis of discrete-time multiserver queueing models with constant service times. *Operations Research Letters*, 15(5):231–236, 1994.

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

[5] M. Chaudhry, B. Madill, and G. Briere. Computational analysis of steady-state probabilities of $M|G^{a,b}|1$ and related nonbulk queues. *Queueing systems*, 2(2):93–114, 1987.

[6] C. Chien, Y. Huang, and C. Hu. A hybrid approach of data mining and genetic algorithms for rehabilitation scheduling. *International Journal of Manufacturing Technology and Management*, 16(1-2):76–100, 2009.

[7] C. Chien, F. Tseng, and C. Chen. An evolutionary approach to rehabilitation patient scheduling: A case study. *European Journal of Operational Research*, 189(3):1234–1253, 2008.

[8] D. Conforti, F. Guerriero, and R. Guido. Optimization models for radiotherapy patient scheduling. *4OR: A Quarterly Journal of Operations Research*, 6(3):263–278, 2008.

[9] D. Conforti, F. Guerriero, R. Guido, and M. Veltri. An optimal decision-making approach for the management of radiotherapy patients. *OR Spectrum*, 33(1):123–148, 2011.

[10] European Neuromuscular Centre. *Retrieved February 29, 2012, from: http://www.enmc.org*, 2012.

[11] J. Goddard and M. Tavakoli. Efficiency and welfare implications of managed public sector hospital waiting lists. *European Journal of Operational Research*, 184(2):778–792, 2008.

[12] H. Gold and P. Tran-Gia. Performance analysis of a batch service queue arising out of manufacturing system modelling. *Queueing Systems*, 14(3):413–426, 1993.

[13] L. Green and J. Soares. Computing time-dependent waiting time probabilities in $M(t)|M|s(t)$ queueing systems. *Manufacturing & Service Operations Management*, 9(1):54–61, 2007.

[14] L. Green, J. Soares, J. Giglio, and R. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[15] D. Gupta and B. Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9):800–819, 2008.

[16] P. J. Hulshof, N. Kortbeek, R. J. Boucherie, and E. W. Hans. Taxonomic classification of planning decisions in health care: a review of the state of the art in or/ms. Technical Report 1944, University of Twente, Enschede, June 2011.

[17] P. Joustra, J. de Wit, V. Struben, B. Overbeek, P. Fockens, and S. Elkhuizen. Reducing access times for an endoscopy department by an iterative combination of computer simulation and linear programming. *Health Care Management Science*, 13(1):17–26, 2010.

[18] N. Kortbeek, M. E. Zonderland, R. J. Boucherie, N. Litvak, and E. W. Hans. Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. Memorandum 1968, Department of Applied Mathematics, University of Twente, Enschede, December 2011.

[19] Medline Plus. *Retrieved February 29, 2012, from: http://www.nlm.nih.gov/medlineplus/neuromusculardisorders.html*, 2012.

[20] M. Neuts. A general class of bulk queues with poisson input. *The Annals of Mathematical Statistics*, 38(3):759–770, 1967.

[21] M. Neuts. Queues solvable without rouche's theorem. *Operations Research*, 27(4):767–781, 1979.

[22] S. Ogulata, M. Koyuncu, and E. Karakas. Personnel and patient scheduling in the high demanded hospital services: a case study in the physiotherapy service. *Journal of medical systems*, 32(3):221–228, 2008.

[23] V. Podgorelec and P. Kokol. Genetic algorithm based system for patient scheduling in highly constrained situations. *Journal of medical systems*, 21(6):417–427, 1997.

[24] K. Schimmelpfeng, S. Helber, and S. Kasper. Decision support for rehabilitation hospital scheduling. *OR Spectrum*, pages 1–29, 2010.

[25] A. Turkcan, B. Zeng, and M. Lawley. Chemotherapy operations planning and scheduling. *Optimization Online*, 12(2):1–30, 2010.

[26] Vereniging Spierziekten Nederland (In Dutch). *Retrieved August 05, 2011, from: http://www.vsn.nl/spierziekten*, 2012.

[27] I. Vermeulen, S. Bohte, S. Elkhuizen, H. Lameris, P. Bakker, and H. Poutré. Adaptive resource allocation for efficient patient scheduling. *Artificial intelligence in medicine*, 46(1):67–80, 2009.

[28] W. Winston. *Operations Research: Applications and Algorithms*. Duxbury press, 2003.

[29] D. Worthington. Queueing models for hospital waiting lists. *Journal of the Operational Research Society*, 38(5):413–422, 1987.