Ann Oper Res (2010) 178: 77–89
DOI 10.1007/s10479-009-0559-7

To pool or not to pool in hospitals: a theoretical and practical comparison for a radiotherapy outpatient department

Paul Joustra · Erik van der Sluis · Nico M. van Dijk

Published online: 20 May 2009

Abstract This paper examines whether urgent and regular patients waiting for a consultation at a radiotherapy outpatient department should be pooled or not. Both queuing theory and discrete event simulation were applied to a realistic case study. The theoretical approach shows that pooling is not always beneficial with regard to the waiting times of urgent patients. Furthermore, the practical approach indicates that the separation of queues may require less capacity to meet the waiting time performance target for urgent as well as regular patients. The results seem to be of general interest for hospitals.

Keywords Computer simulation · Hospitals · Queuing theory · Radiotherapy outpatient department · Waiting lists

1 Introduction

In service practices, the general perception appears to exist that it would be better to merge two (or multiple) queues into a single one, in order to use capacities more efficiently. Indeed, when only one type of service is involved this would be likely. In such an instance, in a single line system, none of the servers can ever be idle when tasks (e.g., patients) to be handled are still waiting. This observation is also supported by the standard $M/M/s$ queuing formula for mean delays; see for instance Tijms (1994) and Cooper (1981). In other words, for systems with one type of service, pooling capacities is clearly a superior strategy in terms of waiting time performance and/or the total capacity required.

However, if two or more different service types are involved, the question of whether capacities (or rather queues) should be pooled—assuming that the servers can handle the different service types—is less obvious and remains to be questioned for either of two reasons:

P. Joustra (✉)
Academic Medical Center, Meibergdreef 11, Amsterdam, The Netherlands
e-mail: p.e.joustra@amc.uva.nl

E. van der Sluis · N.M. van Dijk
University of Amsterdam, Roeterstraat 11, Amsterdam, The Netherlands

1. Different service characteristics (mix ratio);
2. Different service targets (workload ratio).

1.1 Mix ratio

For the first situation (reason), by pooling servers, variability is introduced due to the mix ratio of different means. As essentially based upon Pollaczek-Khintchine's formula, this can have a negative effect. The situation involving two single servers has already been addressed along with counterintuitive examples and analytic results in Whitt (1992) and Wolff (1989). A more extensive analytic and numerical treatment of this counter-intuitive phenomenon can be found in Whitt (1999). And more recently, in Van Dijk and Van der Sluis (2008), it was numerically shown and supported by approximate queuing formula that even for substantially larger numbers of servers it could still be advantageous (say in terms of mean waiting time) to keep capacities and queues separate.

1.2 Workload ratio

The present paper, in contrast, focuses purely on the second situation (reason) involving different service targets. There are two types of service requests with the same duration but with different waiting time performance targets.

1.3 Practical motivation (radiotherapy)

In practical terms, this second situation concerns hospital patients who require a consultation at a radiotherapy outpatient department. The consultations are stochastically identical for all patients. However, two types of patients are to be distinguished:

Type 1: (a small percentage of) urgent (or sub-acute) patients with a high performance target;
Type 2: (a large percentage of) regular patients with a substantially lower target.

The performance target is in terms of waiting time percentiles: namely a certain percentage within a given time. In hospitals, the different performance targets may follow from different recovery and quality criteria as well as financial agreements with insurance companies or rules set by the Ministry of Health. Due to these different performance targets, a separation of capacities might still be preferable, otherwise one group (typically the large group of regular patients) might be forced to pay a price to meet a higher target for the other group. A trade-off may thus have to be made.

Thus far, this second rationale for separate rather than pooled capacities seems to have remained uninvestigated within the queuing literature and has also not been covered in Van Dijk and Van der Sluis (2008). In health care literature, as will be specified in more detail later, it has been partially addressed in recent papers (see Thomas et al. 2001; Murray 2000; Murray and Berwick 2003). This paper, therefore, has a threefold objective.

1.4 Objectives

1. To investigate whether this second trade-off question of pooling is relevant from at least a queuing theoretical point of view for performance (waiting time) improvement;
2. If so, whether the performance (waiting time) improvement can also be obtained at the practical level such as in radiotherapy departments within hospitals, as based upon computer simulation for a case study;
3. If so, to what extent can the observation be applied to reduce capacities within radiotherapy departments.

1.5 Outline and results

First, in Sect. 1.5, a purely queuing theoretic approach is taken by means of a simple but instructive exponential parallel server system to obtain essential insights. By standard queuing formula, it is shown that trade-off points exist to keep the servers separated, depending on workload ratios. Even for this simple case, in the queuing literature no such result seems to have been reported as being of interest in itself.

Next, in Sect. 3, it is investigated to what extent such trade-off points can also be found in a realistic hospital environment. A case study is therefore included for the radiotherapy outpatient department of the Academic Medical Center (AMC) in Amsterdam, The Netherlands. As queuing formulas are no longer available in the more complex situation of the case study, discrete event simulation is used. Furthermore, for their practical interest in this case study a capacity viewpoint was adopted. It is shown that, also in the practical setting of the case study, it can be advantageous to keep capacities for different patient groups separate, as it may lead to an effective reduction of spare capacity. In Sect. 4, additional different scenarios were studied for their practical interest. A discussion completes this paper, which includes a brief evaluation, a review of the health care literature, and conclusions.

2 Queuing insights

Pooling two separate queues is generally perceived to be efficient. Indeed, when two separate queues for one type of service and two separate servers are pooled into a single queue for both servers, neither of the two servers can ever be idle while a customer is still waiting. Pooling thus seems to be the ultimate in efficiency.

More precisely, with $\boldsymbol{W}_P$ and $\boldsymbol{W}_A$ the mean waiting time (excluding service time) for the pooled and separate case, $\tau = 1/\mu$, and $\rho = \lambda/\mu$ the traffic load per server, by straightforward calculations from standard $M/M/1$ and $M/M/2$ expressions, pooling two parallel exponential servers would lead to a reduction factor of at least 50% (since $\rho < 1$) for the mean waiting time as by

$$\frac{\boldsymbol{W}_P}{\boldsymbol{W}_A} = \frac{\tau\rho^2/(1-\rho^2)}{\tau\rho/(1-\rho)} = \frac{\rho}{1+\rho}. \tag{1}$$

This reasoning, however, relies upon the implicit assumption that two identical servers, or rather identical service characteristics, have identical loads.

In Van Dijk and Van der Sluis (2008) it was shown by an approximate formula that similar reduction factors of at least 50% can also be found for larger groups of servers to be pooled provided:

- the service characteristics (mean durations) are the same;
- the workloads are equal.

2.1 Different performance targets

This paper considers another possible reason to keep queues separate: different performance targets in terms of waiting times. When pooling two patient groups, say for urgent and regular patients, all patients have to meet the high performance target for urgent patients. With separate queues, one can distinguish the urgent patient so that the regular patients do not have to meet the high performance target. This may save capacity, although the combined effect cannot be predicted.

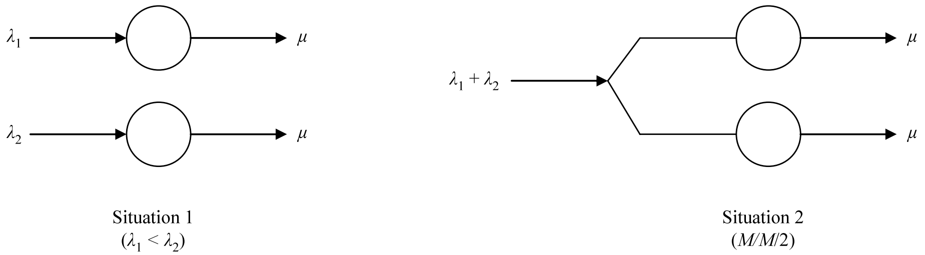

Fig. 1 Representation of situations

It is conceivable that a combination of performance targets for urgent and regular patients exists where separate queues would still be beneficial. First, this trade-off question was studied by a standard queuing formula for three reasons: 1. to illustrate the problem, 2. to theoretically prove that separate queues can be beneficial and 3. as it seems to have remained uninvestigated within the queuing literature, even with a standard queuing formula. This is not obvious, as the efficiency benefit of pooling capacity as seen in (1) may not exceed the efficiency loss, as a higher target is then also required for elective patients. To make a fair comparison, it is assumed that the service durations are identical (as is also the case in the practical AMC radiotherapy case study).

In our situation, a strict performance target for type 1 customers must be met, while type 2 customers are required to meet a substantially lower target. To this end, consider the situation involving two customer arrival streams with the same exponential service times with parameter μ, but different arrival rates λ_1 and λ_2, with $\lambda_1 < \lambda_2$, hence $\rho_1\ (= \lambda_1/\mu) < \rho_2\ (= \lambda_2/\mu)$. Two situations are compared. In situation 1 (separated case), each customer stream has its own single server. In situation 2 (pooled case), the two streams are merged into a single stream with a double server (see Fig. 1).

2.2 Waiting times

Let:

$W_i(\boldsymbol{W}_i)$: the (expected) waiting time of customer type i $(i = 1, 2)$ for the separated case;

$W_P(\boldsymbol{W}_P)$: the (expected) waiting time for the pooled case.

Bold letters are used for expectations (as already used in the previous section). Clearly, by the implicit assumption that $\lambda_1 < \lambda_2$:

$$\boldsymbol{W}_1 < \boldsymbol{W}_2.$$

With pooling, type 2 customers will experience shorter waiting times. The average waiting time for all customers can also be expected to decrease, as the workload is balanced over two servers and because of the pooling factor $\rho/(1+\rho)$ as in (1). The effect of pooling for the type 1 customers is less clear. On the one hand, having two servers available may lead to shorter waiting times; on the other hand, the overall workload becomes larger than for just the type 1 server.

Type 1 customers will not benefit from pooling when the expected waiting time $\boldsymbol{W}_P$ for the pooled system with average workload $\bar{\rho}$ exceeds the expected waiting time $\boldsymbol{W}_1$ for

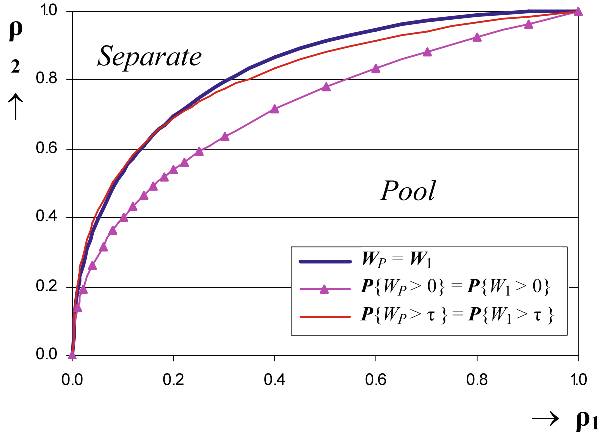

Fig. 2 Trade-off lines

the type 1 customers with workload ρ_1 in situation 1. In formula, by standard $M/M/1$ and $M/M/2$ expressions with $\bar{\rho} = (\rho_1 + \rho_2)/2$:

$$\begin{aligned} &\boldsymbol{W}_1 \leq \boldsymbol{W}_P, \\ &\frac{\rho_1}{1-\rho_1}\tau \leq \frac{\bar{\rho}^2}{(1-\bar{\rho}^2)}\tau. \end{aligned} \tag{2}$$

Here it is noted that (1) does not apply, as $\rho_1 \neq \rho_2 \neq \rho$. It is easy to see that inequality (2) holds for:

$$\bar{\rho} \geq \sqrt{\rho_1} \implies \rho_2 \geq 2\sqrt{\rho_1} - \rho_1. \tag{3}$$

From this inequality, trade-off values for ρ_1 and ρ_2 can be computed which lead to equality in (2) and (3), as illustrated in Fig. 2. Pooling two single servers is thus not always beneficial for all customers. For ρ_2 sufficiently large and ρ_1 sufficiently small, it is thus recommended not to pool.

2.3 Excess waiting time probabilities

In practice (particularly in health care), excess or tail probabilities are often used as a performance measure instead of average waiting times. With W_1 the waiting time of type 1 customers in the separated case and W_P the waiting time in the pooled case, for given value t the waiting time tail probabilities become:

$$\begin{aligned} &\boldsymbol{P}\{W_1 > t\} = \rho_1 e^{-\mu(1-\rho_1)t}, \\ &\boldsymbol{P}\{W_P > t\} = \frac{2\bar{\rho}^2}{(1+\bar{\rho})} e^{-2\mu(1-\bar{\rho})t}. \end{aligned} \tag{4}$$

For $t = 0$ the comparison leads to comparing the probability for having to wait $\boldsymbol{P}(W > 0)$. For type 1 customers, the waiting probability will increase by pooling if:

$$\frac{2\bar{\rho}^2}{(1+\bar{\rho})} \geq \rho_1. \tag{5}$$

Hence, after some manipulations, pooling is no longer useful for type 1 customers if:

$$\bar{\rho} \geq (\rho_1 + \sqrt{\rho_1^2 + 8\rho_1})/4 \quad \Longrightarrow \quad \rho_2 \geq \frac{1}{2}\sqrt{\rho_1^2 + 8\rho_1} - \frac{1}{2}\rho_1. \tag{6}$$

Similar relations exist for tail probabilities $\boldsymbol{P}\{W > t\}$ for other values t. For example, taking $t = \tau$ (with τ the mean service time) and using the performance measure $\boldsymbol{P}\{W > \tau\}$, pooling is not beneficial for customers of type 1, when:

$$\frac{2\bar{\rho}^2}{(1+\bar{\rho})} e^{-2(1-\bar{\rho})} \geq \rho_1 e^{-(1-\rho_1)}. \tag{7}$$

Unfortunately, this does not lead directly to an analytical expression for $\bar{\rho}$ or ρ_2. However, values for which equality in (7) holds are easily found by using a search or goal-seek procedure.

In Fig. 2, the trade-off lines are sketched where the inequalities (3), (5), and (7) hold with equality. A trade-off line indicates a combination of ρ_1 and ρ_2 where the pooled and separate situations perform equally for type 1 patients. For any combination of ρ_1 and ρ_2 in the area above these lines, separation of queues is beneficial for urgent patients.

The upper-left area, hence with a high ρ_2/ρ_1-workload ratio, for which it is preferable to keep capacities separate, typically seems to be applicable for practical situation as described by the practical motivation.

3 A practical case study for radiotherapy

The research for this paper was motivated by the radiotherapy department at the Academic Medical Center (AMC) in Amsterdam, The Netherlands. In this section, therefore, it will be investigated to what extent the theoretical findings of the previous section are applicable to a practical situation and to decide whether the management of the AMC radiotherapy department should keep the capacity pooled.

3.1 Case data

The radiotherapy treatment process consists of three consecutive steps: 1. a first consultation, 2. a preparation phase, and 3. an actual treatment. This study exclusively concerns the first step, the outpatient department. Currently, a small group of urgent patients and a large group of elective patient use the same first consultations. Hence, the capacity is pooled.

The real data of the demand and the available capacity for first consultations of new patients at the AMC radiotherapy department was obtained from the AMC planning system. The arrival pattern of referrals—based upon data from January to May 2006—fits a Poisson distribution with on average 32.5 patients per week. In the specified period, on average 10% of the referrals are urgent patients and the remaining 90% are elective/regular patients (see Table 1). The performance target indicates that 80% of the urgent patients need to have their

Table 1 Performance targets

Type	% Referrals	Performance target
1	10%	80% < 5 days
2	90%	80% < 9 days

first consultation within five calendar days after the date of referral. For regular patients the critical value is nine calendar days. These performance targets for the outpatient department are based upon the targets set by the Dutch Society of Radiotherapy and Oncology.

The capacity is not stable but fluctuates heavily between 25 and 42 appointments a week (such as due to national holidays, attendance at conferences, part-time work, and illness of physicians). The daily number of consultations fits a Poisson distribution (6.635 consultations on average). Both the number of referrals and the number of consultations fit a Poisson distribution. Although the fit is correct, the use of the Poisson distribution implies that both the number of referrals as well as the number of consultations is independent for subsequent days. For the referrals, this independency seems to be a justifiable assumption, but for the consultations it is not. The absolute effect is hard to predict, but as the Poisson distribution was used for the pooled situation as well as the separate queues situation, we assumed this does not influence the outcome of our trade-off question. Furthermore, the effect of the fluctuating capacity on the waiting times is too large to be neglected, so this aspect in the trade-off question had to be incorporated.

Regardless of the type of patient, the scheduled length of time for the first consultation is one hour. Currently both types use the same timeslots at the outpatient department, so in practice both queues are pooled.

3.1.1 Capacity

In practice, the management of the AMC radiotherapy department is only willing to split capacities if it will result in a capacity reduction. Therefore, next to waiting times the minimal required capacities will also be compared.

The available capacity was specified on a daily basis. For the pooled situation, the minimum expected number of daily consultations (the Poisson parameter) is determined up to a decimal fraction (e.g., 6.9 or 7.3) in order to meet the urgent performance target for *all* patients. For the separated case, the total number of consultations was sampled from a Poisson distribution in the same way as the pooled case. In the simulation, this randomly selected daily capacity was divided among the urgent and regular patients by first subtracting the urgent capacity. The remaining number of consultations was dedicated to regular patients. The decimal fraction of the urgent capacity was sampled from a Bernoulli distribution (e.g., with capacity 2.3, two consultations are always available and with a 30% chance a third consultation will be added). This is necessary because the randomly selected number of consultations for urgent patient has to be an integer. The minimum daily number of consultations for urgent patients, in contrast, is determined up to a decimal fraction.

3.2 Simulation

To determine whether the management of the AMC radiotherapy department should keep the capacity pooled, discrete event simulation was used to include the combination of:

- fluctuating capacities (number of consultations) and
- waiting time percentiles.

Due to the combination of both aspects (essential for our practical case study), queuing formulas are no longer available. Nevertheless, as well as for the qualitative behavior, the results from Sect. 1.5 were most useful to ascertain the existence of trade-off points and when to expect them.

Our system can be classified as a non-terminating simulation (see e.g., Law and Kelton 2002). Although the system restarts every day—which is typical for a terminating simulation—on the scheduling level the queue of patients waiting for the first consultation connects the individual days. Additionally, our aim was to investigate the long-term behavior of the system in terms of waiting time performance (in calendar days), which is typical for a *non*-terminating system.

For a non-terminating system, several methods of design of experiments are available. The replication/deletion approach (Law and Kelton 2002) was selected. To solve the problem of the initial transient, a warm-up period was included. Output statistics are only gathered after the warm-up period is over.

To determine the warm-up period (ten weeks of seven days with eight hours each), the method developed by Welch (1981) was used. The run length (including the warm-up period) was set to 50 weeks and the number of replications, based upon a desired half-width of 5% for the 95% confidence interval, was set to 200. To evaluate the performance of a scenario, a confidence interval had to be set up for the percentage of patients that meets the critical value. For the simulation, the performance target is supposed to be met when the lower bound of the confidence interval exceeds the target level of 80%.

After the design of experiments was completed, the simulation model was validated to check whether our model represents practice sufficiently accurate for the purpose at hand (Carson 1986). For the actual validation, the basic inspection approach (Law and Kelton 2002) was selected to compare the average waiting time of the simulation model with the real average waiting time of the AMC radiotherapy outpatient department. The average waiting time of the simulation model (7.3 calendar days) was almost equal to the actual average waiting time (7.2 calendar days) based upon data of the AMC planning system for the period January to May 2006. Therefore, our simulation model was considered to be valid and useful to evaluate the trade-off question in different scenarios.

3.3 Current scenario

As a first scenario, the current situation was executed. In the pooled situation, 35.5 consultations a week are necessary to meet the performance target for all patients. Note that all patients have to meet the high performance target for urgent patients (80% of the patients need to have their first consultation within five calendar days after the date of referral).

The situation of two separate queues also required a weekly capacity of 35.5 visits in order to meet the performance targets from Table 1. This may seem surprising but in Sect. 1.5 the existence of trade-off points (in terms of waiting times) were already proven with queuing theory. Coincidentally, the current practical situation is a trade-off point in terms of required capacity. However, in line with the theoretical results in Sect. 1.5, with equal capacities of 35.5 consultations a week, the separate and pooled situation yield

$$\boldsymbol{W}_1 = 3.2 < \boldsymbol{W}_P = 3.8. \tag{8}$$

In other words, for the realistic AMC case and from the point of view of inequality (2), the capacities should be separated. And indeed, this conclusion corresponds to inequality (3), as $\rho_1 = 65.0\%$ and $\rho_2 = 96.3\%$.

Furthermore, in the pooled situation on average 82% of the patients will have a first consultation within five calendar days. In the situation with separate queues, 91% of the urgent patients will have a first consultation within five calendar days and 84% of the regular patients will have a first consultation within nine calendar days. Hence, for the service levels (**SL**):

$$\begin{aligned}\mathbf{SL}_1 = \boldsymbol{P}\{W_1 < 5\text{ days}\} = 91\% > \boldsymbol{P}\{W_P < 5\text{ days}\} = 82\% = \mathbf{SL}_\mathrm{P}\\ \mathbf{SL}_2 = \boldsymbol{P}\{W_2 < 9\text{ days}\} = 84\% > \boldsymbol{P}\{W_P < 5\text{ days}\} = 82\% = \mathbf{SL}_\mathrm{P}.\end{aligned} \tag{9}$$

For the current scenario with unchanged capacities, separate queues will thus lead to higher percentages of patients who meet their critical value of the performance targets. Also in this respect, keeping the capacities separate can thus be regarded as superior, in accordance with Sect. 1.5.

4 Other scenarios

4.1 Performance target

The current critical values of the performance targets (five days for urgent patients and nine days for regular patient) lead to equal capacity requirements for both situations. This combination of critical values can be regarded as a trade-off point in terms of capacity. In line with Sect. 1.5, it should be possible to determine a trade-off line for the AMC radiotherapy department similar to the trade-off lines based upon queuing theory (see Fig. 3). To determine the trade-off point in terms of utilization rates ρ_1 and ρ_2, the utilization rates for urgent and regular patients were calculated in the separate queues situation. The utilization rate is the average weekly number of referrals of the corresponding patient type divided by the weekly minimum number of consultations required to meet the corresponding performance target (e.g., $\rho_1 = 3.255/5$ and $\rho_2 = 29.295/30.5$ for the current scenario).

Different combinations of critical values in terms of days of the performance target for urgent and regular patients were simulated to find other trade-off points in terms of equal capacity requirements for the pooled and separate queues situation (see Table 2).

(Note that the three-day performance target is the most stringent, as referrals on Friday are not scheduled before the following Monday.) The graph displayed in Fig. 3 shows the trend-line through the trade-off points displayed in Table 2.

Again, for any combination of ρ_1 and ρ_2 in the area above the trade-off line, the separation of queues is beneficial. This implies that, starting from a trade-off point, a smaller ρ_1 (and thus a higher performance target/lower critical value for the performance level of urgent patients) or a larger ρ_2 (and thus a lower performance target/higher critical value for the performance level of elective patients) is a plea for separate queues. Figures 2 and 3 clearly show that the theoretical and practical trade-off lines have a similar shape. Without Fig. 2 for the theoretical case, Fig. 3 would not have been found.

As a special application, to make the trade-off for the AMC radiotherapy department more explicit, different critical values (days) for urgent patients were evaluated, while the critical value for regular patients was kept to 9 days.

Fig. 3 Trade-off line for AMC radiotherapy department

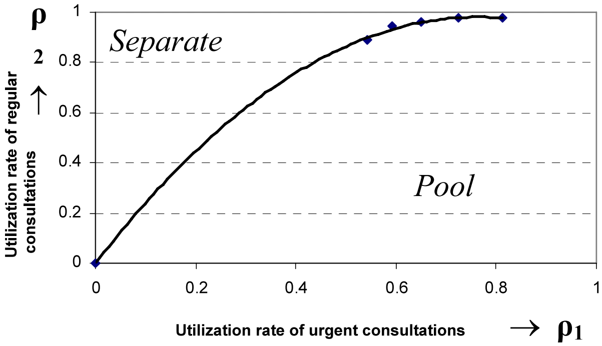

Table 2 Trade-off point AMC radiotherapy department

Critical value (days) for urgent patients	Critical value (days) for regular patients	Urgent capacity	Regular capacity	Total capacity
9	12	4	30	34
6	12	4.5	30	34.5
5	9	5	30.5	35.5
4	7	5.5	31	36.5
3	4	6	33	39

Table 3 Performance target scenario

Critical value for urgent patients	9 days	8 days	7 days	6 days	5 days	4 days	3 days
Pooled situation	34	34	34	34.5	35.5	36.5	39
Separate queues	35	35	35	35	35.5	35.5	36

Table 4 Patient mix scenario

Percentage of urgent patients	5%	10%	15%	20%
Regular capacity	32	30.5	29	27
Urgent capacity	3.5	5	6.5	8.5
Total capacity	*35.5*	*35.5*	*35.5*	*35.5*

The results in Table 3 are the total number of consultations required for the pooled and separated situation in order to meet the performance targets. These results indicate that pooling capacity is no longer beneficial for an urgent performance target of 80% within five days or less. For a three-day performance target for urgent patients, a separation of the patients will reduce the required spare capacity from (39 − 32.5 =) 6.5 to (36 − 32.5 =) 3.5, which is a reduction of nearly **50**%. Such a reduction of spare capacity is significant in health care organizations. Hospital departments are generally efficiency driven and, accordingly, strive for a minimization of scarce capacity, e.g. physicians.

4.2 Patient mix

One might expect that the smaller the fraction of the urgent group, the more separation of capacity might become advantageous. However, as it turns out for the case study, regardless of the percentage of urgent patients, in total 35.5 consultations are needed to meet the performance targets for both groups (see Table 4). Remarkably, changes in urgent capacity are matched exactly by changes in regular capacity. In this case study, the simulation model clearly shows that the trade-off question does not depend on the percentage of urgent patients. This observation is likely to be explained by the high workload for the majority of patients (the group of regular patients).

Table 5 Economy of scale scenario

Critical value urgent patients	5 days	4 days	3 days
Pooled situation	133	133.5	137
Separate queues	132.5	133	134.5

Table 6 Jockeying scenario

Strategy	Required capacity
Pooled queue	35.5
Separate queues	35.5
Separate queues with jockeying	33.5

4.3 Economy of scale

The radiotherapy department has a relatively small outpatient department compared with other departments. To analyze the effect of economy of scale, a larger OPD with four times more referrals was simulated.

With the current performance targets, a separation of the groups saves half a consultation per week (see Table 5). For the more stringent targets, the benefit of two separate queues is identical, as in the situation of a small outpatient department. The economy of scale does not seem to play an important role in the trade-off question. Again, as in Sect. 4.2, this observation seems related to the high workload involved in the present case study.

4.4 Soft blocks or jockeying

Soft blocks indicate that every group has a dedicated capacity but that in special cases a patient of one group can be scheduled in the other block. In our case, a regular patient may use an urgent timeslot when it is not occupied by an urgent patient one day before. In queuing theory, this strategy is also known as one-way-jockeying. Because the utilization rate of the urgent timeslots will be relatively low, this will probably lead to a decrease in the regular capacity required without affecting the urgent patients significantly.

Indeed, the results in Table 6 clearly show that jockeying reduces the required capacity to 33.5 consultations per week. With the current performance targets, this strategy potentially saves two consultations per week. This means that the required spare capacity drops by **67**% compared to the pooled situation.

Jockeying, however, implies that a regular patient has to be scheduled or rescheduled on the morning of the specific day, which may not always be possible in practice. Nevertheless, even with a success rate of only 50% for rescheduling, the necessary spare capacity still drops by over 30%.

4.5 Conclusions for the AMC radiotherapy department

The AMC radiotherapy case leads to the following conclusions:

1. In the current case, the performance is improved by separating the capacities for urgent and regular patients;
2. The more stringent the performance target for urgent patients, the more advantageous separation of queues becomes;

3. With jockeying from the regular to the urgent queue, the required capacity can reduce further. In addition, separation of queues will already be preferable in terms of required capacity in the current situation;
4. Different small fractions of urgent patients do not influence the trade-off question;
5. The economy of scale has only a minor effect on the trade-off question.

5 Discussion

In service industries, dividing capacity among several customers can be beneficial due to a relatively large difference in process times. In this paper, another potentially effective reason to divide capacity is investigated: different performance targets due to a different level of urgency, such as arise in hospitals.

5.1 Our study compared

Within health care, a subdivision of capacity at the outpatient department can be preferable for several underlying reasons:

1. Subspecialization: new patient referrals are divided into several groups with a different medical subspecialization and dedicated capacity for each group;
2. Fast-tracking: reserve dedicated capacity for urgent patients to reduce their waiting times;
3. Follow-up: new patients and follow-up patients use other timeslots at the outpatient department;
4. Geographical: consultations take place at several locations to reduce the travel times for patients.

In Thomas et al. (2001) both subspecialization and fast-tracking are claimed to lead to increased capacity requirements. We agree with the statement regarding subspecialization. However, this paper clearly proves that the subdivision of capacity for different levels of urgency does not necessarily lead to a higher demand for capacity.

In Murray (2000) it is argued that a subdivision of capacity has several disadvantages: 1. the necessary triage to determine whether the patient is indeed urgent has to be effective and accurate and 2. the demand will be less predictable, which implies that additional spare capacity is needed to reach the performance targets. This paper shows that the disadvantage for the small urgent group can be compensated by the large regular group, which could lead to a reduced overall capacity requirement.

In Murray and Berwick (2003) a plea is made for advanced access, which implies one queue with waiting times for urgent patients. Undoubtedly, this is the best strategy for all patients. However, working down the backlog is not an ongoing feature of advanced access, as claimed by Murray and Berwick (2003). Queuing theory shows this is not entirely true. To maintain low waiting times after working down the backlog, the utilization rate must be reduced to deal with the increased variation in demand. In situations with high access times, the demand for care is more stable because there is always a patient waiting to be seen. Unfortunately, the high capacity requirements associated with this strategy often cannot be met in practice where the available capacity/budget is limited. In this paper, it is demonstrated that subdivision for urgency reasons potentially saves capacity in situations where it is too limited to provide advanced access.

5.2 Conclusions

- Queuing theory turned out to be useful to provide basic insights and results to "look for". Using computer simulation, the extent of these results can then also be checked and evaluated in the more complex realistic situation of the case study. In addition, by computer simulation various what-if questions can be investigated, such as on different performance targets, patient mix, economies of scale, and jockeying.
- In the current situation, pooling or separating capacity at the AMC radiotherapy outpatient department requires the same number of consultations. With these equal capacities, however, a separation even slightly improves the performance (mean waiting times service levels).
- A combination of queuing theory and computer simulation led to practical insights and results, and seem highly fruitful.

Acknowledgements The authors wish to acknowledge the radiotherapy department of the AMC in Amsterdam for its cooperation and Geert Jan Kamphuis and Rachid Kolfin for their assistance in building the simulation model. Comments by the anonymous referees were of considerable benefit to the presentation and were highly appreciated.

References

Carson, J. S. (1986). Convincing users of model's validity is challenging aspect of modeler's job. *Industrial Engineering*, *18*, 74–85.

Cooper, R. B. (1981). *Introduction to queuing theory*. Amsterdam: North-Holland.

Law, A. M., & Kelton, W. D. (2002). *Simulation modeling and analysis* (3rd ed.). Singapore: McGraw-Hill.

Murray, M. (2000). Patient care: access. *British Medical Journal*, *320*, 1594–1596.

Murray, M., & Berwick, D. M. (2003). Advanced access: reducing waiting and delays in primary care. *Journal of the American Medical Association*, *289*(8), 1035–1040.

Thomas, S. J., Williams, M. V., Burnet, N. G., & Baker, C. R. (2001). How much surplus capacity is required to maintain low waiting times? *Clinical Oncology*, *13*, 23–28.

Tijms, H. C. (1994). *Stochastic models: an algorithmic approach*. Chichester: Wiley.

Van Dijk, N. M., & Van der Sluis, E. (2008). To pool or not to pool in call centers. *Production and Operations Management*, *17*, 1–10.

Welch, P. D. (1981). *On the problem of the initial transient in steady-state simulation*. Yorktown Heights: IBM Watson Research Center.

Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Management Science*, *38*, 708–723.

Whitt, W. (1999). Partitioning customers into service groups. *Management Science*, *45*, 579–1592.

Wolff, R. W. (1989). *Stochastic modelling and the theory of queues*. Englewood Cliffs: Prentice-Hall.