Article:

Condotta, A and Shakhlevich, NV (2014) Scheduling patient appointments via multilevel template: a case study in chemotherapy. Operations Research for Health Care, 3 (3). 129 - 144. ISSN 2211-6923

https://doi.org/10.1016/j.orhc.2014.02.002

Scheduling Patient Appointments via Multilevel Template: A Case Study in Chemotherapy

A. Condotta[1], N.V. Shakhlevich [2]

School of Computing, University of Leeds, Leeds, LS2 9JT, United Kingdom

Abstract

This paper studies a multi-criteria optimization problem which appears in the context of booking chemotherapy appointments. The main feature of the model under study is the requirement to book for each patient multiple appointments which should follow a pre-specified multi-day pattern. Each appointment involves several nurse activities which should also follow a pre-specified intra-day pattern. The main objectives are to minimize patients' waiting times and peaks of nurses' workload for an outpatient clinic. Our solution approach is based on the concept of a multi-level template schedule which is generated for a set of artificial patients with typical treatment patterns. There are two stages in template generation: the multi-day stage, which fixes appointment dates for all artificial patients, and the intra-day stage, which fixes for each day appointment starting times and patient allocation to nurses. The running schedule is created by considering actual patients one by one as they arrive to the clinic. Booking appointments for each new patient is performed by assigning appropriate dates and times of the template schedule following the prescribed multi-day and intra-day patterns. Additional rescheduling procedure is used to re-optimize intra-day schedules on a treatment day or shortly beforehand. The key stages of the scheduling process are modeled as integer linear programs and solved using CPLEX solver. We demonstrate the effectiveness of our approach through case-based scenarios derived from a real clinic and discuss the advantages that the multi-level template can bring.

Keywords: patient scheduling, treatment patterns, template schedule, integer linear programming

1. Introduction

With increasing number of cancer patients, chemotherapy departments are under pressure to provide more efficient and qualitatively better service. The UK National Chemotherapy Advisory Group [1] and the Department of Health [2] suggest a redesign of the delivery of chemotherapy treatments and set up new targets in terms of patients' waiting times and improved quality of service. In our paper we show that advanced scheduling algorithms can help in providing timely treatments to larger numbers of patients within shorter time-frames, balancing workloads of medical staff and reducing the costs of treatments. Automated scheduling is also helpful in efficient rescheduling which is often needed due to unpredictable events such as changes in patients treatment plans and clinic resources.

In this paper we study a multi-criteria optimization problem which appears in the context of scheduling chemotherapy appointments. The scenario we consider is typical for many chemotherapy outpatient clinics. In its current form, the problem was formulated by the Institute of Oncology at the St. James's University Hospital in Leeds, U.K.

[1]alessandro@condotta.net

[2]n.shakhlevich@leeds.ac.uk (corresponding author, tel. +44 113 343 5444, fax +44 113 343 5468)

Chemotherapy is an important and widely used therapy to treat cancer. It consists of cyclic administrations of drug mixtures delivered to patients under rigid protocols called *regimens*. In an outpatient clinic, chemotherapy is mainly administered orally or through intravenous systems in a day case unit. The two main characteristics of a regimen are drug combinations and delivery patterns. Once a regimen is prescribed, a patient should visit the clinic at treatment days which are separated by a fixed number of rest days. During a treatment day, a patient undergoes a specific treatment procedure which involves nurse activities separated by time intervals of fixed length.

A regimen is characterized by two types of patterns:

- a *multi-day pattern* given as a sequence of treatment days and rest days in-between and
- an *intra-day pattern* given for each treatment day as a sequence of nurse activities and time-lags in-between.

Typically, nurse activities are related to setting up an intravenous machine, its re-setting for a new drug mixture or a check-up of patient's state. The working day of the clinic is split into 15-minute time slots and this is the main time unit of an intra-day schedule: each nurse activity consists of exactly one 15-minute time slot and the length of any time-lag in-between nurse activities is a multiple of 15 minutes.

Upon arrival to the clinic, a patient is allocated to a nurse for delivering treatments of the prescribed intra-day pattern of the current one-day session. An important clinic requirement, related to the quality and safety of service provision, is to assign a single nurse to perform all treatment activities for a patient on one treatment day. It is, however, acceptable that different nurses treat a patient on different visit days.

Normally a nurse treats several patients during a day. If treatment activities assigned to one nurse lead to a *clash* in the nurse's schedule with more than one activity occurring during the same time slot, a special arrangement is made so that clashing activities are covered by additional nurse(s). It is required that each nurse has a 30-minute break (two time slots) in the middle of the day.

Example 1. *Consider a multi-day schedule for nine patients P1, ..., P9, shown in Fig. 1 (a) with treatment days marked by dark boxes. No treatments are planned for weekend days which are dashed. The fragments of schedules for three days and one nurse are shown in Fig. 1 (b). The first two intra-day schedules have clashing activities resulting in a solution of poor quality since clashing treatments cannot be delivered by a single nurse.*

In order to characterize the effect of clashing activities, we use the following two measures of an intra-day schedule for day d and nurse n: the *number of clashing activities* $\Omega_{d,n}$ and the *maximum clash density* $\Delta_{d,n}$. The value of $\Omega_{d,n}$ is defined as the minimum number of activities which, if removed from the schedule of day d, result in a clash-free schedule for nurse n. The value of $\Delta_{d,n}$ is defined as the maximum number of activities assigned to one nurse in one time slot.

Example 1 (cont.) *The schedule for the second day (d =08/07/2008) has clashing activities in time slots 9:15, 9:30, 12:30, 13:30 and 14:45. The schedule becomes clash free if we remove, e.g., the following activities: four activities of patient P8 assigned to time slots 9:15, 9:30, 13:30, 14:45, two activities of patient P1 assigned to time slots 9:15, 14:45, and one activity of patient P5 assigned to time slot 12:30, which implies that* $\Omega_{d,n} = 7$. *The second characteristic* $\Delta_{d,n} = 3$ *reflects the fact that there are three clashing activities in time slots 9:15 and 14:45, and this is the maximum number of activities assigned to one time slot.*

In the rest of the paper we use the term "*appointment*" to indicate one treatment day of a patient. An appointment is scheduled if its date and starting time are fixed and a nurse is allocated to perform all associated treatment activities on that day. An appointment is partially scheduled if only some of these parameters (day, time, nurse allocation) are fixed.

The main decisions associated with scheduling appointments for one patient are as follows:

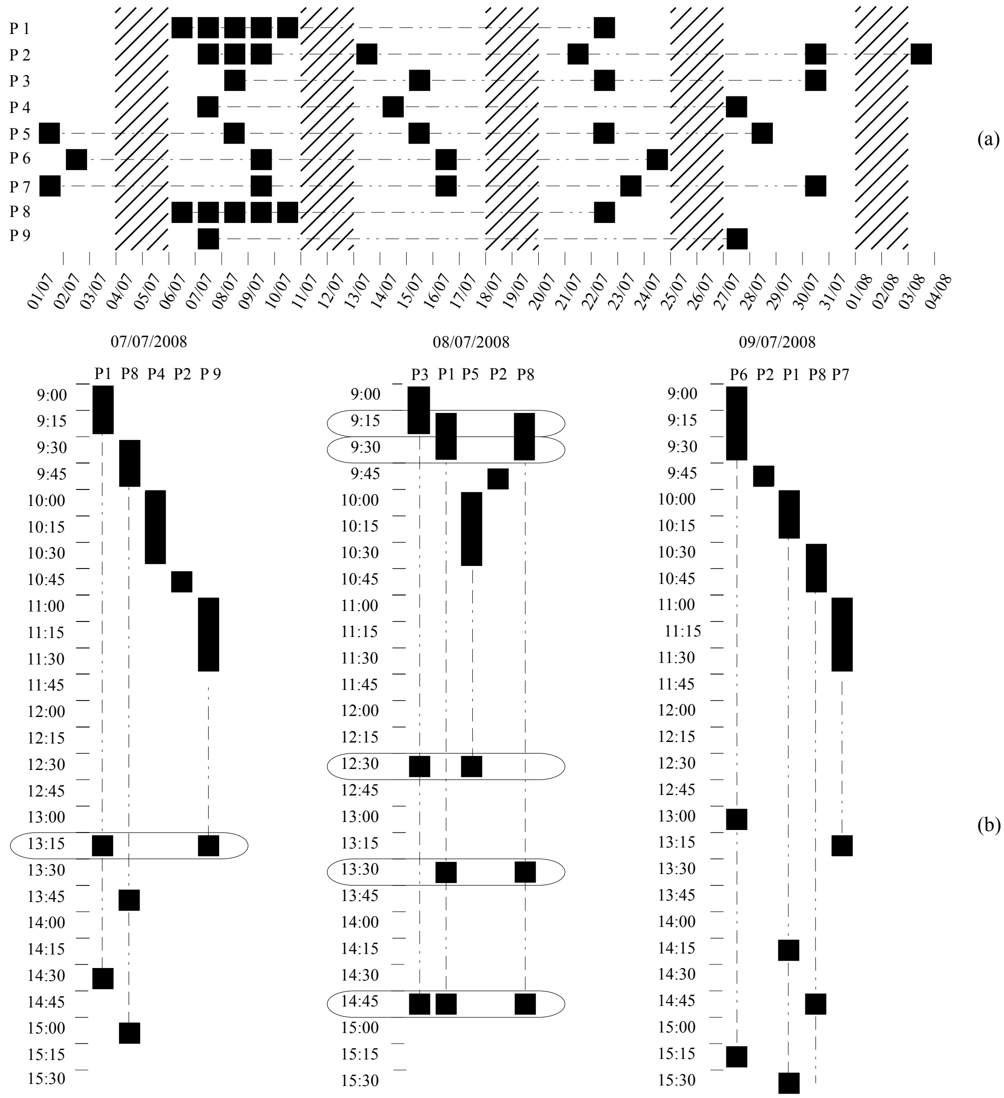

Figure 1: An example of a multi-day schedule (a) and three intra-day schedules (b), each schedule for one nurse

D1: selecting the date of the first appointment, which in fact fixes the dates of all subsequent appointments of the associated multi-day pattern;

D2: selecting the starting time of each appointment; this fixes the time slots for all treatment activities needed for the patient on the visit day;

D3: allocating nurses, one nurse per patient per visit day, to perform all treatment activities for that patient on the day.

The main outcome of the above decision making process is a *multi-day schedule* combined with a series of *intra-day schedules*, one for each day of the time horizon. These schedules fully describe all nurse activities in the clinic.

It should be noticed that each of the above decisions incurs an NP-complete problem even under most simple assumptions. For example, two-component patterns can be modelled as a famous coupled-operation scheduling problem, which has attracted lots of attention of scheduling researchers since the seminal paper by Orman and Potts [23]. In particular, the multi-day problem corresponding to decision D1 with two-component patterns and the assumption that one appointment require a complete day is a special case of the general coupled-operation problem where each appointment is a task to be scheduled on a single machine within a certain deadline. Similarly, the intra-day problem corresponding to decision D2 with two-component patterns to be assigned to one nurse is also a special case of the coupled-operation problem where each activity is a task to be scheduled on a single machine. The NP-completeness of the coupled-operation problem with equal-length operations is proved in [11]. As far as decision D3 is concerned, the NP-completeness of the corresponding problem follows from the NP-completeness of the k-coloring problem, as shown in [10], Chapter 4.

Actual scenarios of chemotherapy scheduling involve complicated patterns with typically more than two activities per pattern; in addition they involve a combination of several metrics which characterize the quality of the combined schedule. The three main metrics are as follows:

F1: the average number of waiting days for all patients, where the number of waiting days for a patient is measured as the delay from a target starting date (predefined by the patient's requirements or by the clinic waiting targets) to the day of the first appointment;

F2: the maximum clash density Δ, which is calculated as the maximum among $\Delta_{d,n}$-values for all nurses $n \in N$ and all days $d \in H$ of the selected time horizon H, i.e. $\Delta = \max\limits_{d \in H, n \in N} \Delta_{d,n}$;

F3: the total number of clashes Ω for all nurses, which is calculated as the sum of the $\Omega_{d,n}$-values for all nurses over the time horizon H, i.e. $\Omega = \sum\limits_{d \in H, n \in N} \Omega_{d,n}$.

The model we propose is based on the requirements gathered in collaboration with the St. James's University Hospital in Leeds, UK. The flexibility of our solution approach makes it possible to introduce various adjustments to deal with an extended version of the model, addressing more complex scenarios. For example, we consider the case where intra-day patterns are the same for different appointment days of a regimen; the general case with different intra-day patterns per regimen could be easily handled by the model. Another possible extension is related to additional requirements of the pharmacy, which supplies drugs for chemotherapy treatments. While some drugs are available off-the-shelf, others need to be prepared in the hospital pharmacy. Since the shelf-life of drugs can be limited, it is often required to produce some drugs just before their administration. Although in our study we mainly focus on metrics F1-F3, the approach we propose is very well suited for further enhancements which can take into account drug preparation.

Our study can be considered as the first attempt to develop an optimization model for scheduling patient appointments which should follow intra-day and multi-day patterns. We explore the benefits of using a template schedule to fix appointment dates and times for patients arriving over time. The *multilevel template schedule* we produce determines possible dates and time slots at which future appointments can be scheduled. The appointments of the template are grouped into levels with a view of producing an efficient running schedule with even nurse workload and the smallest possible number of clashes. The *running schedule* is created by considering arriving patients one by one; for each new patient all required appointments are assigned to appropriate time slots of the template schedule following the prescribed multi-day and intra-day patterns. An additional rescheduling procedure is used to re-optimize each intra-day schedule on a treatment day or shortly beforehand. The key stages of the planning process are modeled as integer linear programs and solved using CPLEX solver.

The paper is organized as follows. We start with the review of relevant sources presented in Section 2. In Section 3 we introduce the concepts of a template schedule and running schedule and give a general overview of our approach. In Sections 4 and 5 we formalize in detail the structure of the template schedule and metrics used to evaluate its quality. In Sections 6 and 7 we describe how the template schedule should be constructed. In particular, Section 6 describes the data generation process and Section 7 formulates two integer linear programs (ILPs): one ILP to schedule appointment dates for future patients and another one to schedule appointment times for each day of the template schedule. In Section 8 we describe how the template schedule is used for creating a running schedule for arriving patients. Section 9 explains how daily rescheduling is done. In Section 10 we evaluate the proposed approach via computational experiments performed on two scenarios typical for a real-world clinic. Finally, conclusions are formulated in Section 11.

2. Literature Review

There are two types of sources related to our study: the papers on scheduling coupled-operation jobs with exact time-lags and the papers on appointment scheduling in health care. The problem with coupled-operation jobs models the simplest case of scheduling multi-day or intra-day patterns when each pattern consists of two components only, see Section 1. The coupled-operation problem is NP-hard even for the case of equal length operations [11]. The literature on heuristics is quite limited (see, e.g., [11, 27]), and the generalization of heuristics proposed for the coupled-operation problem to the case of more complex patterns and several criteria is a nontrivial task.

Appointment scheduling in health care has been studied since the early fifties when Bailey [3] and Lindley [20] published their first results on block appointment systems, using a single queue model to minimize patient waiting times. Since then, the main stream of research focuses on finding appropriate rules for assigning patient appointments to time intervals in order to minimize patients' waiting times, idle time and overtime of physicians or trade-offs between these objectives, see, e.g., [15].

In more recent research, the models are extended by considering additional unexpected events such as no-shows, walk-ins and emergency treatments. A popular approach aimed at reducing effects of those unexpected events is related to overbooking, see, e.g. [30, 33]. A comprehensive survey with a detailed classification of scheduling models for booking outpatient appointments can be found in surveys [6, 18] and in a more recent classification review [19].

In what follows we discuss the papers dealing with scheduling a series of appointments, which is the main feature of our model. Apart from chemotherapy, such problems arise in the context of examination, rehabilitation and radiotherapy treatments.

Examination and *rehabilitation* treatments deal with packages of treatment procedures, see [4, 7, 8, 9, 16, 17, 22, 26]. Intra-day patterns usually reduce to single treatments, while multi-day patterns are characterized by partial precedence constraints: only some treatments require a specific order, others can be sequenced with a large degree of freedom. The time-lags between treatments can also be flexible, contrary to the case of strict chemotherapy patterns. The most popular approach for scheduling examination and rehabilitation appointments is genetic algorithm [8, 9, 26]; there are also examples of construction heuristics [7], local search [16], queueing systems [17] and ILP-based methods [4, 22].

The problem which is most closely related to our problem is scheduling *radiotherapy* appointments. Similar to chemotherapy multi-day patterns, radiotherapy patterns are fixed and cannot be altered. Their structure, however, is quite specific as they typically consist of several consecutive days (the time-lags in-between treatment days are zero). Intra-day patterns are also rather uniform with one treatment per patient per day, see [12, 13, 24, 25]. Typical approaches adopted for scheduling radiotherapy appointments are construction heuristics [24] and local search [25]; the simplified models (e.g., those not taking into account the intra-day scheduling problem, patients' release times or due dates) are addressed via integer linear programming [12, 13].

The authors of the papers listed above emphasize that the subject of scheduling multiple (recurring) appointments is not well studied, whichever the context is: rehabilitation, examination or radiotherapy. In spite of its importance, the problem of scheduling chemotherapy appointments has received even less attention of OR researchers. The majority of publications appear in medical journals; the main focus of those publications is on scheduling strategies adopted by practitioners.

As a rule, scheduling decisions are made by a qualified nurse who selects appointment dates and times based on the knowledge of treatment procedures, resource availability and personal experience. We refer to Turkcan et al. [28] who provide an overview of publications in oncology journals. Here we only mention the paper by Dobish [14] who describes a scheduling strategy based on creating manually a weekly template schedule. The template is used as a basis for scheduling patients arriving over time. Positive feedback of practitioners discussed in [14] and the potential of using a template for scheduling patients in an uncertain environment have inspired our work. Notice that the strategy suggested by [14] is of empirical nature, while in our study the idea of template schedule is developed into a formal optimization model with advanced features.

We are aware of only one publication by Turkcan et al. [28] who consider optimization aspects of delivering chemotherapy treatments. There are a number of points of difference between the model from [28] and the one studied in our paper: intra-day patterns are treated in [28] as contiguous blocks of time rather than sequences of nurse activities separated by idle intervals. Instead of distinguishing within a block between treatment activities and time-lags, the authors introduce a special characteristic of a block which measures the acuity or intensity of treatments of the block, and use that characteristic to restrict nursers' workloads. In our model, we use precise descriptions of intra-day patterns which allows us to make accurate calculations of nurses' workloads and to estimate possible clashes.

The solution approaches are also different. The approach proposed in our paper constructs a dynamic plan for all anticipated appointments of the time horizon; such a plan is then frequently adjusted and tuned in accordance with changes in demand and changes in patients' state of health. All appointment dates and times are scheduled and communicated to a patient at the time a patient is admitted to the clinic. In [28] scheduling is performed at regular times for intervals in-between scheduling sessions. The decisions are made for available patients only, which implies that patients arriving in-between scheduling sessions should always wait for the next session or even longer to get their first appointment. Moreover, appointment times for subsequent visits may not be available all at once as each scheduling session considers a limited time interval until the next session.

In spite of all the differences, both studies, [28] and ours, indicate that the problem of scheduling chemotherapy appointments has multiple constraints, complex objectives and requires further study. To the best of our knowledge the chemotherapy scheduling problem, as defined in this paper, is common to outpatient clinics not only in the U.K., but internationally as well. Although many booking and IT systems are available for the chemotherapy market, their scheduling components are typically not automated and have limited optimization capabilities.

We believe that the approach designed in this paper could be adapted to similar health care settings in the future. Examples of well defined job patterns can be found in nephrology in relation to kidney dialysis treatment performed by a blood cleaning machine supervised by a nurse. With the current tendency to standardize patients' pathways, the scope of application of the proposed methodology can be expanded even further.

3. General Idea of the Solution Approach

In this paper we consider a finite time horizon H in which patients of a set P have to be treated in accordance with their multi-day and intra-day patterns. It is assumed that patients arrive during the first H_a days of time horizon H, which we call the *arrival period*, $H_a \subset H$. Whenever a new patient arrives, the scheduling procedure makes a decision about all appointments for that patient fixing the appointment dates, appointment times and nurse allocation and recording this information in the *running schedule*. The overall length of H is sufficiently large so that the latest patient arriving at the end of H_a can get all required appointments within the remaining part $H \backslash H_a$ of the time horizon.

The continuous arrival of patients is managed by adopting the concept of the rolling time horizon. At some point of the arrival period H_a we abort the scheduling procedure, keeping the current running schedule with appointments fixed so far, and initialize our scheduling procedure for the new time horizon H and new arrival period H_a. Appointments for newly arriving patients are added to the running schedule kept from the previous stage without altering fixed appointments of patients arrived earlier.

The running schedule is created and maintained on the basis of the *multi-level template schedule*, which is pre-calculated prior to the start date of the arrival period H_a using the historical data of the recent patients' arrivals. It contains the appointments for a large number of *artificial patients* and it serves as the basis for fixing the appointments for each newly arriving patient. With careful pre-calculation, we make sure that the template is flexible enough, so that there are sufficient options for arriving patients with different multi-day and intra-day patterns. Moreover, special attention is paid to ensure that the template schedule can potentially produce a high quality running schedule with reduced number of clashing tasks in the nurses' schedules and without violating patients' waiting times limits.

We complement our approach with an additional subroutine of daily rescheduling that allows further improvement of the running schedule.

Formally, the approach we propose consists of the following four stages:

Stage 1. Data Generation for Artificial Patients
(based on the forecast of the arrival rates of new patients);

Stage 2. Template Schedule Generation
(obtained as a solution to the problem of scheduling artificial patients);

Stage 3. Running Schedule: Creating and Maintaining
(based on the template schedule; updated on a daily basis for actual patients arriving over time);

Stage 4. Daily Rescheduling
(to take into account on-the-day changes in nurses' availability and patients' treatment plans).

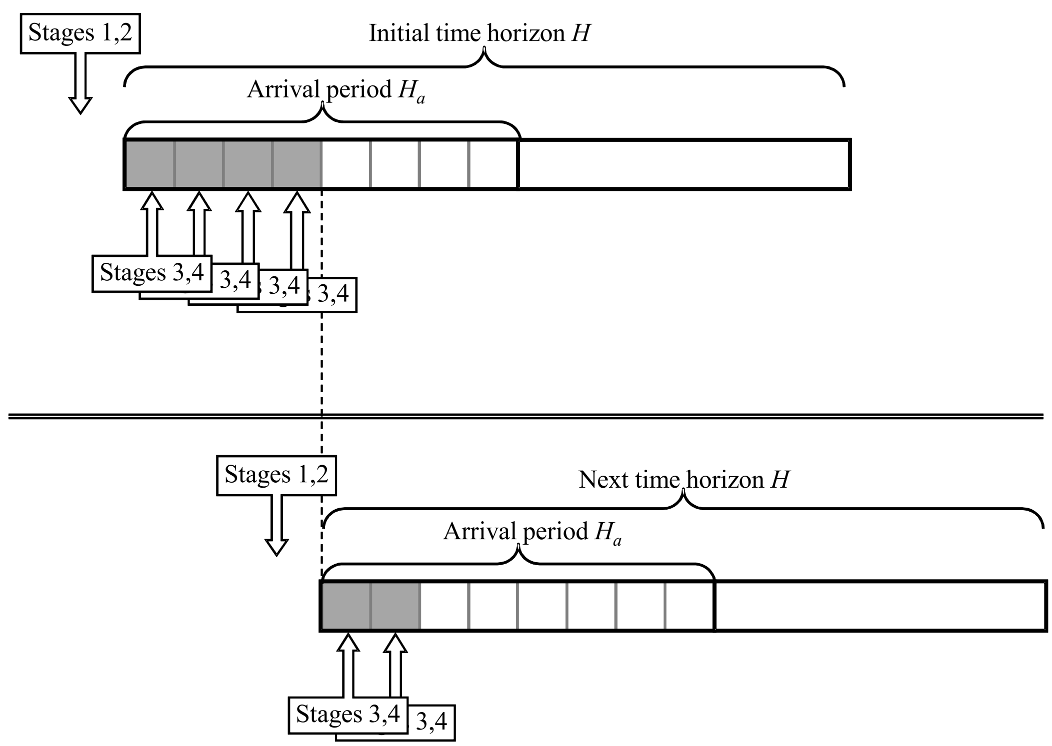

Figure 2: Representation of the 4-stage approach

The application of the described four-stage approach is illustrated in Fig. 2. Stages 1-2 are performed in advance, prior to the first day of the arrival period H_a, and the generated template covers the whole time horizon H. Stages 3-4 are run on a daily basis. After the decision is taken to shift the time horizon, Stages 1-2 are repeated once again: the previous template schedule is fully replaced by a new template calculated on the basis of the most recent information on patient arrival, keeping fixed appointments of actual patients from the previous time horizon unchanged. With the new template schedule for the next time horizon, Stages 3-4 are run again on a daily basis, serving the needs of arriving actual patients.

In what follows we first explain the notion of a multi-level template schedule in more detail (Sections 4-5) and then describe the implementation details for each stage of our approach (Section 6-9).

4. Template Schedule Structure

The main goal of a multi-level template schedule is to create a robust plan for serving patients arriving over time such that the fluctuation in actual demand does not impact too much the quality of the final running schedule. The template schedule determines dates, times and nurse assignment for appointments for a set of artificial patients $P^{\mathcal{A}}$. The number of such patients is the expected demand increased by some factor as explained in detail in Section 6. In brief, the main reason of dealing with a sufficiently large set $P^{\mathcal{A}}$ of artificial patients is to ensure that for each arriving patient the algorithm, which produces a running schedule, will find a set of appointments in the template that match the required multi-day and intra-day patterns. Since the scheduling algorithm, used at the template generation stage, optimizes the quality characteristics of the template, the resulting running schedule, which contains a selection of appointments from the template, inherits those good characteristics of the template.

The template schedule, defined for the time horizon H, can be seen as a combination of

- a multi-day schedule which specifies for each day $d \in H$ the set $P_d^{\mathcal{A}}$ of artificial patients ($P_d^{\mathcal{A}} \subset P^{\mathcal{A}}$) to be treated on that day;
- an intra-day schedule for each day $d \in H$, which specifies for each patient $p \in P_d^{\mathcal{A}}$ the allocated nurse and the appointment starting time.

Due to the overestimated number of artificial patients (introduced in order to provide at Stage 3 good allocation options for all arriving actual patients), intra-day schedules of the template may have a large number of unavoidable clashing activities. Still such a template leads to a successful running schedule at Stages 3-4. This is done by means of the so-called *density sets*, introduced for every intra-day template schedule, which partition artificial appointments in sets according to the number of clashing tasks they generate. When generating the running schedule, the choice of artificial appointments for actual patients is performed on the basis of density sets, aiming to achieve the smallest possible number of clashes in the resulting schedule.

Formally, the appointments of artificial patients of a day $d \in H$ assigned to the same nurse n are grouped into density sets $\mathcal{L}_{d,n}^{(1)}, \mathcal{L}_{d,n}^{(2)}, \ldots, \mathcal{L}_{d,n}^{(K)}$, which satisfy the nested property:

$$\mathcal{L}_{d,n}^{(1)} \subseteq \mathcal{L}_{d,n}^{(2)} \subseteq \cdots \subseteq \mathcal{L}_{d,n}^{(K)}.$$

The upper index k, $1 \le k \le K$, denotes a *density level* which defines the maximum clash density $\Delta_{d,n}$ of appointments $\mathcal{L}_{d,n}^{(k)}$, considered separately. Appointments of the set $\mathcal{L}_{d,n}^{(1)}$ do not have clashing activities, i.e., if the appointments not belonging to $\mathcal{L}_{d,n}^{(1)}$ are removed, the resulting schedule is clash free: $\Delta_{d,n} = 1$. For any k, $1 \le k \le K$, a partial schedule consisting of appointments from the set $\mathcal{L}_{d,n}^{(k)}$ may have clashes of the maximum density $\Delta_{d,n} = k$. Although the total number of density levels K can be, in the worst case, as large as the number of artificial patients scheduled on one day, in our experiments on real world data the template schedule has, as a rule, no more than 5 density levels, so that $K \le 5$ (see Section 10).

Example 1 (cont.) *Consider Fig. 1 as an example of a template schedule. Each of the one-day schedules of the template, generated for one nurse, should provide information on density sets. A possible partition into the density sets for the second one-day schedule corresponding to day* $d =$ 08/07/2008*, can be as follows:*

$$
\begin{aligned}
\mathcal{L}_{d,n}^{(1)} &= \{P1,\ P2,\ P5\}\,,\\
\mathcal{L}_{d,n}^{(2)} &= \mathcal{L}_{d,n}^{(1)} \cup \{P8\} = \{P1,\ P2,\ P5,\ P8\}\,,\\
\mathcal{L}_{d,n}^{(3)} &= \mathcal{L}_{d,n}^{(2)} \cup \{P3\} = \{P1,\ P2,\ P5,\ P8,\ P3\}\,.
\end{aligned}
$$

Let R denote the set of regimens. We distinguish between the density sets for different regimens: for a regimen $r \in R$, the set of appointments belonging to a density set $\mathcal{L}_{d,n}^{(k)}$ is denoted by $\mathcal{L}_{r,d,n}^{(k)}$. In the above example, if patients P1 and P8 have the same regimen r_1, while the regimens of patients P2, P3, P5 are r_2, r_3, r_4, respectively, then each of the sets $\mathcal{L}_{r_i,d,n}^{(k)}$ is as follows:

$$
\begin{array}{lllll}
\mathcal{L}_{d,n}^{(1)}: & \mathcal{L}_{r_1,d,n}^{(1)} = \{\text{P1}\}\,, & \mathcal{L}_{r_2,d,n}^{(1)} = \{\text{P2}\}\,, & \mathcal{L}_{r_3,d,n}^{(1)} = \emptyset, & \mathcal{L}_{r_4,d,n}^{(1)} = \{\text{P5}\}\,,\\
\mathcal{L}_{d,n}^{(2)}: & \mathcal{L}_{r_1,d,n}^{(2)} = \{\text{P1}, \text{P8}\}\,, & \mathcal{L}_{r_2,d,n}^{(2)} = \{\text{P2}\}\,, & \mathcal{L}_{r_3,d,n}^{(2)} = \emptyset, & \mathcal{L}_{r_4,d,n}^{(2)} = \{\text{P5}\}\,,\\
\mathcal{L}_{d,n}^{(3)}: & \mathcal{L}_{r_1,d,n}^{(3)} = \{\text{P1},\ \text{P8}\}\,, & \mathcal{L}_{r_2,d,n}^{(3)} = \{\text{P2}\}\,, & \mathcal{L}_{r_3,d,n}^{(3)} = \{\text{P3}\}\,, & \mathcal{L}_{r_4,d,n}^{(3)} = \{\text{P5}\}\,.
\end{array}
$$

As we show in Sections 7-8, density levels assigned to appointments of the template schedule in Stage 2, play a fundamental role in Stage 3.

5. Quality Metrics of the Template Schedule

The overall performance of our four-stage solution approach is measured in terms of the quality of the running schedule with respect to the objectives F1, F2, F3 introduced in Section 1 as the average number of waiting days, the maximum clash density, and the total number of clashes. In order to achieve a successful running schedule, we propose a number of metrics for the template schedule: two metrics U and V characterize the template schedule at the multi-day level and three metrics X_d, Y_d and Z_d characterize the template schedule at the intra-day level for each day d of time horizon H.

The first metric U specifies the *maximum daily workload excess* of the clinic for the whole template schedule defined over the time horizon H. This metric considers for each day $d \in H$ the difference between the clinic workload W_d and its daily capacity C_d. The workload W_d is the total number of 15-minute nurse activities which appear in the intra-day schedule of the template. It can be calculated by counting for each nurse n the number $W_{n,d}$ of activities performed by that nurse on day d and summing up those values for all nurses $n \in N_d$ (here N_d is the set of nurses on day d):

$$W_d = \sum_{n \in N_d} W_{n,d}.$$

The daily capacity C_d is the maximum number of activities which can be performed by available nurses:

$$C_d = \sum_{n \in N_d} |H_{n,d}|\,,$$

where $H_{n,d}$ specifies working time of nurse n on day d given as a set of 15-minute time slots when the nurse is available.

With an overestimated number of artificial patients and a limited number of nurses, some days of the template schedule may have activities which cannot be covered by available nurses, so that $W_d > C_d$. Then the workload excess of day d is $\max\{W_d - C_d, 0\}$. The maximum daily workload excess U for the whole template schedule is defined for time horizon H as

$$U = \max_{d \in H} \{\max\{W_d - C_d, 0\}\}\,. \tag{1}$$

It is desirable to keep the workload excess as small as possible distributing it evenly over time horizon H.

Stages 3 and 4 of our approach use two additional characteristics related to the daily workload excess: the *relative workload* $\hat{W}_{n,d}$ of nurse n on day d, which quantifies the proportion of time the nurse delivers treatments to patients:

$$\hat{W}_{n,d} = \frac{W_{n,d}}{|H_{n,d}|}, \tag{2}$$

and the *average relative workload* $\hat{W}_d$ of all nurses working on day d:

$$\hat{W}_d = \frac{\sum_{n \in N_d} W_{n,d}}{\sum_{n \in N_d} |H_{n,d}|} = \frac{W_d}{C_d}. \tag{3}$$

Characteristic $\hat{W}_{n,d}$ is used at Stage 4 when daily rescheduling is done aimed at balancing the workloads of different nurses. Characteristic $\hat{W}_d$ is used at Stage 3 when the best possible options are selected from the template schedule to assign appointments of actual patients keeping the maximum relative daily workload as small as possible and evenly distributed over time horizon H.

To define metric V, we calculate for each artificial patient $p \in P^{\mathcal{A}}$ the deviation $|d_p - t_p|$ of the date of the first appointment d_p from a given target day t_p for that patient, and take the average of these values:

$$V = \frac{1}{|P^{\mathcal{A}}|} \sum_{p \in P^{\mathcal{A}}} |d_p - t_p| \,. \tag{4}$$

Observe that the definition of V includes the case in which the first visit day d_p of an artificial patient p precedes the target day t_p.

We now turn to the one-day metrics X_d, Y_d and Z_d which characterize the quality of an intra-day schedule of day d of the template and which are closely related to the notion of a density set introduced in Section 3. Each metric X_d, Y_d and Z_d is in fact a collection of metrics $X_d^{(k)}$, $Y_d^{(k)}$ and $Z_d^{(k)}$, respectively, defined for density levels $k = 1, 2, \ldots, K$.

Consider an intra-day schedule for day $d \in H$, with artificial patients $P_d^{\mathcal{A}}$ treated on that day, which are partitioned into subsets $P_{r,d,n}^{\mathcal{A}}$ depending on regimen $r \in R$, and allocated nurse $n \in N_d$ (here R is the set of all regimens and N_d is the set of nurses on day d). In the intra-day schedule of nurse n, appointments of patients $P_{r,d,n}^{\mathcal{A}}$ are scheduled and allocated to sets $\mathcal{L}_{r,d,n}^{(k)}$, $1 \le k \le K$, belonging to the density set $\mathcal{L}_{d,n}^{(k)}$, see Section 3 for the definition.

For the partial intra-day schedule consisting of appointments from $\mathcal{L}_{d,n}^{(k)}$ we define $\Omega_{d,n}^{(k)}$ as the total number of 15-minute activities, which, if removed from that partial schedule, result in a clash-free schedule.

Using the above notations, we formally introduce three metrics $X_d^{(k)}$, $Y_d^{(k)}$ and $Z_d^{(k)}$ for a density level k and then provide their interpretation and justification. Those metrics are then optimized lexicographically, as explained in Section 7.

The metrics are defined as

$$\begin{aligned} X_d^{(k)} &= \min_{r \in R, n \in N_d} \left\{ \frac{|\mathcal{L}_{r,d,n}^{(k)}|}{|P_{r,d,n}^{\mathcal{A}}|} \right\} && \text{(to be maximized)}, \\ Y_d^{(k)} &= \sum_{r \in R} \alpha_{r,d} \sum_{n \in N_d} \left| \mathcal{L}_{r,d,n}^{(k)} \right| && \text{(to be maximized)}, \\ Z_d^{(k)} &= \sum_{n \in N_d} \Omega_{d,n}^{(k)} && \text{(to be minimized)}, \end{aligned} \tag{5}$$

where parameter $\alpha_{r,d}$ is an additional weight characteristic defined empirically.

Metric $X_d^{(k)}$ measures the proportion of artificial appointments of different regimens allocated to $\mathcal{L}_{d,n}^{(k)}$. Maximizing $X_d^{(k)}$ ensures a fair representation of appointments of regimens in density set $\mathcal{L}_{d,n}^{(k)}$, so that in the resulting template schedule every regimen has appointments in $\mathcal{L}_{d,n}^{(k)}$ and no regimen is overlooked in favor of another regimen.

Metric $Y_d^{(k)}$ counts the total number of appointments allocated to density set $\mathcal{L}_{d,n}^{(k)}$, in the case of unit weights $\alpha_{r,d}$, or the weighted number of appointments allocated to that density set, otherwise. In the weighted version of metric $Y_d^{(k)}$, the most frequent regimens r have higher weights, so that maximizing $Y_d^{(k)}$ is aimed at allocating as many appointments as possible to the density set of level k giving preference to the appointments of the most frequent regimens. In our experiments, we set

$$\alpha_{r,d} = \left|P_{r,d,n}^{\mathcal{A}}\right|. \tag{6}$$

Finally, metric $Z_d^{(k)}$ measures the overall number of clashing tasks determined by appointments in $\mathcal{L}_{d,n}^{(k)}$. Clearly, that value should be as small as possible.

If several intra-day schedules are compared in terms of metrics (5), then in the first place, we give preference to a schedule with the highest proportion $X_d^{(1)}$ of appointments assigned to density level $k = 1$ for each regimen $r \in R$ and each nurse $n \in N_d$. Among the schedules with the same value $X_d^{(1)}$, we give preference to those schedules having the largest total weighted number $Y_d^{(1)}$ of appointments allocated to density level $k = 1$. If there are still several schedules equivalent in terms of $X_d^{(1)}$ and $Y_d^{(1)}$, the preferred schedule has the smallest number $Z_d^{(1)}$ of clashing activities. The comparison then continues for density level $k = 2$, considering metrics $X_d^{(2)}$, $Y_d^{(2)}$ and $Z_d^{(2)}$ in this order. The similar approach is applied for higher density levels $k = 3, \ldots, K$.

The order of consideration of density levels starting from $k = 1$, proceeding to $k = 2$ and so on up to $k = K$, is in agreement with the procedure used in Stage 3. The latter procedure explores for each actual patient density level $k = 1$ in the first place, then density level $k = 2$, etc., in order to ensure that matching appointments are found in the lowest possible density level of the template.

Example 2. *Consider 9 artificial patients P1, P2, P3, P5, P8, P10, P11, P12 and P13 that should be treated on day d by nurse n. The nurse working day starts at 9:00, finishes at 15:45 and consists of 15-minute time intervals numbered from 1 to 27. Patients' regimens and intra-day patterns are given in Table 1. An intra-day pattern is specified as a sequence of time slots, numbered from 1, which require nurse actions. If an appointment is scheduled to start in time slot $t = 1$ (at 9:00), then the intra-day pattern incurs nurse activities in time slots listed in the pattern; if an appointment is scheduled to start at time $t > 1$, the intra-day pattern incurs nurse activities in time slots listed in the pattern incremented by $t - 1$.*

Appointment	Regimen	Intraday Pattern	$\lvert P_{r,d}^{\mathcal{A}}\rvert$
P1 P8	r_1	$1, 2, 18, 23$	2
P2	r_2	1	1
P3	r_3	1,2,15,24	1
P5 P10 P11	r_4	1,2,3,11	3
P12 P13	r_5	1,2	2

Table 1: Input data for the template schedule for day d

The workload excess $\max\{W_d - C_d, 0\}$ *of the selected day d is 2 since all activities require $W_d = 29$ time slots while the nurse can only cover $C_d = 27$ time slots.*

Consider two template schedules S_1 and S_2 given by Table 2 and graphically represented as the first two schedules in Fig. 3. There are two density sets for each schedule $\mathcal{L}_{d,n}^{(1)}$ and $\mathcal{L}_{d,n}^{(2)}$ which are specified in Table 2 and marked in Fig. 3.

S_1	P1	P3	P8	P12	P10	P2	P11	P5	P13
Starting time-slot	1	3	5	7	9	12	14	15	20
$\mathcal{L}^{(1)}_{d,n}$	*	*	*	*	*	*	*		*
$\mathcal{L}^{(2)}_{d,n}$	*	*	*	*	*	*	*	*	*

S_2	P1	P3	P8	P2	P5	P11	P10	P13	P12
Starting time-slot	1	3	5	8	10	11	14	18	22
$\mathcal{L}^{(1)}_{d,n}$	*	*	*	*	*		*		
$\mathcal{L}^{(2)}_{d,n}$	*	*	*	*	*	*	*	*	*

Table 2: Appointment starting times for two template schedules S_1 and S_2 and their density sets

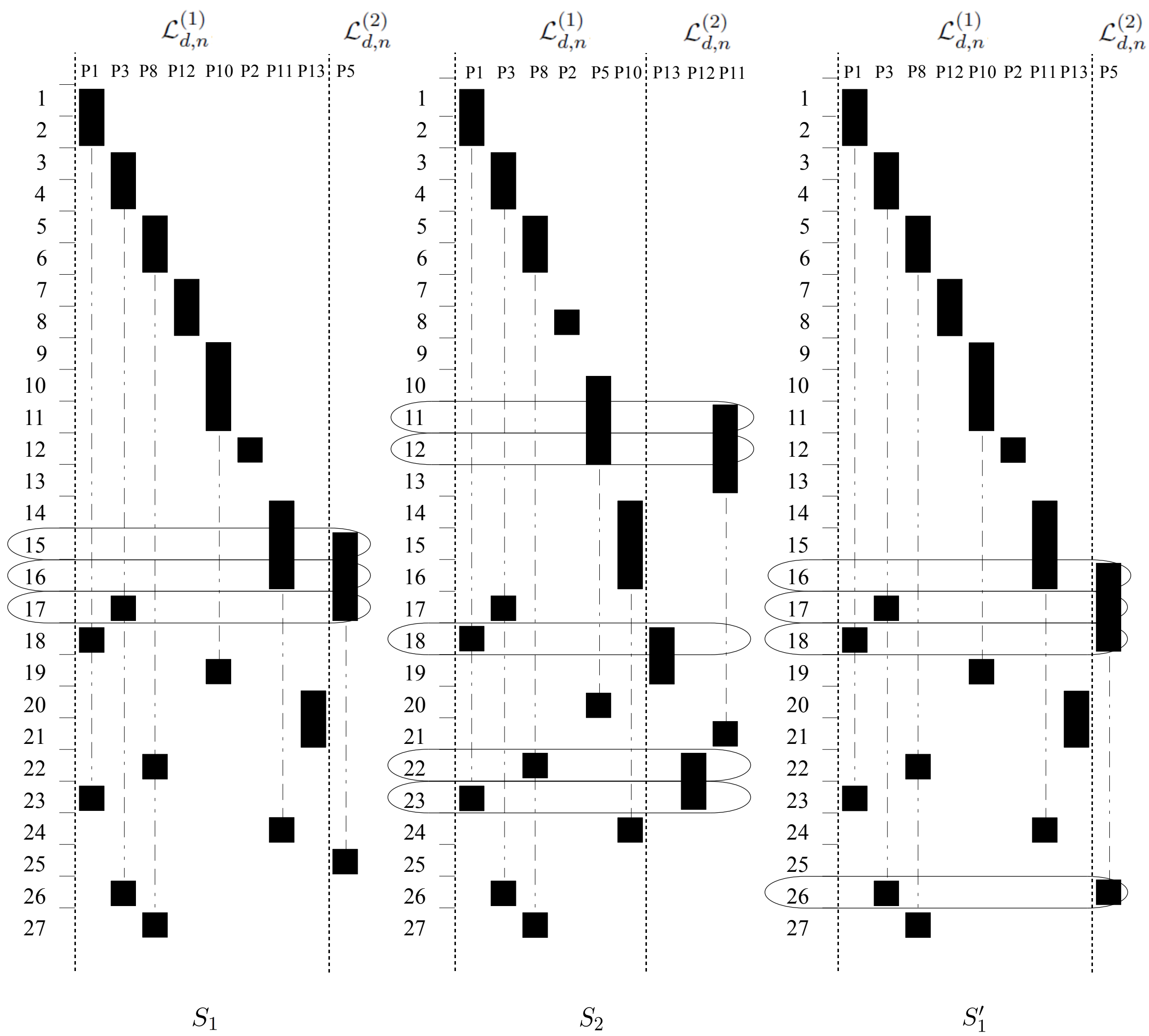

Figure 3: Template schedules S_1 and S_2 given by Table 2 and schedule S'_1 obtained by modifying S_1

The maximum clash density of both schedule S_1 and S_2 is $\Delta_{d,n} = 2$, while the number of clashing tasks is $\Omega_{d,n} = 3$ for S_1 and $\Omega_{d,n} = 5$ for S_2. The values of all metrics for schedules S_1 and S_2 are shown in Table 3, where metric $Y_d^{(k)}$ is calculated with weights $\alpha_{r,d}$ set equal to $|P^{\mathcal{A}}_{r,d,n}|$.

According to metric $X_d^{(1)}$, schedule S_1 is preferred to schedule S_2 since the proportion of appointments allocated to density level $k = 1$ is 0.67 for S_1 and 0 for S_2.

	level k	$X_d^{(k)}$	$Y_d^{(k)}$	$Z_d^{(k)}$
S_1	$k = 1$	0.67	16	0
	$k = 2$	1	19	3
S_2	$k = 1$	0	12	0
	$k = 2$	1	19	5

Table 3: Performance metrics of template schedules S_1 and S_2

Consider now schedule S_1' which is a modification of schedule S_1, where the appointment for patient P5 starts in time slot 16, as shown in Fig. 3. The density sets $\mathcal{L}_{r,d,n}^{(k)}$ are the same for schedules S_1' and S_1, so that S_1 and S_1' are equivalent in terms of metrics $X_d^{(1)}$ and $Y_d^{(1)}$. Still schedule S_1 is preferable in comparison with S_1' due to the metric $Z_d^{(1)}$, as the overall number of clashing activities in density set $\mathcal{L}_{d,n}^{(2)}$ is 3 for schedule S_1 and 4 for schedule S_1'.

6. Data Generation for Artificial Patients

As described in Section 4, a template schedule is generated for a large set of artificial patients $P^{\mathcal{A}}$. The number of artificial patients is always an overestimate in comparison with the expected demand, but the proportion of patients with different regimens and their arrival rates are maintained in agreement with the recent historical data.

Consider the arrival period H_a, when actual patients are due to arrive, and the template generation stage prior to H_a. First we forecast the arrival rate of patients at the clinic for period H_a. The output of this stage is the expected number of patients λ_r (for each regimen $r \in R$) arriving within each week of the arrival period H_a and the distribution function characterizing patient arrival over time. In this paper we focus mainly on scheduling procedures rather than on forecasting techniques. In line with the traditional approach often adopted in similar scenarios, we assume that arrival of patients with a regimen r is a Poisson process with mean and variance equal to λ_r, see, e.g., [5, 6, 21, 29, 31, 32]. The value of λ_r is estimated using the recorded historical data under the assumption that patient arrival rate for the new arrival period H_a does not substantially differ from the past.

The number of artificial patients $\left|P_r^{\mathcal{A}}\right|$ with regimen r for the arrival period consisting of $|H_a|$ working days is defined as

$$|P_r^{\mathcal{A}}| = \xi \lambda_r \times \frac{|H_a|}{5},$$

where $\frac{|H_a|}{5}$ represents the number of weeks in H_a (assuming 5 working days per week) and $\xi \geq 1$ is an overestimation rate, which value is determined empirically. The main purpose is to ensure that for each actual arriving patient considered in Stage 3 there can be found a sufficient number of matching artificial patients with the same regimen to select from, so that the incurred waiting time and the number of clashes are within acceptable limits.

In our experiments the initial value of ξ is derived by scaling the total amount of time needed by the appointments to match as close as possible the capacity of the clinic, ignoring intra-day and multi-day patterns. The latter assumption leads to a template with excessive number of artificial appointments when considering intra-day and multi-day patterns. Further tuning of parameter ξ may be needed in order to find the right level of overestimation for the template to be successful (for a successful template, assigning actual appointments at Stage 3 leads to best results).

An additional adjustment is made for those regimens which happen rarely. In particular, if $\xi \lambda_r < 1$ for a regimen r, then we assign $|P_r^{\mathcal{A}}|$ a higher value:

$$|P_r^{\mathcal{A}}| = \frac{|H_a|}{5},$$

i.e., we force at least one artificial patient with regimen r to be scheduled each week. Such an adjustment ensures that at Stage 3, when an actual patient with a rare regimen r arrives, the waiting time will not be too large, whichever week the patient arrives.

Having defined the number of patients $|P_r^{\mathcal{A}}|$ for each regimen $r \in R$, we generate for each artificial patient $p \in P_r^{\mathcal{A}}$ a target day $t_p \in H_a$ of the first appointment. The values t_p, $p \in P_r^{\mathcal{A}}$, are uniformly distributed in H_a. These values become the main input for generating the template schedule in Stage 2. In the resulting schedule, the day d_p of the first appointment of patient p is selected as close as possible to t_p and either option $d_p \geq t_p$ or $d_p < t_p$ is acceptable.

7. Template Schedule Generation

A template schedule is obtained as a solution to the problem of scheduling appointments of artificial patients $P^{\mathcal{A}}$ over time horizon H. The integrated problem of constructing multi-day and intra-day schedules simultaneously appears to be cumbersome due to its size and the complex combination of multi-day and intra-day patterns. Recall that the problem of scheduling even simplest patterns consisting of two unit-size activities is already NP-complete [11]. On the other hand, it can be naturally decomposed into the following subproblems:

- one subproblem of generating a multi-day schedule for the whole time horizon H;
- $|H|$ subproblems of generating intra-day schedules for each day $d \in H$.

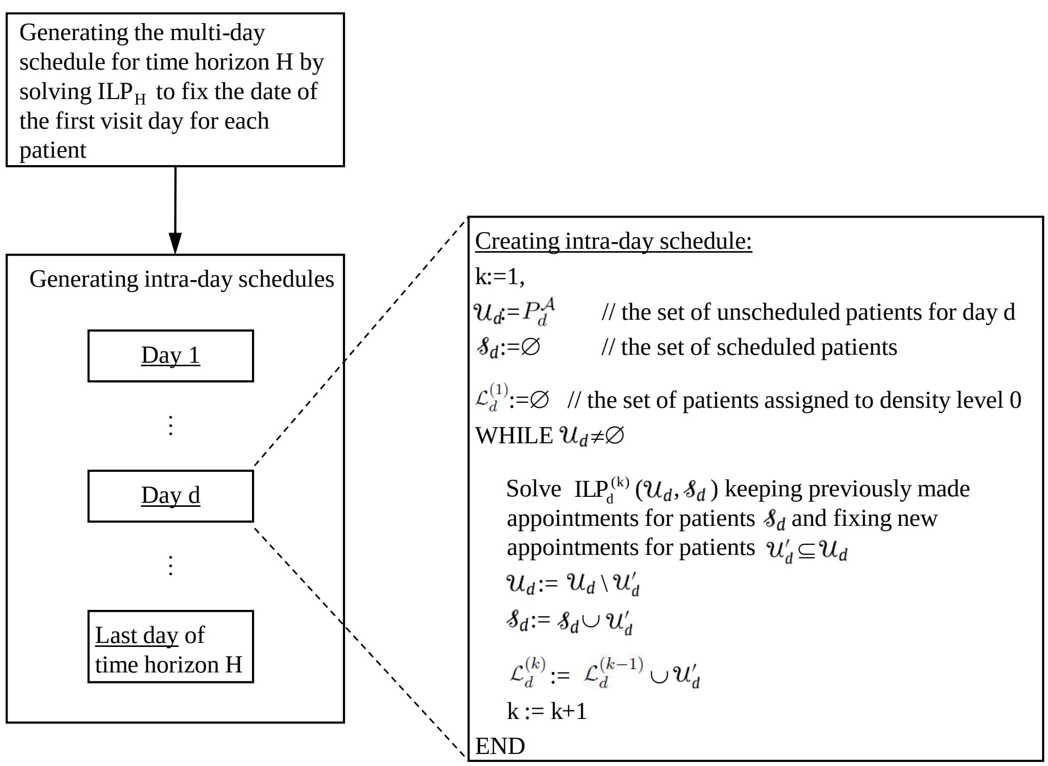

Figure 4: Subproblems and algorithms for generating the template schedule

The multi-day subproblem is formulated as an integer linear program ILP_H defined over time horizon H with time-indexed decision variables

$$x_{p,s} = \begin{cases} 1, & \text{if the first appointment of artificial patient } p \text{ is scheduled on day } s, \\ 0, & \text{otherwise.} \end{cases}$$

For each combination of $s \in H$ and $d \in H$, we define the set of artificial patients $\mathcal{P}^{\mathcal{A}}_{s,d}$ having an appointment on day d if the first visit-day is s, $s \leq d$; in addition for each patient $p \in P^{\mathcal{A}}$ we define constant $w_{p,s,d}$ representing the number of nurse activities necessary to treat that patient on day d if the first visit day is s. Constants $w_{p,s,d}$ are determined by the information given by multi-day and intra-day patterns.

Then metrics U and V defined by (1) and (4) can be represented in the form:

$$
\begin{aligned}
U &= \max_{d \in H} \left\{ \max \left\{ \sum_{s \in H} \sum_{p \in \mathcal{P}^{\mathcal{A}}_{s,d}} (w_{p,s,d} x_{p,s}) - C_d, 0 \right\} \right\}, \\
V &= \frac{1}{|P^{\mathcal{A}}|} \sum_{p \in P^{\mathcal{A}}} \sum_{s \in H} |s - t_p|\, x_{p,s}.
\end{aligned}
$$

We intend to find a solution with the smallest possible value of V among the solutions with the smallest possible value of U. Denoting the lexicographical minimization of U and V by lex$[\min U,\ \min V]$, we formulate the integer linear program ILP_H for the multi-day problem as follows:

$$
\begin{aligned}
ILP_H: \quad & \text{lex} \quad [\min U,\ \min V] \\
\text{s.t.} \quad & U \geq \sum_{s \in H} \sum_{p \in \mathcal{P}^{\mathcal{A}}_{s,d}} w_{p,s,d} x_{p,s} - C_d, && d \in H, \\
& U \geq 0, \\
& V = \tfrac{1}{|P^{\mathcal{A}}|} \sum_{p \in P^{\mathcal{A}}} \sum_{s \in H} |s - t_p|\, x_{ps}, \\
& \sum_{s \in H} x_{p,s} = 1, && p \in P^{\mathcal{A}}, \\
& x_{p,s} \in \{0, 1\}, && p \in P^{\mathcal{A}}, \quad s \in H.
\end{aligned}
$$

Having found the solution to problem ILP_H, we generate input data for intra-day problems: for each day $d \in H$ and each regimen r, we construct the set of artificial patients $P^{\mathcal{A}}_{r,d}$ which should visit the clinic on day d and define $P^{\mathcal{A}}_d = \cup_{r \in R} P^{\mathcal{A}}_{r,d}$.

The intra-day problem consists in selecting the starting times for all appointments of patients $P^{\mathcal{A}}_d$ on day d and finding nurse allocation. The suggested approach considers a series of integer linear programs $ILP^{(1)}_d$, $ILP^{(2)}_d$, $\ldots$, $ILP^{(K)}_d$, as shown in Fig. 4. During the solution process the set of artificial patients $P^{\mathcal{A}}_d$ is partitioned into the subsets $\mathcal{S}_d$ and $\mathcal{U}_d$, which represent the sets of scheduled and unscheduled patients, respectively. Initially $\mathcal{U}_d = P^{\mathcal{A}}_d$ and $\mathcal{S}_d = \emptyset$.

The solution is found iteratively starting from density level $k = 1$. The algorithm fixes the starting times for patients $\mathcal{U}'_d \subseteq \mathcal{U}_d$ by optimizing functions $X^{(k)}_d$, $Y^{(k)}_d$ and $Z^{(k)}_d$; the set $\mathcal{U}_d$ is then updated by moving patients $\mathcal{U}'_d$ to the set of scheduled patients $\mathcal{S}_d$. The corresponding appointments form the density set $\mathcal{L}^{(1)}_d$. Proceeding to density level $k = 2$ with updated sets $\mathcal{U}_d$ and $\mathcal{S}_d$, the algorithm keeps previously fixed appointments and finds the starting times for new patients which are then added to $\mathcal{S}_d$. The density set $\mathcal{L}^{(2)}_d$ is now formed as the union of all previously scheduled appointments $\mathcal{L}^{(1)}_d$ and the appointments scheduled for $k = 2$. The subsequent density sets are considered in a similar fashion until the appointments of all patients are scheduled and $\mathcal{U}_d = \emptyset$.

For each density level k, the intra-day subproblem for day $d \in H$ is formulated as an integer linear program $ILP^{(k)}_d$ with variables

$$
x_{p,s,n,d} = \begin{cases} 1, & \text{if artificial patient } p \in P^{\mathcal{A}}_{r,d} \text{ is assigned to nurse} \\ & n \in N_d \text{ with the first treatment activity in time slot } s, \\ 0, & \text{otherwise.} \end{cases}
$$

Initially, $k = 1$ and all x-variables are free. In subsequent iterations k, $k > 1$, the set of scheduled patients $\mathcal{S}_d$ is non-empty and for $p \in \mathcal{S}_d$, the patient's starting time s_p and allocated nurse n_p are fixed, which implies that the value of the corresponding x-variable is also fixed, $x_{p,s_p,n_p,d} = 1$.

Finding now the expressions for the first two metrics $X_d^{(k)}$ and $Y_d^{(k)}$ from (5) is straightforward:

$$\begin{aligned} X_d^{(k)} &= \min_{r\in R}\left\{\frac{1}{|P_{r,d}^{\mathcal{A}}|}\sum_{p\in P_{r,d}^{\mathcal{A}}}\sum_{n\in N_d}\sum_{s\in H_{n,d}} x_{p,s,n,d}\right\},\\ Y_d^{(k)} &= \sum_{r\in R}\left(\alpha_{r,d}\sum_{p\in P_{r,d}^{\mathcal{A}}}\sum_{n\in N_d}\sum_{s\in H_{n,d}} x_{p,s,n,d}\right),\end{aligned}$$

where $\alpha_{r,d}$ is a parameter defined by (6).

In order to calculate the third metric $Z_d^{(k)}$, we introduce auxiliary variables $y_{i,n,d}$ for each nurse $n \in N_d$ and each time slot $i \in H_{n,d}$ of the nurse's available time slots of day d. The smallest value of $y_{i,n,d}$ is defined as 1 and it represents either of the following two situations which do not incur any penalties:

- nurse n is free in time slot i;
- there is exactly one activity assigned to nurse n in time slot i.

Any greater value $y_{i,n,d} > 1$ represents the number of activities assigned to nurse n in time slot i.

Using the variables $y_{i,n,d}$, the third metric $Z_d^{(k)}$ can be defined as

$$Z_d^{(k)} = \sum_{n\in N_d}\sum_{i\in H_{n,d}} (y_{i,n,d}-1)\,.$$

For calculating $y_{i,n,d}$, we define the set of patients $P_{s,i,d}^{\mathcal{A}}$ having an activity to be performed in time slot i if the appointment starts at time s, and we set

$$y_{i,n,d} \geq \sum_{s=1}^{\tau_{p,n,d}}\sum_{p\in P_{s,i,d}^{\mathcal{A}}} x_{p,s,n,d}$$

with an additional constraint

$$1 \leq y_{i,n,d} \leq k.$$

Here the constant $\tau_{p,n,d}$ defines the latest time slot when the intra-day pattern of patient p may start so that all activities of the pattern can be performed by nurse n within the working hours.

As a result, we obtain the following integer linear program $ILP_d^{(k)}$:

$$ILP_d^{(k)}(\mathcal{U}_d,\mathcal{S}_d):\quad \operatorname{lex}\left[\max X_d^{(k)},\ \max Y_d^{(k)},\ \min Z_d^{(k)}\right]$$

$$\begin{array}{lll}
\text{s.t.} & X_d^{(k)} \leq \frac{1}{|P_{r,d}^{\mathcal{A}}|}\sum_{p\in P_{r,d}^{\mathcal{A}}}\sum_{n\in N_d}\sum_{s\in H_{n,d}} x_{p,s,n,d}, & r\in R,\\
& Y_d^{(k)} = \sum_{r\in R}\left(\alpha_{r,d}\sum_{p\in P_{r,d}^{\mathcal{A}}}\sum_{n\in N_d}\sum_{s\in H_{n,d}} x_{p,s,n,d}\right), & \\
& Z_d^{(k)} = \sum_{n\in N_d}\sum_{i\in H_{n,d}}(y_{i,n,d}-1), & \\
& y_{i,n,d} \geq \sum_{s=1}^{\tau_{p,n,d}}\sum_{p\in P_{s,i,d}^{\mathcal{A}}} x_{p,s,n,d}, & n\in N_d,\quad i\in H_{n,d},\\
& 1\leq y_{i,n,d}\leq k, & n\in N_d,\quad i\in H_{n,d},\\
& x_{p,s_p,n_p,d}=1, & p\in\mathcal{S}_d,\\
& \sum_{n\in N_d}\sum_{s\in H_{n,d}} x_{p,s,n,d}\leq 1, & p\in P_d^{\mathcal{A}},\\
& x_{p,s,n,d}\in\{0,1\}, & p\in P_d^{\mathcal{A}},\quad s\in H_{n,d},\quad n\in N_d.
\end{array}$$

After a solution to problem $ILP_d^{(k)}(\mathcal{U}_d, \mathcal{S}_d)$ is found, the sets of scheduled and unscheduled patients $\mathcal{S}_d$ and $\mathcal{U}_d$ are updated so that $\mathcal{U}_d = \mathcal{U}_d \setminus \mathcal{U}'_d$ and $\mathcal{S}_d = \mathcal{S}_d \cup \mathcal{U}'_d$. A patient p is scheduled, if there exists a combination of s and n such that $x_{p,s,n,d} = 1$; we denote such a combination by s_p and n_p.

The formulation can be extended to include various additional constraints. For example, if a meal break is one of the requirements, then $ILP_d^{(k)}$ can be forced to reserve for each nurse a set of contiguous time slots (two 15-minute slots in our experiments) in the middle of the day such that no patient is treated at that time, leaving the nurse free from treatment activities for that period. This can be achieved by introducing variables $b_{t,n,d}$, which define for nurse n the starting time of a break on day d:

$$b_{t,n,d} = \begin{cases} 1, & \text{if the meal break of nurse } n \in N_d \text{ starts at time } t \in H_{n,d}, \\ 0, & \text{otherwise.} \end{cases}$$

Let $M_{n,d}$ be a set of possible starting times of the meal break of nurse $n \in N_d$. For a patient $p \in P_d^{\mathcal{A}}$, introduce a set $B_{p,t,n,d}$ of appointment starting times which incur a treatment activity during the meal break of nurse n starting at time $t \in M_{n,d}$. Clearly, if the meal break of nurse n is scheduled at time t (i.e., $b_{t,n,d} = 1$), patient $p \in P_d^{\mathcal{A}}$ cannot be allocated to nurse n with appointment starting time $s \in B_{p,t,n,d}$. Then, $ILP_d^{(k)}$ can be adjusted by introducing the following additional constraints:

$$\begin{array}{ll} \sum\limits_{p \in P_d^{\mathcal{A}}} \sum\limits_{s \in B_{p,t,n,d}} x_{p,s,n,d} + b_{t,n,d} \leq 1, & n \in N_d,\ t \in M_{n,d}, \\ \sum\limits_{t \in M_{n,d}} b_{t,n,d} = 1, & n \in N_d. \end{array}$$

Observe that it is easy to ensure that the meal break is of the required length by defining the set $B_{p,t,n,d}$ appropriately.

Suppose the template schedule is created for artificial patients $P^{\mathcal{A}}$ to be treated in time horizon H. In Stage 3, described in the next section, the template schedule is used for generating and maintaining a running schedule for actual patients P by fixing their appointments in time slots reserved for artificial patients. Using the approach of the rolling time horizon, at some point within time horizon H a new template schedule is produced for a new time horizon H' overlapping with H. The new template schedule is found as a solution to a multi-day problem $ILP_{H'}$ and a series of intra-day problems $ILP_d^{(k)}$ for $d \in H'$, $1 \leq k \leq K$, which are similar to problems ILP_H and $ILP_d^{(k)}$, $d \in H$, $1 \leq k \leq K$, with one point of difference: the appointments of actual patients which have been fixed in the running schedule must be kept. This can be achieved by adding the constraints $x_{p,s,n,d} = 1$ for each pre-scheduled actual patient $p \in P$, who should visit the clinic on day d to be treated by nurse n with the first visit day s, $s \leq d$. On the other hand, the appointments of artificial patients which are not booked so far for actual patients can be rescheduled at no cost, so that the corresponding x-variables are free.

8. Running Schedule: Creating and Maintaining

Once a template schedule is generated, it is used to assign appointment dates and times for arriving patients. The booking process for a new patient consists in identifying in the template schedule available appointments of artificial patients with the same regimen and selecting the combination of appointments that satisfies the multi-day pattern and additional preference conditions. Then the dates and starting times of chosen appointments become dates and times of appointments of the new patient. In this section we propose an algorithm that selects appointments from the template schedule with the aim of optimizing the quality of the generated running schedule in terms of metrics F1-F3 (see Section 1).

Consider a newly arrived patient $p \in P$ with regimen $r \in R$ whose first visit date d_p has to be within a given time window $[d_p^{\min}, d_p^{\max}]$ determined by patient's requirements and clinic waiting targets.

The proposed algorithm searches for a possible date d_p in $[d_p^{\min}, d_p^{\max}]$ that allows the selection of appointments from the template schedule satisfying the multi-day pattern of regimen r.

One simple strategy is to choose in the template a single artificial patient with regimen r and to assign all appointments of actual patient p to the dates and times of pre-scheduled appointments of that artificial patient. Although this strategy is quite natural and simple, it has a major drawback: if an appointment of the template is left unused and the date of the first visit has passed, all unused parts of that multi-day pattern, pre-scheduled for some future dates, cannot be used any more by newly arriving actual patients.

In order to overcome such a drawback and to make use of the pre-scheduled appointments in the template more efficiently, we consider the template schedule as a collection of artificial appointments breaking the link to their multi-day patterns. Then, in order to book appointments for actual patient p with regimen r, we consider all artificial appointments with regimen r and select such a combination of appointments that satisfies the intra-day and multi-day patterns of regimen r and our preference criteria (described below). In particular, during the selection we make sure that in the resulting running schedule, the daily workload of the clinic is not exceeded and that the number of clashing activities in each intra-day schedule is as small as possible. Thus, the appointments selected for actual patient p might correspond to appointments of several artificial patients, but they satisfy the requirements of patients p and potentially lead to a good running schedule. Such strategy can be applied in both cases when one or multiple intra-day patterns are used for one regimen.

We describe how to perform a feasibility test for selecting day d_p as the first visit day of actual patient p with regimen r. Let $\pi_r = (\pi_r(1), \pi_r(2), \ldots, \pi_r(z_r))$ be the multi-day pattern associated with regimen r, where $\pi_r(1) = 1$ and

- $\pi_r(j)$ - 1 is the number of days from the first visit to the j^{th} appointment;
- z_r is the total number of appointments of the multi-day pattern of regimen r.

The first visit date d_p of actual patient p should be selected in such a way that for each appointment j, $j = 1, 2, \ldots, z_r$, there exists a matching artificial appointment scheduled in the template on date $d = d_p + \pi_r(j) - 1$. The days and starting times of the selected artificial appointments are then booked for patient p.

An artificial appointment of day d of the the template matches the j^{th} appointment of actual patient p if

(i) it is associated with regimen r,

(ii) it is free, i.e., it has not been used to book an appointment for another actual patient,

(iii) the clinic workload of day d, if increased by treatments activities needed for additional patient p, does not exceed a given threshold.

Verifying the first two conditions is straightforward. In what follows we clarify the last condition.

In order to keep the workload of the clinic within acceptable limits and to reserve some proportion of working hours for additional nurse duties, the relative workload of a clinic on any day must not exceed a given threshold $\sigma < 1$ called *capacity ratio.* The threshold σ represents the proportion of nurses' time which can be booked for treatment activities. Using the notation from Section 5 and the notion of the average relative workload $\hat{W}_d$, the corresponding constraint can be expressed as

$$\hat{W}_d \leq \sigma.$$

In accordance with definition (3) of the relative workload, the capacity requirement can be re-written as

$$\frac{W_d + w_r}{C_d} \leq \sigma,$$

where W_d is the overall time required for treating actual patients who have appointments on day d, C_d is the capacity of the clinic on day d measured as the total number of 15-minute time-slots when the nurses are available, and w_r is the total number of time-slots needed for treatment activities of patient p.

Summarizing, a date $d_p \in [d_p^{\min}, d_p^{\max}]$ is a feasible date for the first appointment of patient p with regimen r if for each day $d_p + \pi_r(j) - 1$, $j = 1, 2, \ldots, z_r$, there exists a matching artificial appointment satisfying conditions (i), (ii) and (iii).

The minimization of the number of clashing tasks is fundamental in order to achieve a successful running schedule (see metric F3). For this reason our algorithm chooses a feasible starting date d_p such that the maximum density level k of the selected artificial appointments is minimum. By the definition of a density set, selecting appointments belonging to the set $\mathcal{L}_d^{(k)}$ results in an intra-day schedule with clash density not exceeding k.

The formal description of the algorithm is given by procedure 'Match-Appointments(p,σ)' presented below. It is assumed that the template schedule is represented by the set of artificial appointments grouped in density sets $\mathcal{L}_{r,d}^{(k)}$ for each day $d \in H$ and each regimen $r \in R$. The algorithm uses the 'Feasibility-Test(r, j, d, k, σ)' which verifies whether for the jth visit day of the patient with regimen r there exists a matching artificial appointment of set $\mathcal{L}_{r,d}^{(k)}$ satisfying conditions (i), (ii) and (iii).

Procedure 'Match-Appointments(p, σ)'

$k := 0$;

WHILE appointments $\pi_r(1)$, $\pi_r(2)$, $\ldots$, $\pi_r(z_r)$ for patient p are not booked DO

  Set the density level $k := k + 1$;

  FOR $d_p = d_p^{\min}$ TO $d_p^{\max}$ DO

    IF 'Feasibility-Test(r, j, d, k, σ)' confirms for each $j = 1, 2, \ldots, z_r$, that the jth appointment of patient p can be assigned to the corresponding day $d = d_p + \pi_r(j) - 1$ and the matching artificial appointments belong to density set $\mathcal{L}_{r,d}^{(k)}$

    THEN book all appointments for patient p with the first visit day d_p; STOP

  END FOR

END WHILE

Notice that in some extreme cases a set of feasible dates and times may not be found by procedure 'Match-Appointments(p,σ)'. For example, if patient arrival rates substantially diverge from the forecast of Stage 1, artificial appointments for arriving patients may not be available in the template schedule. Our experiments show that this difficulty can be overcome by using an appropriate overestimation rate of patient arrival at the stage of template generation.

In general, the matching strategy and the method used to estimate the number of arrivals (at the template generation stage) provide a strong mechanism for adjusting the model to address further enhancements, related to additional constraints and optimization criteria. For example, an additional objective of minimizing the number of patients waiting longer than a certain threshold can be achieved by adjusting the time window $[d_p^{\min}, d_p^{\max}]$ and tuning the overestimation parameter accordingly.

9. Daily Rescheduling

In this section we develop an integer linear program to adjust a one-day schedule in order to achieve an improvement in the running schedule in terms of metrics F2, F3 introduced in Section 1. In particular, we reduce the number of clashing activities and the maximum clash density of each day by full re-allocation of nurses, introducing minor shifts of nurses' meal breaks and minor delays in starting times of the pre-booked appointments. Observe that changing nurse allocation does not incur any cost as a patient can be treated by different nurses on different visit days.

Rescheduling for day d can be performed when complete information about all booked appointments for that day is known. The integer linear program presented below is a reformulation of $ILP_d^{(k)}$ introduced in Section 7 with slightly modified objective functions and constraints.

Consider a set of patients P_d visiting the clinic on day d. For each patient $p \in P_d$, let s_p be the starting time of the pre-booked appointment and $n_p \in N_d$ be the nurse allocated to patient p on that day. The rescheduling problem consists of finding for each patient p a new starting time s'_p and a new nurse allocation $n'_p \in N_d$ such that the maximum clash density

$$\Delta_d = \max_{n \in N} \Delta_{d,n} \tag{7}$$

and the total number of clashing activities

$$\Omega_d = \sum_{n \in N} \Omega_{d,n} \tag{8}$$

are minimized (see Section 1 for the definition of $\Delta_{d,n}$ and $\Omega_{d,n}$).

Since only small delays in patients' starting times are acceptable, a new starting time s'_p of the appointment of patient p can take values from a restricted set $H'_{p,d} \subseteq H_d$, where H_d is the set of 15-minute time intervals of day d. For example, if it is acceptable to delay starting times by at most two 15-minute time slots, then

$$H'_{p,d} = \{t \in H_d \mid s_p \le t \le s_p + 2\}.$$

Consider now reallocation of patients to nurses. Introduce a set of patients $P_{s,i,n,d}$ such that the intra-day pattern of patient $p \in P_{s,i,n,d}$ incurs an activity for nurse $n \in N_d$ in time slot i, if the appointment starts at time s. Allocation of patient p with appointment starting time s'_p to nurse n'_p may be infeasible if it is not possible to complete all treatments of the intra-day pattern of patient p within the working hours of nurse n'_p. Therefore we limit our consideration to a set of nurses $N_{p,s,d} \subseteq N_d$ whose working hours on day d allow to perform all treatments of patient p with starting time s.

Similar to formulation $\text{ILP}_d^{(k)}$, we define the decision variables $x_{p,s,n,d}$ and $b_{t,n,d}$:

$$x_{p,s,n,d} = \begin{cases} 1, & \text{if the appointment time of patient } p \text{ is } s \text{ and the allocated nurse is } n, \\ 0, & \text{otherwise.} \end{cases}$$

$$b_{t,n,d} = \begin{cases} 1, & \text{if nurse } n \text{ has a meal break starting at time } t, \\ 0, & \text{otherwise,} \end{cases}$$

and auxiliary variables $y_{i,n,d}$ to measure the number of clashing activities. Recall that

$$y_{i,n,d} \ge 1,$$

where $y_{i,n,d} = 1$ represents either of the cases: the nurse is free in time slot i or performs one activity in time slot i and there are no other clashing activities, see Section 7. The case $y_{i,n,d} > 1$ corresponds to the number of clashing activities of nurse n happening in time slot i and it is calculated as

$$y_{i,n,d} \ge \sum_{s \in H_{n,d}} \sum_{p \in P_{s,i,n,d}} x_{p,s,n,d}.$$

The two main objective functions of the rescheduling problem are the maximum clash density Δ_d and the number of clashing tasks Ω_d, defined by (7) and (8), respectively:

$$\Delta_d = \max_{n \in N_d,\ i \in H_{n,d}} \{y_{i,n,d}\},$$

$$\Omega_d = \sum_{n \in N_d} \sum_{i \in H_{n,d}} (y_{i,n,d} - 1).$$

An additional objective is aimed at balancing the workload of different nurses and it is measured as the maximum difference W_d^{diff} in nurses' workloads:

$$W_d^{\text{diff}} = \max_{n_1, n_2 \in N_d} \left\{ \left| \hat{W}_{n_1,d} - \hat{W}_{n_2,d} \right| \right\}.$$

Recall that the relative workload $W_{n,d}$ of nurse n on day d is defined by (2) and it can be calculated as

$$\hat{W}_{n,d} = \frac{1}{|H_{n,d}|} \sum_{s \in H_{n,d}} \sum_{i \in H_{n,d}} \sum_{p \in P_{s,i,n,d}} x_{p,s,n,d}.$$

Since the above three criteria are conflicting, we establish an order of their importance for lexicographical optimization. Our first priority is to minimize the maximum clash density Δ_d; secondly, among the solutions with the smallest clash density, we give priority to those with the smallest number of clashing activities Ω_d; finally, among the solutions with the smallest Δ_d and Ω_d we select those with the minimum workload difference W_d^{diff}.

Summarizing, the resulting integer linear program ILP_d is of the form:

$$ILP_d: \quad \text{lex}\left[\min \Delta_d,\ \min \Omega_d,\ \min W_d^{\text{diff}}\right]$$

$$\begin{array}{llll}
\text{s.t.} & \Delta_d \geq y_{i,n,d}, & n \in N_d, & i \in H_{n,d}, \\
& \Omega_d = \sum\limits_{n \in N_d} \sum\limits_{i \in H_{n,d}} (y_{i,n,d} - 1), & & \\
& W_d^{\text{diff}} \geq \hat{W}_{n_1,d} - \hat{W}_{n_2,d}, & n_1, n_2 \in N_d, & \\
& \hat{W}_{n,d} = \frac{1}{|H_{n,d}|} \sum\limits_{s \in H_{n,d}} \sum\limits_{i \in H_{n,d}} \sum\limits_{p \in P_{s,i,n,d}} x_{p,s,n,d}, & n \in N_d, & \\
& y_{i,n,d} \geq \sum\limits_{s \in H_{n,d}} \sum\limits_{p \in P_{s,i,n,d}} x_{p,s,n,d}, & n \in N_d, & i \in H_{n,d}, \\
& y_{i,n,d} \geq 1, & n \in N_d, & i \in H_{n,d}, \\
& \sum\limits_{s \in H'_{p,d}} \sum\limits_{n \in N_{p,s,d}} x_{p,s,n,d} = 1, & p \in P_d, & \\
& \sum\limits_{p \in P_d} \sum\limits_{s \in B_{p,t,n,d}} x_{p,s,n,d} + b_{t,n,d} \leq 1, & n \in N_d, & t \in M_{n,d}, \\
& \sum_{t \in M_{n,d}} b_{t,n,d} = 1, & n \in N_d & \\
& x_{p,s,n,d} \in \{0,1\}, & p \in P_d, & s \in H'_{p,d}, \quad n \in N_{p,s,d}, \\
& b_{t,n,d} \in \{0,1\}, & t \in M_{n,d} & n \in N_{p,s,d}.
\end{array}$$

Notice that the pre-assigned appointment starting time s_p and nurse allocation n_p for every patient $p \in P_d$ define a feasible solution to ILP_d, which can be used as an initial solution for that problem. An optimal solution to ILP_d determines new appointment times s'_p and nurse allocations n'_p for all patients $p \in P_d$.

10. Results of Computational Experiments

In this section we describe typical problem instances in a real clinic, our design of experiments and computational results.

At the chemotherapy outpatient clinic of the St. James's Hospital in Leeds, an average of 800 appointments are scheduled every month. Treatments are performed by 19 nurses which have different working hours. The clinic works 5 days a week and every day about 40 patients are treated by 8 nurses. According to the recorded data, the time from the decision to treat to the date of the first appointment is 14 days on average. In a typical monthly schedule, there are about 450 clashing tasks while in a typical intra-day schedule, there are about 20 clashing tasks. Patients' waiting times on treatment days can be as large as 2 hours. The percentage of time a nurse spends on treatment activities in relation to overall nurse's working hours is 47% on average.

We perform two types of experiments evaluating the operation of the clinic

- if only the rescheduling procedure (Stage 4 of our approach) is used in addition to the existing manual scheduling policy,
- if the whole four-stage scheduling approach is adopted.

The evaluation is based on historical data recorded at the St. James's University Hospital in Leeds during the period from 1st May 2008 to 1st September 2009.

In our first set of experiments we show how the rescheduling procedure corresponding to Stage 4 of our approach can be used to improve daily schedules produced in the hospital manually. We use actual daily schedules for one month of the recorded historical data (May 2008). The results are summarized in Table 4 where we compare the average number of daily clashes over the one month period depending on rescheduling constraints.

Our experiments demonstrate that operation of the clinic can be substantially improved if nurse reallocation is done (compare the figures in the first two columns of Table 4). Further improvement can be achieved if delays are allowed in starting times of patients' appointments (see the figures in the last four columns of Table 4). Observe that such delays do not affect all patients and they are not necessarily of maximum duration (see Table 5).

	Actual schedule	Nurse reallocation; no delays in appointments starting times	Nurse reallocation and possible delays in appointment starting times			
			15 min delay	30 min delay	45 min delay	complete rescheduling
Number of clashing tasks per month	454	110	41	30	22	0
Max clash density	6	4	4	2	2	1
Daily avg. clash density	3.2	1.95	1.55	1.25	1.20	1

Table 4: Comparison of actual daily schedules of May 2008 used at the St. James's Hospital (Leeds, U.K.) and those obtained via rescheduling

	Nurse reallocation and possible delays in appointment starting times		
	15 min delay	30 min delay	45 min delay
patients not waiting	53.98 %	44.96 %	40.19 %
patients waiting 15 mins	46.02 %	22.94 %	18.81 %
patients waiting 30 mins	-	32.10 %	17.78 %
patients waiting 45 mins	-	-	23.22 %

Table 5: The effect of rescheduling on patient waiting times

In the second set of experiments, we evaluate the four-stage approach on two scenarios, one with patients' arrival rates similar to actual rates in the real clinic (119 patients per week on average) and another one with higher arrival rates (about 141 patients per week). We consider the 160 regimens which occur in the real clinic most often. Our experiments cover only one time horizon H consisting of 134 days with the first 60 days corresponding to the arrival period H_a (see Section 3) which includes a warm-up period of 30 days. Solution quality is evaluated for the period between the 30^{th}

and the 60^{th} day, excluding the warm up period, where the simulation does not consider patients arrived in the past, as well as the final period, where no new patients are scheduled. The need to generate the data is caused by the presence of the noise in historical data. Since the generated data closely follows the characteristics of the historical data, we compare our results with the actual performance characteristics of the clinic.

Stages 1-2 are performed only once producing one multilevel template schedule, which is then used to generate running schedules for both scenarios. The data for artificial patient is generated considering the arrival rate λ_r calculated as the arithmetic mean of weekly arrival rates recorded in the historical data. The value of parameter ξ, which defines the level of overestimation, is set to 5.

Stages 3 and 4 are evaluated on 100 datasets, which describe patients' arrivals simulated using a Poisson process. Each dataset we generate contains arrival dates of patients for all regimens. For the first scenario, arrival rates are the same as λ_r-values used in Stage 2. For the second scenario the arrival rates are increased by about 20%.

All scheduling decisions for a patient have to be made on the arrival date. Depending on the category of a patient, the first appointment should happen within 7, 14 or 28 days from the arrival date. We denote the corresponding groups of patients as A, B and C, assuming that the patients are split in proportion 73% , 15% and 12%, respectively, reproducing the statistical trends in historical patient data.

Both scenarios use the same nurses' weekly rota which is shown in Table 6. The shifts have different lengths giving nurses some flexibility in negotiating preferred working hours. Notice that on Monday nurse shifts start no earlier than 11 a.m. since the hospital pharmacy needs additional set up time after a weekend in order to prepare drugs for treatments. It is assumed that each nurse needs a 30-minute lunch break between 11 a.m. and 2 p.m.

The capacity ratio σ used in Stage 3 to generate and maintain running schedules is set to 0.9 bounding the total number of treatment activities performed by a nurse. This implies that in addition to a 30-minute lunch break, nurse's schedule should contain at least 10% of unused time slots for additional duties.

	Mon	Tue	Wed	Thu	Fri
Shift 1	11:00-18:00	9:00-18:15	9:00-18:15	9:00-18:00	9:00-16:45
Shift 2	11:00-18:00	9:00-18:15	9:00-18:15	9:00-18:00	9:00-16:45
Shift 3	11:00-18:15	9:00-18:15	9:00-18:15	9:00-18:00	9:00-16:45
Shift 4	11:30-18:15	9:00-18:15	9:00-18:15	9:00-16:45	9:00-16:45
Shift 5	12:00-18:30	9:00-17:00	9:00-18:15	9:00-16:45	9:00-16:15
Shift 6	12:00-19:00	9:00-17:00	9:00-17:30	9:00-16:45	9:00-16:15
Shift 7	-	9:00-17:00	9:00-17:00	9:00-16:45	9:00-16:15
Shift 8	-	9:00-17:00	9:00-17:00	9:00-16:45	9:00-16:15
Shift 9	-	-	-	-	9:00-16:45

Table 6: Weekly nurse rota for scenario 1 and 2

	Monthly number of appointments	Daily Clashing Tasks			Daily Clash Density			Relative Workload (%)		
		Avg.	Dev.	Max.	Avg.	Dev.	Max.	Avg.	Dev.	Max.
Real clinic	992	22.7	10.43	43	3.20	0.80	6	47.28	12.71	62.90
Scenario 1	1024	1.25	2.07	15	1.42	0.51	3	62.71	13.84	84.80
Scenario 2	1277	1.95	2.55	18	1.56	0.52	3	71.13	12.53	85.86

Table 7: Characteristics of schedules produced for Scenario 1 and 2

Tables 7 and 8 characterize the quality of the generated running schedules. In the first scenario, our approach is able to schedule appointments reducing the average number of waiting days (see

	Waiting days							Waiting times (minutes)	
	Total	Group A		Group B		Group C			
	Avg	Avg.	Dev.	Avg.	Dev.	Avg.	Dev.	Avg.	Dev.
Real Clinic	14	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Scenario 1	4.69	3.80	4.85	5.79	4.83	8.97	6.50	11:11	12:28
Scenario 2	5.26	4.42	5.46	6.28	5.26	9.30	6.73	10:59	12:25

Table 8: Patients' waiting days and times

Table 8), eliminating almost all clashing tasks (see Table 7) and, therefore, reducing their density and increasing nurses' relative workloads defined by (2).

The results of experiments for the second scenario are quite remarkable. With an increased number of arriving patients it is indeed inevitable that the waiting time characteristics and the number of clashes should increase. It appears that for a 20% increase in the number of patients (about 89 additional patients per month or 253 additional appointments) the deterioration in schedule quality is marginal: for each category of patients the number of waiting days increases on average by no more than 1 day, while the average number of clashing tasks changes from 1.25 to 1.95 only. The comparison with the actual booking process of the clinic is even more dramatic: the quality of actual schedules generated manually for 992 appointments is much worse than the quality of schedules generated by our approach for 1277 appointments.

The experiments were performed on Intel Pentium Core 2 Quad CPU 2.5GHz and 3GB RAM. The ILP programs of Stages 2 and 4 were solved by a single thread version of CPLEX 11.2. Time estimates for the four stages of our approach are as follows.

- Stage 1 is implemented by calculating averages of patients' arrival rates; the time required by these calculation is negligible.
- Since the template schedule is the major factor which affects the quality of running schedules, we allowed more time for solving the associated ILP programs: 1 hour for the solution of the multi-day problem and 8 hours for the solution of all intra-day problems of the time horizon. In the environment of a real clinic, long computation time of Stage 2 is acceptable as template schedules are generated rarely and serve for sufficiently long time. These calculations can be performed, for example, overnight.
- Stage 3 is a fast heuristic requiring less than 1 second to book all appointments for a patient.
- Stage 4 is implemented as a single ILP problem for each day. We set up a time limit of 120 seconds for rescheduling each intra-day schedule; however optimal solutions are often found within 60 seconds.

In order to set up the time limit for the series of intra-day problems of Stage 2, we use the following approach. The 8-hour time limit (28800 seconds) is divided among the days of the time horizon H in proportion to the number of daily appointments assigned. This implies that the time limit T_d for solving the intra-day problem for day d can be defined as

$$T_d = 28800 \times \frac{|P_d^{\mathcal{A}}|}{\sum_{d \in H} |P_d^{\mathcal{A}}|},$$

where $\left|P_d^{\mathcal{A}}\right|$ is the number of artificial patients to be treated on day d.

Generating an intra-day schedule for one day of the template involves a number of integer linear programs $ILP_d^{(k)}$, one for each clash density level $k = 1, 2, \ldots, K$. Although the total number of density levels K in an intra-day schedule for day d can be as large as the number of artificial patients $\left|P_d^{\mathcal{A}}\right|$ scheduled on that day, for our datasets K can be bounded by a much smaller number defined empirically:

$$K \leq 5.$$

Due to this we set up a time limit for each program $ILP_d^{(k)}$ as

$$T_d^{(k)} = \frac{T_d}{5}.$$

Finally, since there are three objective functions for each problem $ILP_d^{(k)}$ optimized lexicographically, a time limit of $T_d^{(k)}/3$ is imposed for optimizing one function.

Notice that, due to the size and complexity of the ILP problems used for creating the template schedule, introducing the time limits is important to achieve admissible computation time (which, otherwise, may be as large as several days). Although we cannot guarantee the optimality of the template schedule, the resulting running schedules are still of high quality.

Summarizing we observe that the computational experiments demonstrate the advantages of the proposed approach evaluated against the performance metrics of the model. Further experiments run in a real clinic in parallel with the existing manual system might provide additional evidence of operational benefits which the streamlined scheduling procedures can bring.

11. Conclusions

In this paper we have introduced a new appointment scheduling problem which arises in the context of chemotherapy outpatient clinic. If all information about patients is known in advance, producing good schedules for nurses and patients which treatments should follow a combination of multi-day and intra-day patterns with constraints on waiting times appears to be a difficult task. The on-line nature of the problem with patients arriving over time adds even more complexity to the scheduling process.

The approach we propose consists of four stages: data generation for artificial patients, generating a template schedule, producing a running schedule and rescheduling. The major underlying idea is essentially based on the concept of multilevel template schedule which represents a well thought through plan. The template schedule contains more pre-booked appointments than anticipated, providing flexibility in selecting the most appropriate options for arriving patients and for handling unexpected arrivals. It is obtained as a solution to a series of integer linear programs with multiple objectives optimized lexicographically. Due to this, the template ensures that the final running schedule is potentially of high quality and satisfies the requirements of patients and nurses.

In order to use the template for booking actual appointments and creating a running schedule, we design a matching procedure which takes into account characteristics of appointments of the template and requirements of arriving patients. Finally, a running schedule is further improved via rescheduling.

Thus, the novelty of our work lies in

- the introduction of a new scheduling model with jobs consisting of repetitive tasks satisfying given multidimensional patterns;
- the formulation of the concept of multi-level template for scheduling patients in an uncertain environment and in its development into a formal optimization model with advanced features;
- the integration of daily rescheduling procedures into long-term planning.

The proposed approach can be enhanced with various additional features. We have demonstrated how it can be extended to take into account the requirements of meal breaks for nurses. Further enhancements may include patients' preferences on appointment starting times (morning, midday or afternoon) or requirements of pharmaceutical suppliers and hospital pharmacy on drug preparation.

Computational experiments demonstrate that our approach can potentially bring substantial improvement in operation of a real clinic in different ways:

- maintaining patients' waiting times within required limits;
- improving nurses' schedules by reducing the number of clashing tasks (from an average of 20 clashing tasks per day to less than 2 clashes),
- increasing the clinic capacity in terms of additional patients (in our experiments based on real-world data, up to 89 patients can be treated monthly in addition to the current 476 patients on average) without extending nurses' working hours and avoiding essential deterioration in schedule quality.

We strongly believe that the concept of template schedule can be used as a powerful algorithmic tool to tackle complex online scheduling problems, especially those which involve multi-operation jobs with given patterns.

It will be interesting to consider alternative approaches for generating template schedules, for example, an integrated approach for solving multi-day and intra-day problems simultaneously rather than sequentially. Due to the size and complexity of the integrated problem, it will be appropriate to develop metaheuristics and to compare the quality of the resulting schedules with those produced by the current ILP-based approach which treats multi-day and intra-day problems sequentially. Another possible improvement could be achieved in Stage 1 by developing advanced models for prediction of future demands.

12. Acknowledgement

The authors would like to thank the Department of Informatics and the Outpatient Chemotherapy Clinic of the Leeds Teaching Hospitals NHS Trust. Our special thanks to Andy Shaw who introduced the problem to us, provided the clinical data and support.
This research was supported by the Doctoral Training Grant EP/P503868/1.
The authors are grateful to the referees for careful reading of the text and suggested improvements.

References

[1] *Chemotherapy Services in England: Ensuring Quality and Safety.* November 2008.

[2] *Cancer Reform Strategy.* December 2009.

[3] N. Bailey. A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society*, A14:185–189, 1952.

[4] A. Braaksma, N. Kortbeek, G.F. Post, and F. Nollet. Integral multidisciplinary rehabilitation treatment planning. Memorandum 1985. *Department of Applied Mathematics, University of Twente, Enschede*, pages 1–18, 2012.

[5] M. Brahimi and D. J. Worthington. Queueing models for out-patient appointment systems – a case study. *Journal of the Operational Research Society*, 42:733–746, 1991.

[6] T. Cayirli and E. Veral. Outpatient scheduling in health care: A review of literature. *Production & Operations Management*, 12:519–548, 2003.

[7] C.-C. Chern, P.-S. Chien, and S.-Y. Chen. A heuristic algorithm for the hospital health examination scheduling problem. *European Journal of Operational Research*, 186:1137–1157, 2008.

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

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

[10] A. Condotta. Scheduling with Due Dates and Time-Lags: New Theoretical Results and Applications. *PhD Thesis, School of Computing, Universlity of Leeds*, 2011.

[11] A. Condotta and N.V. Shakhlevich. Scheduling coupled-task jobs with unit-time jobs. *Discrete Applied Mathematics*, 160:2370–2388, 2012.

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

[13] D. Conforti, F. Guerriero, and R. Guido. Non-block scheduling with priority for radiotherapy treatments. *European Journal of Operational Research*, 201:289–296, 2010.

[14] R. Dobish. Next-day chemotherapy scheduling: a multidisciplinary approach to solving workload issues in a tertiary oncology center. *Journal of Oncology Pharmacy Practice*, 9:37–42, 2003.

[15] R. B. Fetter and J. D. Thompson. Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research*, 1:66–90, 1966.

[16] J.D. Griffiths, J.E. William, and R.M. Wood. Scheduling physiotherapy treatment in an inpatient setting. *Operations Research for Health Care*, 1:65–72, 2012.

[17] J.D. Griffiths, J.E. William, and R.M. Wood. Modelling activities at a neurological rehabilitation unit. *European Journal of Operational Research*, 226:301–312, 2013.

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

[19] P.J.H. Hulshof, N. Kortbeek, R.J. Boucherie, E.W. Hans, and P.J.M Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems*, 1:129–175, 2012.

[20] D.V. Lindley. The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48:277–289, 1952.

[21] S. Ogulata, M. Cetik, E. Koyuncu, and M. Koyuncu. A simulation approach for scheduling patients in the department of radiation oncology. *Journal of Medical Systems*, 33:233–239, 2009.

[22] S.N. 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*, 32:221–228, 2008.

[23] A.J. Orman and C.N. Potts. On the complexity of coupled-task scheduling. *Discrete Applied Mathematics*, 72:141–154, 1997.

[24] S. Petrovic, W. Leung, X. Song, and S. Sundar. Algorithms for radiotherapy treatment booking. In R. Qu, editor, *Proceedings of the 25th Workshop of the UK Planning and Scheduling Special Interest Group (PlanSIG'2006)*, pages 105–112, 2006.

[25] D. Petrovich, M. Morsed, and S. Petrovich. Multi-objective genetic algorithms for scheduling of radiotherapy treatments for catagorised cancer patients. *Expert Systems with Applications*, 38:6994–7002, 2011.

[26] V. Podgorelec and P. Kokol. Genetic algorithm based system for patient scheduling in highly constrained situations. *Journal of Medical Systems*, 21:417–427, 1997.

[27] C.N. Potts and J.D. Whitehead. Heuristics for a coupled-operation scheduling problem. *Journal of the Operational Research Society*, 58:1375–1388, 2007.

[28] A. Turkcan, B. Zeng, and M. Lawley. Chemotherapy operations planning and scheduling. *IIE Transactions on Healthcare Systems Engineering*, 2:31–49, 2012.

[29] I. Vermeulen, S. Bohte, P. Bosman, S. Elkhuizen, P. Bakker, and J. La Poutré. Optimization of online patient scheduling with urgencies and preferences. *Lecture Notes in Computer Science*, 5651:71–80, 2009.

[30] J. Vissers. Selecting a suitable appointment system in an outpatient setting. *Medical Care*, 17:1207–1220, 1979.

[31] A. Wijewickrama and S. Takakuwa. Simulation analysis of appointment scheduling in an outpatient department of internal medicine. In M.E. Kuhl, N.M. Steiger, F.B. Armstrong, and J.A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 2264–2273. Winter Simulation Conference, 2005.

[32] P. Williams, G. Tai, and Y. Lei. Simulation based analysis of patient arrival to health care systems and evaluation of an operations improvement scheme. *Annals of Operations Research*, 178:263–279, 2010.

[33] B. Zeng, A. Turkcan, J. Lin, and M. Lawley. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Annals of Operations Research*, 178:121–144, 2010.