An adaptive priority policy for radiotherapy scheduling

Li, Siqiao; Koole, Ger; Xie, Xiaolan

published in
Flexible Services and Manufacturing Journal
2020

DOI (link to publisher)
10.1007/s10696-019-09373-4

document version
Publisher's PDF, also known as Version of record

Flexible Services and Manufacturing Journal (2020) 32:154–180
https://doi.org/10.1007/s10696-019-09373-4

An adaptive priority policy for radiotherapy scheduling

Siqiao Li[1,2] · **Ger Koole**[2] · **Xiaolan Xie**[1,3]

Published online: 4 November 2019

Abstract
In radiotherapy, treatment needs to be delivered in time. Long waiting times can result in patient anxiety and growth of tumors. They are often caused by inefficient use of radiotherapy equipment, the linear accelerators (LINACs). However, making an efficient schedule is very challenging, especially when we have multiple types of patients, having different service requirements and waiting time constraints. Moreover, in radiotherapy a patient needs to go through a LINAC multiple times over multiple days, to complete the treatment. In this paper we model the radiotherapy treatment process as a queueing system with multiple queues, and we propose a new class of scheduling policies that are simple, flexible and fair to patients. Numerical experiments show that our new policy outperforms the commonly used policies. We also extend the policy to an adaptive one to deal with unknown and fluctuating arrival rates. Our adaptive policy turns out to be quite efficient in absorbing the effects caused by these changes. Due to the complexity of our problem, we select the parameters of the policies through simulation-based optimization heuristics. Our work may also have important implications for managers in other service systems such as call centers.

Keywords Healthcare · Adaptive · Routing policy · Patient scheduling · Simulation-based heuristic

✉ Siqiao Li
aprilsiqiao@sjtu.edu.cn; l.s.q.li@vu.nl

Ger Koole
ger.koole@vu.nl

Xiaolan Xie
xie@emse.fr

[1] Shanghai Jiao Tong University, Shanghai, China

[2] Vrije Universiteit Amsterdam, Amsterdam, Netherlands

[3] Ecole des Mines de Saint-Étienne, Saint-Etienne, France

1 Introduction

According to the World Health Report Organization (2017), about 8.8 million people died from cancer in 2015 and the annual number of new cases in the world-wide reached 14.1 million. It is estimated that these numbers will increase by 70% in the next 20 years. As one of the main treatment methods for cancer, radiotherapy uses different levels of radiation (e.g., X-rays) to kill tumors. It is the leading treatment for several cancer types (e.g., head and neck cancer). It also can be combined with other therapies such as surgery and chemotherapy to further reduce and remove tumors. Indeed, about 40% of all cancer patients receive radiotherapy as part of their treatment.

Increasing demand for radiotherapy force hospital managers to face the challenge of how to deliver in-time treatment to patients for the lowest possible costs. It has been shown in Chen et al. (2008) and other papers that long waiting times can lead to patient anxiety and unexpected growths of tumors which have a negative impact on clinical outcomes. In fact, a waiting time target (WTT) measured in days is often assigned to each patient type. It can be interpreted as the patient's longest acceptable waiting time. The main costs are the Linear Accelerators (LINACs). Therefore, the focus is on how to schedule multiple types of patients on a limited number of LINACs so as to meet various WTT requirements.

Let us first take a closer look at the radiotherapy process. Before a referral patient receives treatment on a LINAC (the so-called treatment phase), he/she needs to go through consultation, examination, tumor location, and simulation, which altogether are called the pre-treatment phase. At the end of the pre-treatment phase, a diagnosis-based treatment protocol is proposed for this patient whose status then becomes "ready to treat". The waiting time we consider is the time between the "ready to treat" date and the real start date of treatment. Indeed, the treatment phase is usually the bottleneck in the whole cancer-treatment process Legrain et al. (2015).

In the treatment phase, patients have a typical "re-entrance" behavior. This behavior is due to the fact that healthy cells around tumors need to be protected as much as possible during irradiation. Therefore, the total needed dosage of a patient is not delivered once and for all, but via multiple equal fractions. Once a patient starts the treatment, he/she needs to receive a fixed small dosage of rays (i.e., a fraction) every day from a LINAC until all needed fractions are carried out. Except for a routine break in the weekend, interruptions are not recommended during the whole treatment. Every fraction occupies a fixed time slot (e.g., 15 min) of a LINAC. We neglect the slight difference in fraction durations between patients since we could average it at the scheduling level. This "re-entrance" behavior leads to a very challenging scheduling problem because once we schedule a patient for a time slot, the same time slot will be occupied in the following days until he/she departs. The system is even more difficult to analyze if we consider multiple types of patients. However, in reality, radiotherapy patients are categorized by the number of fractions and the WTT requirements according to tumor position, treatment intent, growth level, and urgency degree. A small

Table 1 Heterogeneity of radiotherapy patients

Patients	Cancer type	Arrival rate	Number of fractions	WTT (days)
P1	Prostate	1.43	33	20
P2	Lung	0.18	2	5

example with two patient types is shown in Table 1, which can help to understand the heterogeneity of patients. The "Arrival Rate" column represents the average daily workloads, which can be very different. Patients within the same patient type have the same number of fractions, shown in the "Number of fractions" column. Note that we only consider identical LINACs here, because LINACs are becoming more and more advanced nowadays and are capable to treat almost all types of patients.

Due to the complexity of the system, no analytical results exist for such a system. This also explains why related work is limited and why most of the work gave myopic solutions or case-oriented solutions that cannot be applied to a more general case. In this paper, one of the contributions is solving the patient scheduling problem by proposing a class of "routing policies" that can automatically help managers to decide which patient to treat next once a time slot becomes available. This decision can be made several days ahead because the number of fractions a patient needs is prescribed in the treatment protocol before the treatment starts.

We model the treatment process as a queueing system with multi-server and multi-type queues. This is based on the "slot server" framework we proposed in Li et al. (2015), where time slots of a LINAC are considered as servers in parallel. Therefore, the number of fractions is nothing more than a patient's service time, the days he/she occupies the time slot. The patients who find all time slots are occupied will wait in their queues (based on patient types). This queueing model gives us the insight that we can solve the patient scheduling problem by deciding which queue to treat first. Hence we propose a "routing policy" to automatically assign the patient to the "right" time slot on a LINAC, based on the current system state, considering different WTT requirements. The fairness between patient types is also considered in the objective function.

A standard way to optimize routing policies is Dynamic Programming (DP). However, the problem is intractable in case of a large number of patient types. Although some approximation methods can be used, they are computationally very demanding, give little insight and are usually too complex to implement. Furthermore, this only works for the homogeneous arrival rate case. We propose a new type of policy, called the *Highest Waiting Index First (HWIF)* policy, which only needs to update the Waiting Index (WI) of each patient type and choose the queue with the highest WI to treat next. The WI is simply the sum of the longest waiting time in the queue and a certain priority factor. The priority factor of each queue is calculated by a simulation-based heuristic that optimizes the long-run performance of the system. The policy is efficient, simple and easy to understand.

Next we consider the situation of non-homogeneous arrival rates, which is hardly considered in the healthcare systems literature. Although the cancer incidence rates in an area might not change much, the arrival rates of radiotherapy patients of a cancer center often do. Typical reasons are changes in treatment, changes in patient preferences, the opening or closing of competing centers in the neighborhood, etc. Indeed, as is often the case in healthcare processes, fluctuations induced in the process are often bigger than those generated by external events. From discussions with practitioners, we learned that radiotherapy is no exception. The arrival rate of each patient type may not change in the same pattern. Some patient types may remain relatively constant arrival rates while some others change. Moreover, the changing patterns are usually unknown and difficult to forecast based on the historical data. We hence extend our policy to an adaptive one, called the *Adaptive Highest Waiting Index First (AHWIF)*. The priority factors are adjusted adaptively to absorb the effects caused by the changes in arrival rates. The construction of the adaptive policy is explained in Sect. 5. Again, the policy is efficient and easily understandable. In our numerical experiments, the impact of not using an adaptive policy under changing arrival rates will be illustrated by comparing the two types of policies under various scenarios.

2 Related literature

In this section, we first give a brief literature review on the patient scheduling problem in radiotherapy. Then some literature from other fields related to our model and methods are discussed.

Petrovic and Leite-Rocha (2008) proposed two simple heuristic algorithms to schedule patients forward from the first feasible start date (ASAP algorithm) or backward from the last feasible start date (due date) in order to minimize the total number of patients whose waiting times are longer than their WTTs. Conforti et al. (2008) constructed a series of deterministic integer programming models to schedule patients in a given waiting list in order to maximize the number of scheduled patients without breaching WTTs. Burke et al. (2011) also constructed a deterministic mathematical model with multiple hierarchical objectives. None of the above models is stochastic, and the optimization is made only based on current patients.

An exception is the work of Saure et al. (2012), in which the treatment process is modeled as a Markov Decision Process (MDP) to decide the optimal threshold of the patients that can be scheduled for each patient type per day. Their objective function involves the overall weighted waiting time. Due to the curse of dimensionality, the MDP model cannot be solved directly so that they employed the Approximate Dynamic Programming (ADP) approach. Legrain et al. (2015) also consider the uncertainty of arrivals. They developed a hybrid method combining a stochastic programming and an online optimization algorithm to maximize the utilization of resources.

The studies of routing algorithms in the healthcare field are quite limited, mostly around the dynamic bed allocation problem. However, the trade-off in the bed allocation problem is between refused admissions and overall occupation of beds, which

leads to a different model. The studies closest to ours actually can be found in call centers, which also have multi-type queues with different waiting time targets.

Leveraging on queueing theory, some asymptotically optimal routing policies are studied mostly in a heavy-traffic regime, when agent occupancies converge to 100%, see for example Gurvich and Whitt (2010); Ward and Armony (2013); Tezcan and Dai (2010). Their routing policies cannot be used for radiotherapy planning because they fail in cases with queues having low volumes and their objective does not involve WTTs.

A more relevant paper is Chan et al. (2014). In this paper, a routing policy, which is also index-based, was proposed. They defined the index function as an affine combination of customer waiting times and agent idle times, of which the coefficients are chosen to maximize the service level of the call center. The performance of the proposed policy is excellent in the simulation model. However, the results are very difficult to interpret.

An adaptive policy was put forward in the work of Legros et al. (2015), which considered a call center having single-skill inbound calls and an infinite amount of emails. They adaptively changed the number of agents reserved for calls to deal with the non-homogeneous arrival rates of calls. The structure of their adaptive policy resembles ours, but their problem setting is quite different.

The remainder is organized as follows. In Sect. 1, we define our model. In Sect. 4, we give the structures of various policies involved in the paper. In Sect. 5, we explain the method that we have used to decide the priority factors in the HWIF policy. In Sect. 6, the construction of the AHWIF policy is introduced. In Sect. 7, we discuss the results of several numerical experiments. Because there are no closed-form solutions for complex systems with different policies, all performance evaluations in our paper are carried out by simulation. A conclusion follows in Sect. 8.

3 Model

We model the treatment phase of radiotherapy as a queueing system with I queues and c identical slot servers in parallel. For simplicity, we use $\mathcal{I}$ to denote the set of patient types. Each queue is dedicated to a patient type. The patients within the same queue are treated in a First Come First Serve (FCFS) manner and will not leave the queue until they get treatment. This is due to the fact that the rates of no-show, abandonment, and cancellation are very low in radiotherapy.

Note that the arrivals in our model are the patients who become ready to treat. In reality, the pre-treatment phase usually needs at least 1 day to finish (the exceptions are emergency cases which are less than 1%), hence we can assume in the model that the new arrivals and also the newly released slot servers are observed at the beginning of the day. Afterward, the routing decisions are made for the idle slot servers, if any, according to some given policy π. This assumption will not affect the evaluation of the performance since we measure by days. From a practical point of view, the assumption also fits the way the planners work. We use $A_{i,n}$ to represent the arrival date of patient n of type i, and $S_{i,n}(\pi)$ to represent the date that the patient starts treatment (i.e.,

is assigned to an idle server). Then the access time (i.e., waiting time) of the patient is $S_{i,n} - A_{i,n}$.

The arrival rate of patient type i is assumed to follow a Poisson distribution with a mean λ_i. In the non-homogeneous situation, the arrival rates may change over time. As we discussed in the previous section, the changes in the arrival rates can be triggered by different events. In radiotherapy, instead of seasonality, we focus more on trend (i.e., a gradual change) and leap (i.e., a sudden change) which can be caused by new policies of doctors/ hospitals or adding LINAC resources. Service times are deterministic for each patient type, denoted by μ_i, which are equal to the number of fractions needed. The WTT of a patient type is denoted by ω_i.

Only work-conserving policies are considered in this paper, which implies that we do not allow servers to remain idle as long as there is any patient waiting. Although non-work-conserving policies can perform better in some specific situations, in the healthcare system, work-conserving policies are preferable because managers want to make the best use of the expensive equipment. Moreover, keeping capacity idle is not fair to the patients currently waiting in the queue. It only makes sense when we have emergency patients. However, less than 1% of radiotherapy patients are emergency patients who usually have very short service time (1–2 days). The capacity of LINACs is somewhat flexible so that we can always treat them using overtime.

3.1 Performance measures

In this subsection, we explain two important performance measures we have used to evaluate the system: the expected *tardiness* of each patient type and the *service level*.

The tardiness of a patient either is 0, which implies he/she started treatment within the required WTT or is the part of the waiting time that exceeds the required WTT. For every scheduled patient, his/her tardiness is calculated as $T_{i,n}(\pi) = \max\{S_{i,n}(\pi) - A_{i,n} - \omega_i, 0\}$. When we have a stable system with homogeneous arrival rates, the expected tardiness of type i patients under policy π can be measured by

$$\mathbb{E}[T_i(\pi)] = \lim_{N\to\infty} \frac{\sum_{n=1}^{N} T_{i,n}(\pi)}{N}. \tag{1}$$

By setting N large enough, this value can be well approximated in our simulation model. In the non-homogeneous situation, the system is no longer stable. We pay attention to the transient performance. The expected tardiness over a given time period d (e.g., 1 year) is given by

$$\mathbb{E}[T_i(\pi, d)] = \mathbb{E}\left[\frac{\sum_{n=1}^{N_d} T_{i,n}(\pi)}{N_d}\right]. \tag{2}$$

This value is estimated by averaging the results of numerous simulation runs. Note that N_d is the number of patients that have started treatment during the d days.

The service level is defined as $SL_i(\pi) = \mathbb{P}\{S_{i,n} - A_{i,n} \leq \omega_i\}$. We only consider it as an additional performance measure since it can be more intuitive for some managers than the tardiness. It is not part of our objective function due to its drawback: there is no incentive to treat patients once they waited longer than their WTT.

The service level for the homogeneous and non-homogeneous situations can be calculated by

$$SL_i(\pi) = \lim_{N\to\infty} \frac{\sum_{n=1}^{N} \mathbf{1}(S_{i,n} - A_{i,n} \leq \omega_i))}{N}, \tag{3}$$

and

$$SL_i(\pi, d) = \mathbb{E}\left[\frac{\sum_{n=1}^{N_d} \mathbf{1}((S_{i,n} - A_{i,n}) \leq \omega_i))}{N_d}\right], \tag{4}$$

respectively.

3.2 Objective function

Our paper aims at scheduling patients on LINACs more efficiently so as to meet the WTT requirements. Therefore, an objective function should represent how good the requirements are met. Later, the parameters (priority factors) in the WI-based policy are optimized according to the objective function. In the objective function, we consider patients' expected tardiness and assign a penalty weight p_i to each patient type. This is because the same length of tardiness should be penalized more for the patients having a shorter WTT. An intuitive setting is $p_i = 1/\omega_i$. Furthermore, we also want to consider the fairness between patient types. In other words, we don't want to sacrifice the performance of any patient type for the overall performance. As a result, the objective function is set to minimize the maximal weighted tardiness among all patient types, which can be defined as follows:

$$G(\pi) = \max_{\mathcal{I}} (p_i \cdot \mathbb{E}[T_i(\pi)]). \tag{5}$$

In prior studies, a (weighted) average of the tardiness/waiting time of all patient types is usually used as the objective function, which hides the different performances among patient types so that fairness cannot be considered. Instead of considering it as our objective function, we use it to show the overall performance which is useful information for managers. To distinguish with the objective function $G(\pi)$, we use $T_o(\pi)$ to represent the overall tardiness, defined as:

$$T_o(\pi) = \sum_{i=1}^{I} \left(\frac{\lambda_i}{\sum_{i=1}^{I} \lambda_i} \cdot p_i \cdot \mathbb{E}[T_i(\pi)]\right), \tag{6}$$

Note that $T_o(\pi)$ is not used to set any parameters. In numerical experiments, we will see the WI-based policy outperforms other benchmark policies in terms of both the

objective $G(\pi)$ and the overall performance $T_o(\pi)$. The expected tardiness and service level of each patient type are shown as well to understand more about fairness. In the non-homogeneous situation, we replace $\mathbb{E}[T_i(\pi)]$ by $\mathbb{E}[T_i(\pi, d)]$ so that $G(\pi)$ becomes $G(\pi, d)$ and $T_o(\pi)$ becomes $T_o(\pi, d)$.

4 Routing policies

In this section, we want to explain the structure of the WI-based policy. The WI of each patient type is the sum of the current longest waiting time in the queue and the priority factor. Interestingly, the benchmark policies which are commonly used in practice are special cases of the WI-based policy. Our WI-based policy is more flexible since the priority factors in the WI function are optimized for each given scenario. The algorithm we used to optimize the factors will be introduced in the next section. Here we only focus on the structures and how they are implemented. The abbreviations in the parentheses will be used later in the results of numerical experiments.

4.1 Global FCFS (GF)

In this simple and intuitive policy, all patient types are treated equally, so that the decisions are made only based on waiting times, independent of patient types. This policy can be easily implemented via a single FCFS queue: at the beginning of a day, we first update the queue by adding the new arrivals and then each idle server will pick the patient from the head of the queue to treat. In fact, it has poor performance when there are multi-type queues with different priorities (i.e., WTTs). However, we regard it as the basic benchmark policy since it is still often used in healthcare systems.

4.2 Static priority (SP)

In this policy, the routing decisions are made based on a static priority list of patient types. At the beginning of a day, we first add new arrivals to their corresponding queues. Then the assignment is done in the order of the priority. In our problem, some patient types may have the same priority levels. Within these patient types, we use the GF policy. According to the way we determine the static priority list, four different SP policies are considered. They are defined as follows.

Shortest WTT First (SWF): In healthcare systems, the static priority list is normally determined based on the urgency degree (i.e., WTTs). In the SWF policy, the patient type with the shortest WTT has the highest priority level.

Shortest Service First (SSF): This policy comes from the classic scheduling policy in which the patient type that needs the shortest service time has the highest priority level.

Shortest WTT & Service First (SWSF): This policy is an extension of the SWF policy. The static priority list is first determined based on WTT. Then the patient types with the same WTT will be further ordered with respect to their service times.

Shortest Service & WTT First (SSWF): This policy is an extension of the SSF policy. The static priority list is first determined based on the service time. Then the patient types that need the same service time are further ordered according to their WTTs.

Table 2 shows the priorities of three patients under different SP policies. Note that the smaller the number is, the higher the priority.

4.3 Earliest due date (EDD)

This policy is often used in manufacturing systems, where products need to be delivered before some due date fixed in the contract. However, in our problem, we only focus on the date a patient starts treatment. Therefore, a patient's due date is defined as the arrival date plus the WTT. To implement this policy, an idle server will scan the patient at the head of each queue and pick the one with the earliest due date to treat. If all patients' due dates are the same, the server will pick the one with the shortest WTT. Compared to the SWF policy which always gives the shorter WTT patients higher priority, the EDD policy is more flexible. It can mitigate the imbalance between patient types caused by different priorities. For instance, consider two patients P1 and P2 from Table 2 both waiting in the queue. P1 has waited only 1 day but P2 has already waited 10 days. If there is only one idle server, under the SWF policy, the server will pick P1 ignoring the fact that P2 cannot wait any longer. However, under the EDD policy, P2 will be picked because his/her due date is earlier (4 days earlier than P1).

4.4 Highest waiting index first (HWIF)

In WI-based policies, routing decisions are made according to the current waiting index (WI) of each queue. We let X_i denote the WI of patient type i, which is defined as:

$$X_i = W_i^f + \gamma_i. \tag{7}$$

W_i^f is the current longest waiting time of queue i, which is the waiting time of the first patient in queue i. If the queue is empty, we let $W_i^f = -\infty$. γ_i is a priority factor assigned to queue i. Under the HWIF policy, an idle server will scan the patient at the head of each queue and pick the one with the highest WI to treat. In case of a tie, the server will pick the one with the largest γ_i.

Table 2 Comparison of the SP policies

	WTT (days)	Service time (days)	SWF	SSF	SWSF	SSWF
P1	5	16	1	2	1	2
P2	10	16	2	2	3	3
P3	10	2	2	1	2	1

The essential idea underlying the structure of our WI is that using the W_i^f to indicate the current urgency of queue i and making an adjustment by adding a priority factor γ_i so as to consider various WTT requirements. On average, the larger the γ_i is, the less the patient of type i needs to wait to start treatment. For example, we again consider two patients P1 and P2 from Table 2 both waiting at the head of each queue. We assume $\gamma_1 > \gamma_2, \gamma_1 - \gamma_2 = 3$, then P2 can start the treatment first if he/she has waited 4 days longer than P1 has.

In the homogeneous situation, the arrival rates are assumed constant and given (e.g., estimated using historical data). Therefore, in any given scenario, we can determine γ_i for each patient type i so as to minimize the objective function (5). Once the priority factors are determined, the HWIF policy can be easily implemented to schedule patients. Managers just need to update the current WI of each queue every time they have made a routing decision. However, due to the complexity of the system, we cannot derive a closed-form solution of γ_i. Instead, the calculation of γ_i is carried out by a simulation-based algorithm, which will be discussed in Sect. 4. We find that γ_i is influenced by the configuration of all patient types' workloads (i.e., service time and arrival rates) and WTT requirements. This also explains why we need to adjust γ_i adaptively when we have non-homogeneous arrival rates because the configuration of patient types' workloads will change.

We already mentioned that the benchmark policies mentioned above fit in the framework of the WI. Their γ_i and $X_i = W_i^f + \gamma_i$ are shown in Table 3, where M is a large number (e.g., 1000); $\mu_i' = \max_{\mathcal{I}}\{\mu_i - \mu_j\}$; $\omega_i' = \max_{\mathcal{I}}\{\omega_i - \omega_j\}$ and γ_i^* represents the optimal γ_i that minimized the objective function (5). This property also makes the explanation of γ_i^* (e.g., to managers) much easier.

5 Determination of the priority factors

In this section, we introduce the approaches used to determine the priority factors γ_i in the HWIF policy. As we mentioned, the priority factors should be calculated to minimize the objective function $G(\pi)$. We use $\boldsymbol{\gamma}$ to represent a solution. Firstly, two properties of $\boldsymbol{\gamma}$ need to be considered.

(1) Due to the structure of the HWIF policy, an idle server will choose the patient with the highest WI to treat. Therefore, the routing decisions are actually affected by the relative differences of γ_i between patient types. It implies that we only need to optimize $I - 1$ dimensions of the decision vector $\boldsymbol{\gamma}$ because we can always fix a certain γ_i (e.g., to 0).

Table 3 Special cases of the HWIF policy

	HWIF	GF	SWF	SSF	SWSF	SSWF	EDD
γ_i	γ_i^*	0	$\omega_i' M$	$\mu_i' M$	$(\omega_i' M + \mu_i')M$	$(\mu_i' M + \omega_i')M$	ω_i'
X_i	$W_i^f + \gamma_i^*$	W_i^f	ω_i'	μ_i'	$\omega_i' M + \mu_i'$	$\mu_i' M + \omega_i'$	$W_i^f + \omega_i'$

(2) Since the waiting time W_i^f is counted in days, it makes little sense to set γ_i as a fraction. Hence γ_i is considered to be integer.

The objective value under any given $\boldsymbol{\gamma}$ is evaluated via simulation. Then a simulation-based optimization (SIMOPT) method can be employed to derive the optimal or near optimal solution. However, general-purpose SIMOPT methods are slow. To increase speed and accuracy we have developed heuristics. In Sect. 5.1, six scenarios with only two patient types are analyzed. The gained numerical insights play an important role in developing the heuristic algorithm which is explained in Sect. 5.2. Later, the insights also help us to construct the adaptive policy.

5.1 Numerical insights

Due to the properties of $\boldsymbol{\gamma}$, when we consider a scenario with 2 patient types, the solution can be simplified to one integer variable: $\Delta\gamma = \gamma_2 - \gamma_1$. To understand how $\Delta\gamma$ affects the objective value, we evaluate $G(\Delta\gamma)$ for any $\Delta\gamma \in [-\max(\omega_1 + \mu_1, \omega_2 + \mu_2), \max(\omega_1 + \mu_1, \omega_2 + \mu_2)]$ with brute force, where $G(\Delta\gamma)$ is the objective value under the HWIF policy with $\Delta\gamma$.

Table 4 lists the scenarios considered. To compare the results of various scenarios, the number of servers c is set to achieve similar traffic loads, shown in column ρ, so that the influence caused by different traffic loads can be eliminated. The last two columns give the optimal solution $\Delta\gamma^*$ and the value of the objective function $G(\Delta\gamma^*)$.

In Fig. 1, we show the weighted tardiness $p_1\mathbb{E}(T_1(\Delta\gamma))$, $p_2\mathbb{E}(T_2(\Delta\gamma))$ and also $T_o(\Delta\gamma)$ of Scenario 1–4. The points are only connected to show the trend clearly. As we mentioned in Sect. 4.4, other benchmark policies can be regarded as special cases of the HWIF policy. In the plots, the performances under the GF policy and the EDD policy can be easily shown. Two vertical dashed lines with different colors are used to point out the corresponding $\Delta\gamma$ (GF: $\Delta\gamma = 0$, EDD: $\Delta\gamma = \omega_1 - \omega_2 = -5$). The corresponding performances are the values of the intersections of the three lines and the dashed lines.

The following numerical insights are gained through the analysis of various scenarios.

Table 4 Scenarios of model with 2 patient types

Scenarios	$(\mu_1/\omega_1/\lambda_1)$	$(\mu_2/\omega_2/\lambda_2)$	c	ρ (%)	$\Delta\gamma^*$	$G(\Delta\gamma^*)$
S1	5 / 5 / 1.2	5 / 10 / 1.2	13	92.30	− 6	0.0125
S2	20 / 5 / 0.12	10 / 10 / 1.2	16	96	− 7	0.02
S3	20 / 5 / 1.2	10 / 10 / 0.12	26	96.90	− 14	1.28
S4	5 / 5 / 0.12	20 / 10 / 1.2	26	94.50	− 10	0.13
S5	20 / 5 / 1.2	18 / 10 / 0.12	28	92.30	− 7	0.24
S6	20 / 5 / 1.2	5 / 10 / 0.12	26	92.30	− 11	0.47
S7	20 / 10 / 0.12	5 / 10 / 1.2	10	93.30	0	0.01

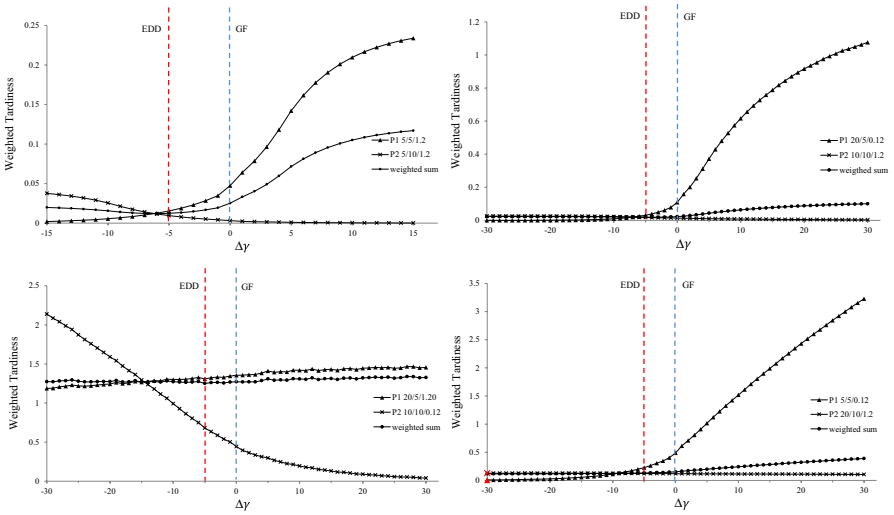

Fig. 1 Comparison of scenarios with 2 patient types: Scenario 1–Scenario 4

Observation 1: The expected weighted tardiness of patient type i is non-increasing in γ_i.

Indeed, in Fig. 1, we find that the expected weighted tardiness of patient type 1 (patient type 2) in all scenarios decreases as $\boldsymbol{\Delta\gamma}$ decrease (increase). However, we only can prove this property in 2 specific situations: identical service times or $c = 1$.

Theorem 1 *If patients have identical service times then the expected weighted tardiness of patient type i is non-increasing in* γ_i.

Proof For a given $\boldsymbol{\gamma}$ consider an arbitrary sample path $A_{i,n}$ of the arrivals. Because the service times are constant and deterministic this fully determines the WI and the order and times in which the patients are treated. Now suppose we increase γ_i. Because the policies are work-conserving the number of patients treated is the same for any time, only the order can change. For any type i patient, due to the form of the WI, patients that are later in the order will never get a WI higher than the type i patient. Thus all type i patients keep their position or move to an earlier position. Therefore they are all treated at the same time or earlier. □

Theorem 2 *If there is only 1 server in the system then the expected weighted tardiness of patient type i is non-increasing in* γ_i.

Proof Consider again an arbitrary sample path $A_{i,n}$ of the arrivals. Because the policies are work-conserving, the amount of work done in the system is the same under any policy. Using the same argument as for Theorem 1 we conclude that type i patients can only move forward and therefore their waiting times can only get shorter. □

Unfortunately, it is impossible to use the sample path method above to give a proof for the situation with both unequal service times and multiple servers. A counterexample can be easily found, see Fig. 2. On day 3, server 2 becomes idle so that a routing decision needs to be made. In the queues, there are 3 patients: P1, P2, and P3 who have waited 1 day, 2 days and 0 days respectively. P1 belongs to patient type 1 with $\mu_1 = 2$ and the other two patients are from patient type 2 with $\mu_2 = 3$. On day 3, $X_1 = 1 + \gamma_1$ and $X_2 = 2 + \gamma_2$.

In the first case, we assume $\gamma_1 = 3, \gamma_2 = 1$ so that $X_1 = 4, X_2 = 3$. Hence P1 is picked first to start the treatment. On day 4, server 1 becomes idle, then P2 will start the treatment. Since P1 will occupy server 2 for 2 days, P3 will start the treatment on day 5. However, in the second case, we increase γ_2 and let $\gamma_1 = \gamma_2 = 3$, which leads to $X_1 = 4, X_2 = 5$. Hence this time, the routing decisions are P2 → P1 → P3. One sees that P3 will start the treatment on day 6, which is even later than he/she does in the first case.

Observation 2: The rate at which the expected weighted tardiness changes with γ_i is different for each patient type. The minimum of the objective function $G(\boldsymbol{\gamma})$ is obtained when all $p_i T_i(\boldsymbol{\gamma})$ are similar.

This observation is based on the comparison of the plots, which show quite different ranges of the expected weighted tardiness based on the same range of $\Delta\boldsymbol{\gamma}$. However, the optimal solution always has similar values of $p_i T_i(\boldsymbol{\gamma})$.

Observation 3: The expected weighted tardiness of each patient type is bounded, and the upper/lower bound can be derived by applying the SP policy with the lowest/highest priority.

This is because we only consider work-conserving policies. In plot 1 and plot 2 of Fig. 1 one can see that the weighted tardiness of patient type 1 increases to an upper bound. In fact, the weighted tardiness of patient type 1 in plot 3 and plot 4 also converges to an upper bound but at a much slower rate, outside of the plot.

Observation 4: The EDD policy outperforms the GF policy in all scenarios with different WTTs, and is close to the optimal solution.

Under the GF policy, all patient types are treated equally, having the same priority factors. It leads to poor performance of the patients with a shorter WTT. On the contrary, the γ_i in the EDD policy is set based on the WTTs. Due to its

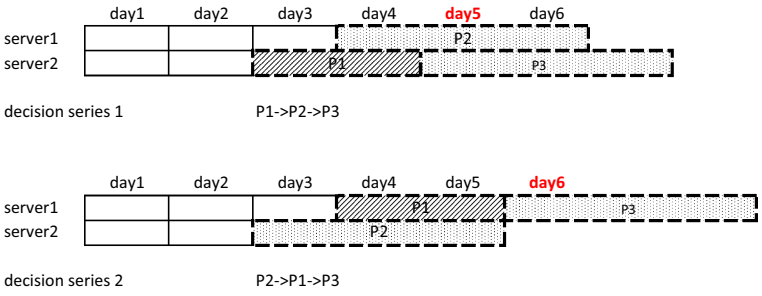

Fig. 2 Comparison of scenarios of 2 patient types

good performance, we choose the EDD policy as the starting point (i.e., the initial solution) to search for the optimal $\boldsymbol{\gamma}^*$ in the algorithm.

Observation 5: When patients have identical WTT, we observed that the optimal solution of $\boldsymbol{\gamma}^*$ is $\mathbf{0}$, which implies that there is no difference between applying the EDD, GF and HWIF policies. However, for different WTTs, the optimal $\boldsymbol{\gamma}$ is no longer $\mathbf{0}$ and it depends on both the service times and the arrival rates.

An example of having identical WTT is Scenario 7 in Table 4. Although the attributes of the patient types are different, the optimal solution gives them the same priority factor. This is due to the objective function $G(\boldsymbol{\gamma})$: a penalty occurs only if the waiting time exceeds the WTT; instead of the weighted average tardiness, the maximal tardiness among all patient types is minimized. However, in reality, various WTTs need to be considered. Then the optimal solution is no longer $\mathbf{0}$. Scenario 5 and Scenario 6 are compared to show how the service time influences the solution: the service time of patient type 2 is different in two scenarios so that the corresponding optimal solution is different as well. Similarly, the arrival rate influence can be found by comparing Scenario 2 and Scenario 3.

Observation 6: If we choose the weighted average tardiness $T_o(\boldsymbol{\gamma})$ as the objective function, we can get very unfair performances between patient types.

In the last plot of Fig. 1, we see that the weighted average tardiness remains small while the tardiness of patient type 1 becomes very large. Vice versa, the solution $\boldsymbol{\gamma}^*$ that minimizes $G(\boldsymbol{\gamma})$ still gives a low $T_o(\boldsymbol{\gamma})$. This is why we think G is the preferred objective function: it is fair and also gives a low overall tardiness.

5.2 A hill-climbing algorithm

Combining the above numerical insights, a hill-Climbing-based local search algorithm is developed. The algorithm needs four input parameters: an initial solution $\boldsymbol{\gamma}^{init}$, the patient type with the priority factor fixed i^{fix}, a step size α and tolerance e. As we mentioned in Observation 4, we choose the priority factors of the EDD policy as the initial solution: $\gamma_i^{init} = \omega_i'$, see Table 3. Due to the first property of $\boldsymbol{\gamma}$, we need to fix the priority factor of a certain patient type: i^{fix}. To avoid having any negative γ_i in the end, we let $i^{fix} = \arg\min_{\mathcal{I}} p_i T_i(\mathbf{0})$. As we known, $\boldsymbol{\gamma} = \mathbf{0}$ implies treating all patient types equally. Hence for the patient type with the minimal weighted tardiness (i.e., the best performance), there is no need to assign any additional priority factor, compared to other patient types. Due to the second property of $\boldsymbol{\gamma}$, the step size α is set to 1 and the tolerance e is a very small value (e.g., 0.0001) used to deal with the noise in the simulation.

To search for the optimal solution, we first evaluate each patient type's weighted tardiness under the current solution, then we find the patient type with the worst performance i^{max} to improve. In every step, we add 1 to $\gamma_{i^{max}}$. The procedure continues until $i^{max} = i^{fix}$. During the search, we record the "current top" of the "hill" to compare with the next "top" and keep the better one (i.e., the minimal one). In the end, we output the "current top".

Input: $\boldsymbol{\gamma}^{init}$, i^{fix}, step size $\alpha = 1$, tolerance e
Output: $\boldsymbol{\gamma}^*$, $G(\boldsymbol{\gamma}^*)$
Initialization: $\boldsymbol{\gamma}^1 = \boldsymbol{\gamma}^{init}$, $i^{max} = \arg\max_{\mathcal{I}}\{p_i T_i(\boldsymbol{\gamma}^1)\}$, $G(\boldsymbol{\gamma}^*) = -\infty$;
for $n = 2; i^{max} \neq i^{fix}$ **do**
  $\gamma^n{}_i = \gamma^{n-1}{}_i + \alpha$, for $i = i^{max}$;
  $\gamma^n{}_j = \gamma^{n-1}{}_j$, for $j \neq i$;
  $i^{max} = \arg\max_{\mathcal{I}}\{p_i T_i(\boldsymbol{\gamma}^n)\}$;
  if $G(\boldsymbol{\gamma}^n) > G(\boldsymbol{\gamma}^{n-1}) + e$ **then**
    if $G(\boldsymbol{\gamma}^{n-1}) < G(\boldsymbol{\gamma}^*) - e$ **then**
      $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^{n-1}$;
    end
  end
  $n = n + 1$;
end
return $\boldsymbol{\gamma}^*$, $G(\boldsymbol{\gamma}^*)$;

Algorithm 1: Hill-climbing-based local search algorithm

To verify the algorithm, we first apply the algorithm to the scenarios in Table 4. The same optimal solution was found in all scenarios within several minutes. To further analyze the efficiency of the algorithm, we apply it in a scenario with 7 patient types (PT1-PT7), which are chosen from a real scenario. Figure 3 gives the search path of the proposed algorithm, where the red line connects the solutions that have been recorded. In Table 5, we show the expected weighted tardiness of each patient type, the objective value $G(\pi)$, and the γ_i under each policy. It is obvious that the HWIF policy outperforms the other two policies. The detailed information of each patient type can be found in Table 6, marked in bold.

6 The construction of the adaptive policy

In this section, we will discuss the adaptive policy, which is developed for the non-homogeneous situation. As we mentioned, not all arrival rates change in the same pattern. Sometimes, the change of the arrival rates can be gradual, which is called a

Table 5 Results of the local search algorithm

	PT1	PT2	PT3	PT4	PT5	PT6	PT7	$G(\pi)$
GF	0.1342	0.1337	0.0005	0.0005	0.5488	0.5467	0.005	0.5488
γ_i	0	0	0	0	0	0	0	–
EDD	0.0046	0.0045	0.0013	0.0012	0.0090	0.0088	0.0012	0.0090
γ_i	30	30	0	0	35	35	0	–
HWIF	0.0015	0.0015	0.0015	0.0015	0.0014	0.0014	0.0015	0.0015
γ_i	38	38	0	0	49	49	0	–

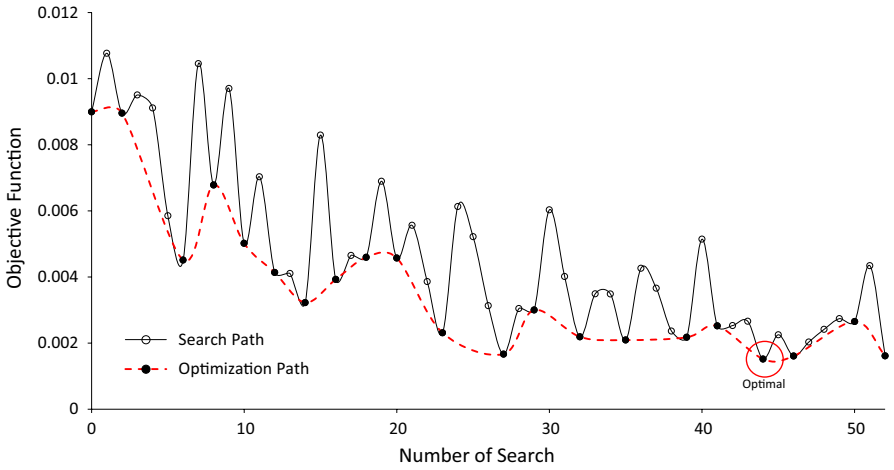

Fig. 3 Search paths of the algorithm

Table 6 Patient information

Patient type	Cancer description	Urgency situation	λ_i	μ_i	ω_i
1	Lung	Radical	0.19	5	5
2		Urgent	0.11	4	5
3		Palliative	0.11	1	5
4	Head and neck	Radical#1	0.29	20	10
5		Radical#2	0.21	35	10
6	Breast	Adjuvant#1	1.43	16	40
7		Adjuvant#2	0.59	20	40
8		Adjuvant#3	0.45	12	40
9		Palliative#1	1.6	1	40
10		Palliative#2	1.36	5	40
11		Non-urgent#1	0.57	10	40
12		Non-urgent#2	0.38	4	40
13		Non-urgent#3	0.18	15	40
14	Prostate	Radical#1	0.34	33	40
15		Radical#2	0.44	37	40

trend. A sudden change can also happen, for example, some patients from a partner hospital may be distributed to the hospital because there is a new LINAC. In our paper, we call the latter case a leap. According to Observation 5, the changes in arrival rates will lead to a different $\boldsymbol{\gamma}^*$ in the HWIF policy. We could calculate the new $\boldsymbol{\gamma}$ for every new situation, but, in reality, the changes are often unknown and difficult to forecast. Therefore, an adaptive policy is more appropriate.

Our adaptive policy is a generalization of the HWIF policy. It is also easy to use and simple to explain. We use the same WI structure. We start with calculating $\boldsymbol{\gamma}$ using Algorithm 1 based on historical data. When the arrival rates changes, the $\boldsymbol{\gamma}$ is

no longer optimal. Hence we need to adjust it to adapt to the changes in the arrival rates. Inspired by Observation 1 and the results of the HWIF policy, the idea of the AHWIF policy is adding a small factor h to the patient type with the worst performance. This is done every day, and the performance is calculated on the basis of the last v days.

Now let's focus on how to set the value of v and h. We assume both of them are constant. Unfortunately, choosing v and h is very difficult because the rate of changes in the arrival rates are unknown. Instead, we propose several combinations of v and h and later test them under various scenarios. The numerical experiments show that a reasonable combination of v and h already can lead to very good results.

To derive such a combination of v and h, we should understand the relationship between the parameters and the performance. Firstly, the longer the v is, the more difficult capturing the changes in the system becomes, including both the change in arrival rates and the change caused by our adaptive policy. However, if the v is set too short, the re-evaluated performance is less reliable because of the high variability. Hence we suggest three options: $v = 10$, $v = 30$, and $v = 90$. Secondly, h implies the amount of adjustment every time. Different from the step size $\alpha = 1$ that have defined in the local search algorithm, we allow h to be a fraction. Although a fractional increment in γ_i will not affect the routing decisions immediately due to the second property of $\boldsymbol{\gamma}$, a fractional increment can help avoid too much changes in the policy. For example, if we set $h = 0.5$, a patient type needs to have the worst performance among all patient types continuously for 2 days to get a higher priority than before (i.e., its γ_i increased by 1 and others remain the same). Moreover, a large h (e.g., $h > 1$) can increase the fluctuation in the performance of each patient type. However, a too small h is also not preferred because it cannot properly respond to the changes in arrival rates. As a result, three options are considered: $h = 0.1$, $h = 0.5$, and $h = 1$. In reality, managers can choose a combination of v and h based on the understanding of the system. For instance, if they understand there will be a sudden increase in the number of patients referred to their hospital, $v = 10$ and $h = 1$ can be a good option since a quick response to the change is needed. We found that the performance of the AHWIF is relatively robust to the choice of v and h.

7 Numerical experiments

7.1 Experimental setting

To test the efficiency of the policies for realistic parameters, a series of numerical experiments are conducted based on a scenario with 15 patient types. The information about the patient types is provided in Table 6, which is derived from the work of Saure et al. (2012). Three different WTTs are assigned to different patient types, and the workloads of patient types are asymmetric. We regard the patient types which share the same cancer position (e.g., lung, prostate, etc.) as a group. For example, the breast group refers to patient type 6-13. The number of slot servers is given as 104 so the total traffic load equals to 96.68%.

Our simulation model is built in C++. We set the confidence level to 97%. Moreover, we also use the Common Random Numbers Method to compare different policies more efficiently. In the homogeneous situation, we simulated 10^7 days with 10^3 days warm up. In the non-homogeneous situation, the focus is on the weighted expected tardiness over d days. We set $d = 1000$ and run the simulation model 10^4 times with different seeds to get a reliable performance. We also run the simulation 10^3 days first with homogeneous arrival rates to avoid it starts from an empty system.

7.2 Results under homogeneous arrival rates

We first compare the performance of the HWIF policy with the benchmark policies mentioned in Sect. 3. The expected weighted tardiness of each patient type, $p_i\mathbb{E}(T_i(\pi))$, the maximal weighted tardiness, $G(\pi)$ and the weighted average tardiness, $T_o(\pi)$, are shown in Table 7.

From the results, we can easily observe that the HWIF policy outperforms all the benchmark policies. Moreover, the HWIF policy provides a very balanced performance. We first look at the overall performance of the system: $T_o(\pi)$. By applying the HWIF policy, it has been reduced by 98.5% and 44.4%, compared to the GF policy and the EDD policy respectively. Among the four different SP policies, the SWSF policy gives the best performance. It implies, in our scheduling problem, that the WTT constraints are more important than the service time. However, the comparison between the SWSF policy and the SWF policy shows that the service time should also be taken into consideration. The EDD policy is the top one in the

Table 7 Weighted tardiness under various policies

Patient Type	GF	SSF	SWF	SSWF	SWSF	EDD	HWIF
1	0.5111	0	0	0	0	0.0043	0.0003
2	0.5106	0	0	0	0	0.0044	0.0003
3	0.5109	0	0	0	0	0.0044	0.0003
4	0.1216	0.0019	0	0.0010	0	0.0022	0.0004
5	0.1209	0.0741	0	0.0743	0	0.0022	0.0004
6	0.0005	0	0.0033	0	0	0.0006	0.0005
7	0.0005	0	0	0	0	0.0006	0.0005
8	0.0005	0	0	0	0	0.0006	0.0005
9	0.0005	0	0.0456	0	0	0.0006	0.0005
10	0.0005	0	0.0020	0	0	0.0006	0.0005
11	0.0005	0	0	0	0	0.0006	0.0005
12	0.0005	0	0	0	0	0.0006	0.0005
13	0.0005	0	0	0	0	0.0006	0.0005
14	0.0005	0	0	0	0	0.0006	0.0005
15	0.0005	0.0222	0	0.0223	0.0243	0.0006	0.0005
$G(\pi)$	0.5111	0.0741	0.0456	0.0743	0.0243	0.0044	0.0005
$T_o(\pi)$	0.0332	0.0031	0.0098	0.0031	0.0013	0.0009	0.0005

benchmark policies. Now we focus on the maximal weighted tardiness. Compared to the GF policy and the EDD policy, $G(\pi)$ has been reduced by 99.9% and 88.6%. As we mentioned before, we do not want to sacrifice the performance of any patient type to achieve a beautiful overall performance. For example, under the SWSF policy, only the last patient type has a very bad performance. By minimizing $G(\pi)$, the results are more balanced so that the fairness between patient types is considered.

Since the service level of each patient type $SL_i(\pi)$ can be an important performance measure for managers, we show the results in Table 8. Similar to $T_o(\pi)$, we also calculated the weighted average service level $SL_o(\pi)$ to show the overall performance in terms of the service level. We find the HWIF policy also gives an excellent performance. The rank of the policies is almost the same as the one considering tardiness. The only difference is that the SWSF becomes the best policy, even slightly better than the HWIF policy, if we only look at the overall service level. This is due to the fact that the service level only offers the probability of breaching the WTT requirements. It cannot tell us how long the patients have waited after the WTT. Under the SWSF policy, again the last patient type is sacrificed.

In Fig. 4, the performance of each patient type is compared one by one under the EDD policy and the HWIF policy. For any patient type, both performances are better when the HWIF policy is implemented. The service levels of patient type PT1-PT5 are higher than other patient types. This is because no weight is assigned to the service level.

Except for the benchmark policies, it is also interesting to compare the proposed HWIF policy with the ones suggested by the previous work. As we mentioned before, our experiments are based on the scenario from the work of Saure et al. (2012), which also considered radiotherapy scheduling problem and

Table 8 Service level under various policies

Patient type	GF (%)	SSF (%)	SWF (%)	SSWF (%)	SWSF (%)	EDD (%)	HWIF (%)
1	64.42	100.00	100.00	100.00	100.00	99.68	99.96
2	64.51	100.00	100.00	100.00	100.00	99.68	99.96
3	64.46	100.00	100.00	100.00	100.00	99.68	99.97
4	83.26	99.31	100.00	99.65	100.00	99.68	99.91
5	83.33	88.98	100.00	88.95	100.00	99.68	99.92
6	99.77	100.00	98.83	100.00	100.00	99.65	99.67
7	99.77	100.00	100.00	100.00	100.00	99.65	99.68
8	99.77	100.00	100.00	100.00	100.00	99.65	99.67
9	99.77	100.00	93.22	100.00	100.00	99.65	99.67
10	99.77	100.00	99.28	100.00	100.00	99.64	99.67
11	99.77	100.00	100.00	100.00	100.00	99.65	99.67
12	99.77	100.00	100.00	100.00	100.00	99.65	99.67
13	99.78	100.00	100.00	100.00	100.00	99.66	99.68
14	99.77	100.00	100.00	100.00	99.99	99.65	99.68
15	99.77	95.20	100.00	95.30	94.91	99.65	99.68
$SL_o(\pi)$	97.02	99.44	98.36	99.46	99.73	99.65	99.70

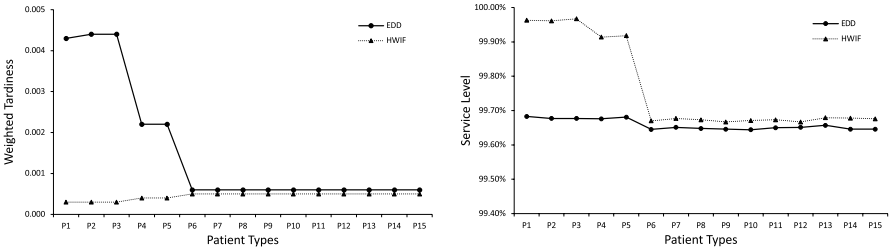

Fig. 4 EDD versus HWIF: weighted tardiness (left), service level (right)

proposed an ADP based policy. Although their problem setting is slightly different from ours, a reasonable comparison can be conducted in the following way. Their numerical results show that all patient types can start their treatment within 15 workdays with 111 slots. However, only around 80% of patients from the first three patient types can receive the first treatment within 5 workdays. Note that the number of slots needed (i.e., 111) is calculated by removing the average number of new start patients every day, which can be approximated by the overall arrival rate in the steady system, from the original number of slots needed in total (i.e., 120 slots). This is because they have assumed, different from us, that the first treatment of each patient needs one more slot. For comparison, we set the WTT of patient type 6–15 to 15 workdays and others remain the same. The results of the SL by applying the proposed HWIF policy are shown in Table 9, where one can observe that, with only 108 slots, all patients can start their treatment within

Table 9 Service level comparison between HWIF and ADP

Patient type	WTT	HWIF (%)			ADP (%)
		c = 106	c = 107	c = 108	c = 111
P1	5	99.9	100	100	81
P2	5	99.9	100	100	84
P3	5	99.9	100	100	80
P4	10	99.8	99.9	100	95
P5	10	99.8	99.9	100	94
P6	15	99.6	99.9	100	100
P7	15	99.6	99.9	100	100
P8	15	99.6	99.9	100	100
P9	15	99.6	99.9	100	100
P10	15	99.6	99.9	100	100
P11	15	99.6	99.9	100	100
P12	15	99.6	99.9	100	100
P13	15	99.6	99.9	100	100
P14	15	99.6	99.9	100	100
P15	15	99.6	99.9	100	100

15 workdays, moreover, patients from patient type 1–3 (4–5) can start within 5 (10) workdays.

In the end, we want to show the robustness of the proposed HWIF policy by testing it in different environments.

Firstly, we change the extent of similarity and difference between patient types in terms of their WTTs, arrival rates and service times respectively to see how the variety of patient types affects the results. In specific, for each attribute, we structure 3 different cases: Same, Low, and High. For example, considering the arrival rate, "Same" represents all patient types have the same arrival rate equals to the average (i.e., 0.55 in our case); "Low" represents low variety which is achieved by reducing(increasing), by 20%, the arrival rates of the patient types which are higher(lower) than the average; "High" represents high variety which is achieved by increasing(reducing), by 20%, the arrival rates of the patient types which are higher(lower) than the average. Again, we let all scenarios have similar traffic loads by adjusting the number of slots so that the effect of the system load is small. The maximal weighted tardiness in each case is shown in Table 10. One can easily find that the HWIF policy outperforms other benchmark policies in all cases. The performance in terms of the overall weighted tardiness gives the same observation, which is not shown here for brevity. Another interesting observation is that it is more important to use the proposed HWIF policy in cases with high variety, especially when we have various WTTs. When all patient types have the same WTT, the improvement gained by using the HWIF policy is not that obvious. The EDD policy also becomes the same as the GF policy. In reality, large hospitals or cancer centers often tend to have more variety in patient types which needs advanced scheduling policies (e.g., the HWIF policy) more. Moreover, although they have more capacity than small institutions, they also have higher arrival rates so that often higher traffic loads. Next, we want to discuss the effect of the overall system workload.

In Table 11, we give both the maximal weighted tardiness and the overall tardiness under different traffic loads. With the same patient settings, when we have more slot servers, the overall system traffic load is lower. The HWIF policy is still the top choice, however, the differences in performance between policies become less with the increase of slot servers. In other words, when the capacity is sufficient, the

Table 10 $G(\pi, d)$ with different variety of patient types

	Arrival rate			Service time			WTT		
	Same	Low	High	Same	Low	High	Same	Low	High
GF	0.4745	0.5324	0.4569	0.4511	0.6987	0.4280	0.0362	0.4974	1.1256
SSF	0.2178	0.2051	0.3411	0.1041	0.4917	0.1785	1.2638	0.5854	0.1958
SWF	0.0158	0.7276	0.5561	0.0935	1.0444	0.4380	2.2777	1.1617	0.4904
SSWF	0.2093	0.2146	0.3264	0.0924	0.4234	0.1762	1.2300	0.5733	0.1944
SWSF	0.2349	0.2034	0.3313	0.0991	0.4415	0.1896	1.2433	0.5942	0.2005
EDD	0.0072	0.0047	0.0049	0.0030	0.0147	0.0026	0.0362	0.0293	0.0034
HWIF	0.0009	0.0007	0.0002	0.0005	0.0012	0.0002	0.0350	0.0060	0.0001

Table 11 $G(\pi, d)$ and $T_o(\pi, d)$ under different traffic loads

	c=103		c=104		c=105	
	$G(\pi, d)$	$T_o(\pi, d)$	$G(\pi, d)$	$T_o(\pi, d)$	$G(\pi, d)$	$T_o(\pi, d)$
GF	2.3323	0.2005	0.3558	0.0298	0.1721	0.0100
SSF	1.5886	0.0922	0.1102	0.0056	0.1121	0.0066
SWF	4.8671	1.0433	0.0679	0.0164	0.1344	0.0289
SSWF	1.4289	0.0834	0.1105	0.0054	0.1100	0.0063
SWSF	1.4334	0.0765	0.0509	0.0027	0.0630	0.0034
EDD	0.3600	0.0701	0.0036	0.0035	0.0000	0.0000
HWIF	0.0422	0.0402	0.0033	0.0030	0.0000	0.0000

improvement could be gained by applying a better policy is small. Although adding capacity can easily improve the system performance, however, in reality, the medical capacity is often costly so that high utilization is often the case. Therefore, it is essential to develop effective policies.

7.3 Results under non-homogeneous arrival rates

Finally, the non-homogeneous situation is considered. Numerical experiments are carried out to discuss the efficiency of the AHWIF policy. Especially, we want to look into two interesting questions: how much the performance can be improved, by implementing the AHWIF policy instead of the HWIF policy in the non-homogeneous situation, and how the system performance changes along with the changes in arrival rates. The first question is studied by comparing the system performance over $d = 1000$ days. The second question is studied by looking at the system performance of every 100 days during the 1000 days. Considering the real situation in radiotherapy department, 5 changing patterns of arrival rates are designed.

Increasing Trend (IT): We consider a patient group (i.e., the breast group) having an increasing trend, which means their arrival rates are increasing slowly over several years.

Increasing and Decreasing Trend together (IDT): This pattern is designed to test the AHWIF policy the ability to handle an increase and a decrease in arrival rates.

Leap (L): The arrival rate of a patient group increases quickly (or suddenly). It can be caused by transferred patients.

Leap caused by additional capacity (L^+): A sudden increase in arrival rates can also be caused by introducing more capacity. Compared to L, L^+ will not increase the traffic load ρ much.

Leap and Drop (LD): This pattern is designed to simulate multiple changes. The arrival rate of a patient group increases while another group's decreases.

In Sect. 6, several pairs of (v, h) were proposed. We first test different parameters and patterns in a toy scenario. Two patient types are considered: P1($\mu_1 = 20/\omega_1 = 5/\lambda_1 = 1.2$) and P2($\mu_2 = 10/\omega_2 = 10/\lambda_2 = 1.2$). We set $c = 38$ so $\rho = 94.7\%$. To understand how the arrival rates change, see Table 13. However, in this toy scenario, we assume only the arrival rate of P1 changes, except for the

Table 12 $G(\pi, d)$ under different settings of (v, h)

Scenario	$v = 10$			$v = 30$			$v = 90$		
	h = 0.1	h = 0.5	h = 1	h = 0.1	h = 0.5	h = 1	h = 0.1	h = 0.5	h = 1
IT	0.27	0.25	0.27	0.24	0.27	0.27	0.28	0.29	0.30
IDT	0.23	0.21	0.24	0.22	0.23	0.25	0.24	0.26	0.28
L	15.47	14.02	13.99	15.38	14.06	14.07	15.59	14.35	14.40
L^+	1.26	1.21	1.23	1.29	1.25	1.26	1.30	1.29	1.30
LD	2.65	2.62	2.61	2.66	2.64	2.64	2.69	2.70	2.70

Table 13 How arrival rates change in each pattern

Patterns	Details
IT	λ_b gradually increases by 20% over 1000 days
IDT	λ_b gradually increases by 10% over the first 500 days and then gradually decreases by 10%
L	λ_b suddenly leaps to 120% after 300 days
L^+	λ_b suddenly leaps to 120% after 300 days, and 10 slot servers are added
LD	λ_b suddenly leaps to 120% after 300 days and λ_p suddenly drops down to 80% after 600 days

pattern LD. The objective value $G(\pi, d)$ under different settings of (v, h) is shown in Table 12, where $d = 1000$. The results are not very sensitive to the choice of (v, h). In the patterns having leap, a larger h seems to result in slightly better results. In contrast, a smaller h suits the patterns having Trend. $v = 90$ is a little bit long in this scenario. Since $(v = 10, h = 0.5)$ works well almost in all patterns, it is used in the numerical experiments later for a realistic scenario. Although we agree that there may exist a combination of v and h which can lead to a better performance, in this paper, we focus more on the improvement that can be gained by applying the AHWIF policy instead of the HWIF policy. The numerical experiments show that the results with $(v = 10, h = 0.5)$ are already satisfying.

The realistic scenario is constructed in the following way. We consider the arrival rates of the breast group and the prostate group are non-homogeneous, while others remain the same. For simplicity, we use λ_b (λ_p) to represent the arrival rates of the breast (prostate) group. How the arrival rates change are listed in Table 13, which are also shown visually in Fig. 5. Since the arrival rates are now changing with time, so is the workload of the system. To show how the changing arrival rates affect the system performance, we let the average workload within the considered period to be the same as it in the homogeneous situation. The only exception is pattern L^+, which suddenly increases the capacity in the system.

We use the output of the local search algorithm as the initial value of γ_i in the AHWIF policy. The results are shown in Table 14, for three different measurements. Since the arrival rates are non-homogeneous now, when we evaluate $T_o(\pi, d)$ and $SL_o(\pi, d)$, the workload percentage of each patient type is calculated with its average arrival rate.

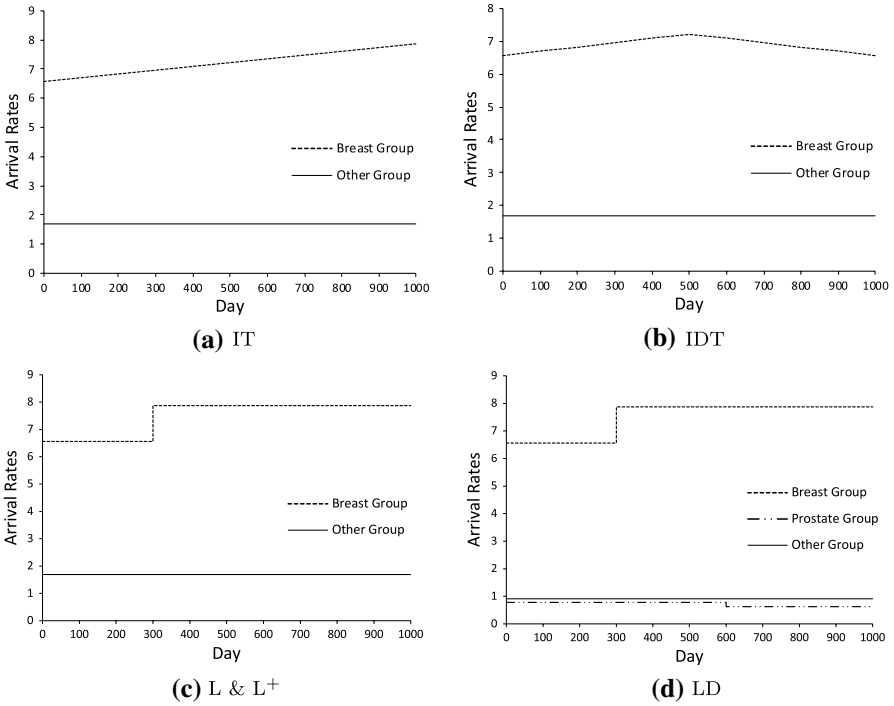

Fig. 5 Arrival patterns

Table 14 Performance of the AHWIF policy over $d = 1000$ days

	$G(\pi, d)$		$T_o(\pi, d)$		$SL_o(\pi, d)$	
	HWIF	AHWIF	HWIF	AHWIF	HWIF (%)	AHWIF (%)
IT	0.0995	0.0469	0.0491	0.0451	84.61	84.74
IDT	0.0165	0.0119	0.0118	0.0113	94.46	94.36
L	0.8878	0.2842	0.2682	0.2322	59.72	59.94
L^+	0.0005	0.0004	0.0005	0.0004	99.63	99.71
LD	0.2875	0.1164	0.1223	0.1122	67.84	68.00

According to the results, we first observe that the system performance in the non-homogenous situation is worse than the performance in the homogenous situation, although the average workload is the same. In other words, we should not only look at the average, ignoring the change of arrival rates. We also find that the system performance indeed can be improved by using the AHWIF policy. The objective value $G(\pi, d)$ has been reduced by 53%, 28%, 68%, 20%, and 60%, in each pattern respectively. The overall performance $T_o(\pi, d)$ is also reduced. We find that the improvement is more obvious in the pattern of IT, L, and LD. This is because the traffic loads in these cases are higher. When the capacity is sufficient,

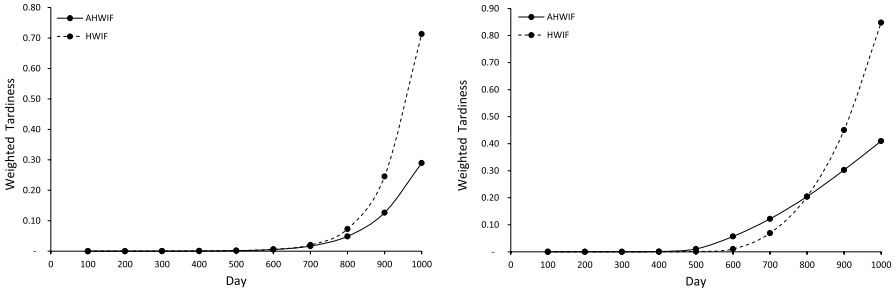

Fig. 6 maximal weighted tardiness in every 100 days

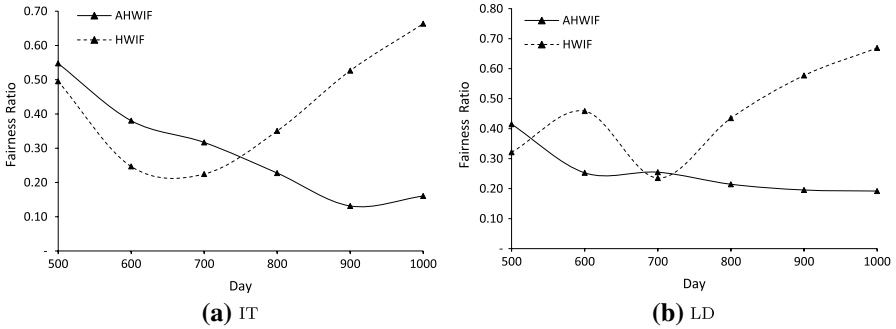

Fig. 7 Fairness ratio in every 100 days

the benefits gained by using a better scheduling policy are less. This can be drawn by comparing the results in the pattern of L and L^+. It is also inline with the conclusion in the previous workload discussion. In the pattern of LD, although the overall performance has not been improved too much since the changes are too mild, the big decrease in $G(\pi, d)$ implies the adaptive policy leads to a more balanced performance.

So far, we have looked at the expected tardiness over 1000 days. Now we look at the maximal weighted tardiness every 100 days during 1000 days. For brevity, we only show the results in the IT and LD patterns, see Fig. 6. In general, under the HWIF policy the maximal weighted tardiness increases more quickly. In the LD pattern, during day 500 to 800, the value of the maximal tardiness is larger under the AHWIF policy than it is under the HWIF policy. It may be caused by adjusting too much so that the performance of another patient type becomes the worst. A better combination of (h, v) may improve this. In both patterns, the value of the maximal tardiness starts to increase quickly around day 500. It is due to the traffic load exceeding 100% after day 500. Additionally, a fairness ratio is defined to represent the fairness between patient types. The fairness ratio is defined as the difference between the maximal and minimal weighted tardiness divided by

the highest one. Therefore a small fairness ratio value is preferred. As shown in Fig. 7, the value of the fairness ratio under the AHWIF policy is getting smaller. It implies, by balancing the performances between patient types, the AHWIF policy is efficient in absorbing the effects caused by the changes in arrival rates.

8 Conclusion

In this paper, we have discussed the scheduling problem for radiotherapy, considering multi-type patients with various service times and WTT requirements. The scheduling problem is solved by proposing a WI-based routing policy (i.e., the HWIF policy), which can help managers to decide which queue to treat next on the LINACs. The WI is the sum of the waiting time and the priority factor assigned to each queue. A simulation-based heuristic is proposed to determine the priority factors. In the objective function, we choose to minimize the maximal weighted tardiness so that the fairness between patient types can be considered. The HWIF policy turns out to be very efficient. We also propose an adaptive policy (i.e., the AHWIF policy) to deal with non-homogeneous arrival rates. The adaptive policy is constructed by two steps: the re-evaluation of the performance over the last v days and adding an adjustment h to the patient type with the worst performance. The results under non-homogeneous arrival rates have shown that the AHWIF policy works well in balancing the performances between patient types and leads to a better overall performance compared to the HWIF policy. In the future research, some variants of the adaptive policy can be studied, for example, dynamic v and h can be considered. Moreover, it would be interesting to test the policies using real data.

References

Burke EK, Leite-Rocha P, Petrovic S (2011) An integer linear programming model for the radiotherapy treatment scheduling problem. arXiv preprint arXiv:11033391

Chan W, Koole G, L'Ecuyer P (2014) Dynamic call center routing policies using call waiting and agent idle times. Manuf Serv Oper Manag 16(4):544–560

Chen Z, King W, Pearcey R, Kerba M, Mackillop WJ (2008) The relationship between waiting time for radiotherapy and clinical outcomes: a systematic review of the literature. Radiother Oncol 87(1):3–16

Conforti D, Guerriero F, Guido R (2008) Optimization models for radiotherapy patient scheduling. 4OR 6(3):263–278

Gurvich I, Whitt W (2010) Service-level differentiation in many-server service systems via queue-ratio routing. Oper Res 58(2):316–328

Legrain A, Fortin MA, Lahrichi N, Rousseau LM (2015) Online stochastic optimization of radiotherapy patient scheduling. Health Care Manag Sci 18(2):110–123

Legros B, Jouini O, Koole G (2015) Adaptive threshold policies for multi-channel call centers. IIE Trans 47(4):414–430

Li S, Geng N, Xie X (2015) Radiation queue: meeting patient waiting time targets. IEEE Robot Autom Mag 22(2):51–63

Organization WH (2017) World's health ministers renew commitment to cancer prevention and control. Cancer report WHO 2017 http://www.who.int/cancer/media/news/cancer-prevention-resolution/en/. Accessed 30 May 2017

Petrovic S, Leite-Rocha P (2008) Constructive and grasp approaches to radiotherapy treatment scheduling. In: World congress on engineering and computer science 2008, WCECS'08. IEEE, pp 192–200
Saure A, Patrick J, Tyldesley S, Puterman ML (2012) Dynamic multi-appointment patient scheduling for radiation therapy. Eur J Oper Res 223(2):573–584
Tezcan T, Dai J (2010) Dynamic control of n-systems with many servers: asymptotic optimality of a static priority policy in heavy traffic. Oper Res 58(1):94–110
Ward AR, Armony M (2013) Blind fair routing in large-scale service systems with heterogeneous customers and servers. Oper Res 61(1):228–243

Siqiao Li is a PhD Student in the Department of Industrial Engineering at Shanghai Jiaotong University, she has submitted the thesis on the topic of "Capacity Allocation and Patient Scheduling for Radiotherapy Process with Re-entrance and Random Arrivals", applying for the degree of Doctor. She is also currently a PhD Student in the Department of Mathematics at Vrije Universiteit Amsterdam. Her research interests include forecasting, modeling and simulation for improving workforce management in contact center systems.

Ger Koole is Full Professor in Applied Probability at Vrije Universiteit Amsterdam. His research is centred around the control of queueing systems, and applications of that in various areas, especially call centers, health care and revenue management. Next to his academic work, he also works for PICA, the VU University/VU medical center knowledge center on health care logistics. Moreover, he has also co-founded CCmath, a call center optimization company and founded Adscience, which focuses on internet advertisement optimization.

Xiaolan Xie is the head of the Department of Healthcare Engineering Center for Biomedical & Healthcare Engineering (CIS) at Mines Saint-Etienne, France. As a Full Professor, his research interests are centred around health care systems and services. Xie was named a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) in 2015 for his contributions to systems engineering for health care and manufacturing.