Proceedings of the 9th International Conference on
Applied Operational Research (ICAOR 2017)
Chung Yuan Christian University
Taoyuan, Taiwan
18-20 December 2017

Dynamic appointment scheduling in priority queueing systems with access time targets

Carrie Ka Yuk Lin [1]
[1] Department of Management Sciences, College of Business, City University of Hong Kong, Hong Kong (SAR), China
mslincky@cityu.edu.hk

Abstract

Multi-class priority queuing systems with access time targets and dynamic arrivals are common in many environments, including emergency, healthcare, logistics and hospitality services. This study focuses on the waiting line management of new cases in the specialist outpatient clinics of the Hospital Authority of Hong Kong. The objectives are to assign appointment dates in a way to fulfill the access time targets for urgent (highest priority) and semi-urgent (second highest priority) new cases while maintaining acceptable access time for non-urgent (routine) new cases. A dynamic appointment scheduling algorithm is developed and tested through simulated instances created from the waiting time data and hospital attendance data for a one-year horizon. The experimental results are compared with the reported performance statistics. Performances are also compared with results of the deterministic problem assuming the arrival data were given. For certain ordered set of objective weights relevant to the current problem, the deterministic problem can be formulated as a transportation model and solved optimally by the classical transportation algorithm with an additional checking procedure.

Keywords: Real-time scheduling heuristic; Healthcare; Appointment systems; Access time service level; Transportation model

Background

This study is motivated by the long waiting time for specialist outpatient clinics in public hospitals reported in recent years. In the current system, new cases are triaged into three categories: urgent, semi-urgent or non-urgent (routine) cases. After the Hospital Authority has adopted a number of measures (including diversion of stable, less urgent cases to family medicine specialist clinic, general out-patient clinic; setting up public-private partnership; and cross-cluster referral arrangement for selected patients), the median waiting time for urgent and semi-urgent new case bookings have been improved to within their respective target (of two and eight weeks, respectively) [1]. The tradeoff is the long waiting time for patients with less severe conditions and non-urgent cases seeking for follow-up appointments. Due to new and continuing cases, the total attendance figures of specialty outpatient clinics have been increasing in recent years. Occasional programs on clearing backlog were launched such as offering additional appointment sessions to the non-urgent new cases. To minimize additional effort, this study aims to provide a real-time scheduling tool for outpatient appointment booking systems to help manage the limited resources supply in anticipation of growing service demand. Moreover, it was suggested during government meetings that a performance pledge should be set on the median waiting time for routine cases of major specialties [2]. A dynamic scheduling tool could enable testing of such feasibility with evaluation of appropriate access time target.

Problem statement

The online appointment booking system operates like a single-server priority queueing system offering multiple units of capacity in consecutive time periods. In each time period, arrivals (patients) of different categories occur dynamically and each demands one unit of capacity in the current or a future time period. The decision is to assign an appointment time period (date) for each arrival to satisfy a specific target service level. It should ensure a specified proportion (or more) of patients in each urgent category can receive an appointment within a predetermined target time. Non-urgent category may not be imposed an access time target but the arrivals are expected to receive an appointment as early as possible. The objectives can be summarized in the following decreasing order of priority:

(i) Maintain the specific target service level by ensuring a specified proportion of patients (or more) in relevant categories can get access to service within the target time
(ii) Assign an appointment time period as early as possible to patients in categories with no access time target

(Naturally, objective (ii) can be handled as objective (i) with zero minimum specified proportion and a large access time target.) The system constraints are the service capacity in each time period based on availability of resources. A medium-term planning horizon (one year) comprising a number of time periods (workdays) is studied in which the arrivals are allowed to have appointments beyond this planning horizon.

The current problem is related to several classical problems. The demand and supply relationship is similar to a dynamic transportation model where the dynamic arrivals are each requesting for a unit capacity and suppliers correspond to the time periods (workdays) providing multiple units of service capacity. Objective (ii) is related to the parallel machine scheduling problem with the weighted completion time of jobs (patients) as the objective function. Each job has its release time (arrival time) and demands unit time on a machine but the number of machines varies in different time periods. The weights can represent the priority (importance) of categories. These two related problems inspire the development of (a) an optimal solution in the deterministic problem for performance evaluation; and (b) the dynamic scheduling algorithm.

Related studies

Other analytical methods for similar problems include stochastic dynamic programming which formulates a multi-day appointment problem to maximize the number of same-day appointments made with walk-in patients and the assignment frequencies of patients to their familiar physicians [3]. An adaptive threshold rule with a look-ahead time window is also proposed for practical use. Similarly, the problem of an appointment system where prospective inpatients will abandon depending on their waiting time is formulated as a Markov decision process [4]. In solving practical problems of realistic size, the large state space required has to be compromised by use of an approximate dynamic programming algorithm which could approximate the optimal solution in acceptable computational time. Other problems adopting this approach include scheduling diagnostic service, such as computer tomography scan in [5], in which the objective is to minimize the sum of appointment cost, patient diversion cost and delay cost over an infinite horizon.

As alternatives, an adaptive method is developed to adjust the resource (computer tomography scans) in a radiology department [6]. The operating hours are adjusted on a weekly basis and the time slots can be reallocated between patient groups on a daily basis. A hierarchical linear goal programming approach was adopted in an offline radiotherapy pre-treatment scheduling problem [7]. The waiting time target was set by patient type and priority class. The non-preemptive goals comprise of minimizing proportion of patients not meeting the waiting time target over all patient groups, and minimizing two lateness measures. Due to the large problem size, dispatching rules are employed to provide an initial solution to each phase after which the best objective obtained within a time limit will be passed onto the subsequent phase as an additional constraint.

Model Development

The current problem can be formulated by a stochastic goal programming model with uncertain demand arrival times. The solution of the deterministic equivalent model is used to assess the solution quality of the dynamic scheduling algorithm, given the arrival times of requests. This section first presents the model characteristics based on the local public specialist outpatient clinics.

Demand characteristics

- Patients arrive randomly and those with urgent conditions requiring early intervention are treated with priority. A triage system is implemented to classify patients into one of three categories.
- Each patient demands one unit of service (appointment) regardless of their category.

Server characteristics

- The daily service capacity (maximum number of appointments) for a specialist outpatient clinic is set by considering various factors, including the service demand, manpower available and capacity of physical facilities.

Operating targets

- Patients triaged into one of the urgent categories will expect an access (or waiting) time limit to an appointment. The management expect a significant proportion (e.g., at least half) of patients in these categories can get access to service within the target time. The actual statistics serve the purpose of monitoring, control and performance reporting.
- Stable (routine) new cases have no target access time but should be treated as early as possible. (The access time of a majority (e.g., 90th percentile) of these patients is also published regularly on the public website.) Without loss of generality, this category is defined as the last category in the problem.

Model parameters

- Deterministic

  M = number of patient categories in decreasing order of urgency

  D = number of workdays recording arrivals in the planning horizon

  T_m = the target access time to an appointment for patients in category $m = 1,\ldots, M$ (For non-urgent category M with no target access time, T_M can be set to a large positive value.)

  α_m = minimum proportion of patients in category m expected to have access time to service within the target T_m, $m = 1,\ldots, M$ (For non-urgent category M with no target access time, α_M can be set to 0.)

  P_m^-= priority weight (or penalty cost) of underachieving the (primary) goal on access time target of category $m = 1,\ldots, M$-1, where $P_1^- > P_2^- > \cdots > P_{M-1}^-$

  P_m^+= priority weight (or penalty cost) of not achieving the (secondary) goal on early appointment time of category $m = 1,\ldots, M$, where $P_1^+ \geq P_2^+ \geq \cdots \geq P_M^+$

  Q_t = units of service capacity on workday $t = 1,\ldots, D'$
- Stochastic

  N = number of patients arriving in interval [1, D] of the planning horizon

  N_m = number of patients of category $m = 1,\ldots, M$, where $N = \sum_{m=1}^{M} N_m$

  a_i = arrival day of patient $i = 1,\ldots, N$

  u_i = category of patient $i = 1,\ldots, N$

  L_m = list of category m patients sorted in ascending order of their arrival times, $m = 1,\ldots, M$, where $N_m = |L_m|$

  D' = total number of workdays required to provide appointments for the N patients (based on first-come-first serve regardless of patient category)

Decision variables

x_{it} = 1 if patient i receives an appointment on day t; 0 otherwise ($i = 1,\ldots, N$, $t = 1,\ldots, D'$)

d_m^- = underachievement of the minimum number of patients in category m achieving the target access time, $m = 1,\ldots, M$-1

Formulation of Retrospective Model (Model *R*)

$$\text{Minimize}\, Z = \sum_{m=1}^{M-1} P_m^- \cdot d_m^- + \sum_{m=1}^{M} P_m^+ \cdot \left(\sum_{i \in L_m} \sum_{t=a_i}^{D'} (t - a_i) \cdot x_{it} \right) \tag{1}$$

subject to:

$$\sum_{t \geq a_i}^{D'} x_{it} = 1,\; i = 1,\ldots, N \tag{2}$$

$$\sum_{i \in \{1,\ldots,N \mid t \geq a_i\}} x_{it} \leq Q_t,\; t = 1,\ldots, D' \tag{3}$$

$$\sum_{i \in L_m} \sum_{t=a_i}^{a_i+T_m} x_{it} + d_m^- \geq \lceil \alpha_m \cdot N_m \rceil,\; m = 1,\ldots, M\text{-}1 \tag{4}$$

$$x_{it} \in \{0, 1\},\; i = 1,\ldots, N,\; t = 1,\ldots, D',\; d_m^- \geq 0,\; m = 1,\ldots, M\text{-}1 \tag{5}$$

Model *R* consists of both deterministic and stochastic parameters. The first expression in the objective function (equation (1)) penalizes the underachievement of cases in urgent categories (1,…, *M*-1) with respect to the target time limit and desired proportion achieving the target. This serves as the primary goal of the model. The secondary goal is to schedule appointments as early as possible in each category. This is represented by the second expression minimizing the sum of access times of patients in each category (1,…, *M*). Demand fulfilment and service capacity is formulated by equation (2) and equation (3), respectively. The appointment day (*t*) of a patient (*i*) is scheduled after the request day (a_i). The minimum number of patients achieving the target access time in an urgent category (*m* = 1,…, *M*-1) is formulated as a goal constraint in constraint (4). Variables in the model are declared in constraint (5).

When the stochastic parameters are known in advance, model *R* becomes a deterministic linear goal programming model that can be solved exactly. It can be further transformed into a *variant* of the transportation model (*R*′) with constraint (4) incorporated into the objective function (1):

Formulation of Model *R*′

$$\text{Minimize}\, Z = \sum_{m=1}^{M-1} P_m^- \cdot \max\left\{ 0, \lceil \alpha_m \cdot N_m \rceil - \sum_{i \in L_m} \sum_{t=a_i}^{a_i+T_m} x_{it} \right\} + \sum_{m=1}^{M} P_m^+ \cdot \left(\sum_{i \in L_m} \sum_{t=a_i}^{D'} (t - a_i) \cdot x_{it} \right) \tag{1′}$$

subject to: (2), (3) and

$$0 \leq x_{it} \leq 1,\; i = 1,\ldots, N,\; t = 1,\ldots, D' \tag{5′}$$

In the current problem, the primary goal of ensuring a specified proportion of patients in relevant categories can achieve the access time target is treated with higher priority than the secondary goal of providing early access of service to all patients. This implies the following ordered set of weights is appropriate for the current problem:

$$P_1^- > P_2^- > \cdots > P_{M-1}^- > P_1^+ \geq P_2^+ \geq \cdots \geq P_M^+ \tag{6}$$

For the weight parameters satisfying (6), model *R’* can be solved optimally by the classical transportation simplex method for large problems (e.g., *N* = 18,000 patients, *M* = 3 categories, *D* = 247 workdays) with an additional checking procedure. A good initial solution is to apply the first-come-first-serve (FCFS) rule to patients in the most urgent category as early as possible when capacity is available. Repeat the same FCFS rule on patients in the next less urgent category until the last. The underachievement penalty cost (P_m^-) for a category (*m*) exists as objective coefficient for all its assignment decisions (x_{it}) when the target level ($\lceil \alpha_m \cdot N_m \rceil$) is not fulfilled in the incumbent solution. (Otherwise, the coefficient will be zero.) The additional checking applies to every incumbent solution when finding an improved solution. The non-basic cell with the most negative reduced cost will be checked to ensure the total cost of the new solution (after identifying a cycle of cells starting with the non-basic cell) really leads to a lower total cost as given by equation (1′). Otherwise, the non-basic cell with the next smallest (negative) reduced cost will be examined and the checking procedure repeats. When all reduced costs are non-negative at some stage, the optimal result provides a benchmark for evaluating the performance of the dynamic scheduling algorithm when the stochastic parameter values are revealed dynamically.

Dynamic scheduling algorithm

The design of the dynamic algorithm consists of two main features: (a) reserving a portion of the daily capacity for urgency categories and (b) allowing rescheduling of appointments to make better use of idle capacity by looking ahead into the near future. The flowchart in Figure 1 shows the logical design of the algorithm with the major parameters defined in Table 1.

Table 1. Algorithm parameters

Description	Notation	Design rationale	Value
• **Delay** in assigning appointment to category *m* patients (*m* = 1,…, *M*)	δ_m	No delay for the most urgent category 1; delay of category 2 patients = T_1 + 1 week; delay of category 3 patients = T_2 – 4 weeks	$\delta_1 = 0$, $\delta_2 = 3$, δ_3 = 4 weeks
• Proportion of daily capacity **reserved** for all urgent categories (*m* = 1,…, *M*-1)	*r*	Previous year demand, plus 5% growth	0.37
• **Look-ahead** period to reschedule appointments earlier	*h*	Operational feasibility and convenience	3 days
• Proportion of patients in non-urgent category willing to have their appointments **rescheduled** earlier	*p*	Assume between 5 to 6 rescheduled appointments per workday	0.8
• Number of instances with demand simulated	*n*	Achieving a maximum %error of less than 10% in estimating most performance measures with 95% confidence	30

Computational experiments and results

The computational experiments are designed to test the performances of the dynamic scheduling algorithm and model *R′* for a one-year period based on annual reported statistics on access time of new cases and hospital attendance figures [8-10] from the Hospital Authority. The recent four-year period from 2012-16 and the ophthalmology specialist outpatient clinic in the Kowloon East Cluster are selected due to past research work conducted in this clinic. The demand data for the previous year and current year will be simulated based on past pattern while the operating parameters will remain constant. Previous year patients with appointments carried over to the current year will be deducted from the daily capacity (number of appointments allowed).

Operating parameters

- M = 3 categories, D = 247 workdays = number of workdays in previous year.
- (T_m, α_m) = (2 weeks, 50%), (8 weeks, 50%) and (-, -) for m = 1, 2 and 3, respectively.
- Average Q_t = 59 appointments (from annual patient attendance of 14,472 in 2015-16). Assume the range of daily capacity = 6 appointments.

Demand-related parameters

- Previous year: Demand = 18,240 (2014-15), standard deviation (from annual demand in 2012-15)

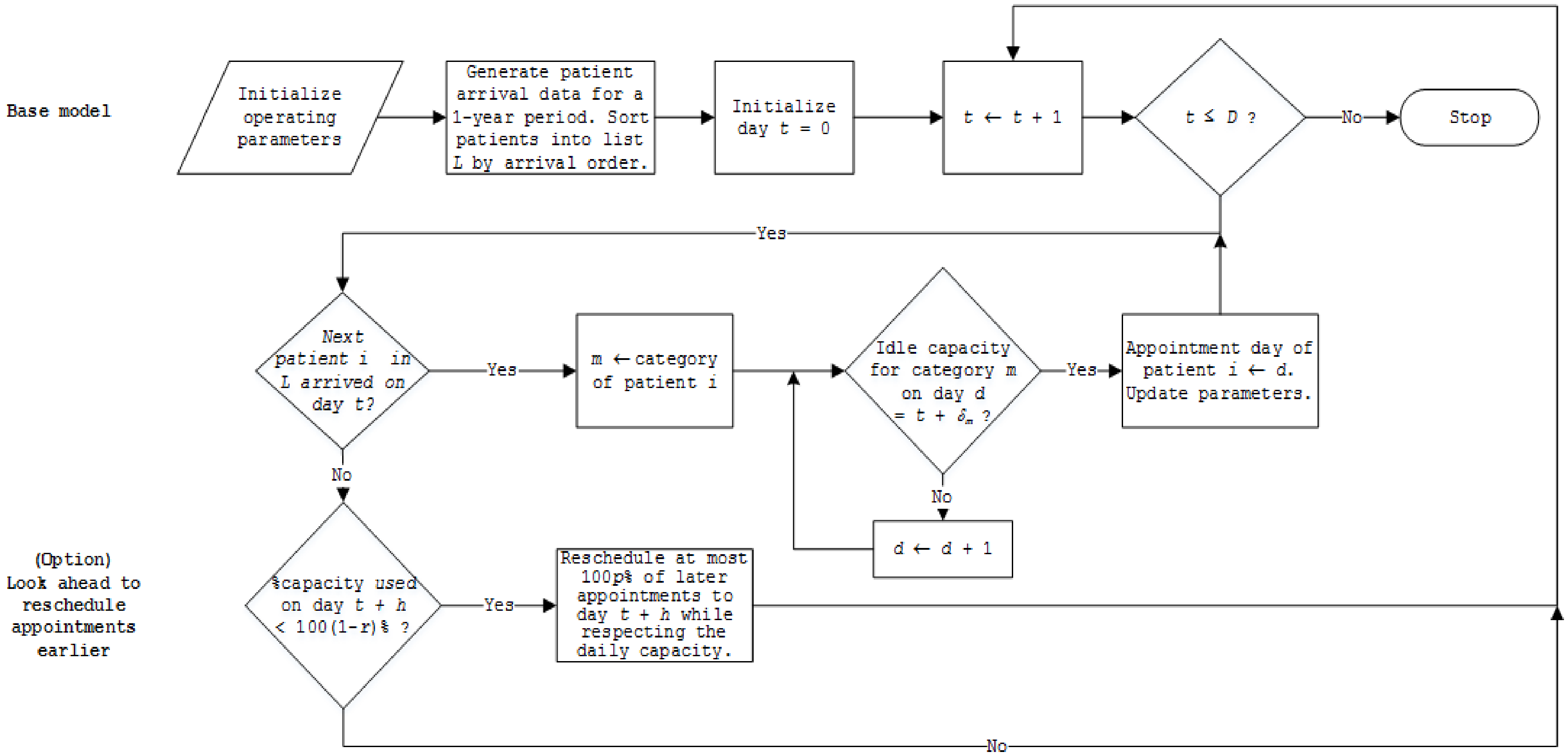
Base model
Initialize operating parameters
Generate patient arrival data for a 1-year period. Sort patients into list L by arrival order.
Initialize day $t = 0$
$t \leftarrow t + 1$
$t \leq D$?
No
Stop
Yes
Next patient i in L arrived on day t?
Yes
$m \leftarrow$ category of patient i
Idle capacity for category m on day $d = t + \delta_m$?
Yes
Appointment day of patient $i \leftarrow d$. Update parameters.
No
$d \leftarrow d + 1$
No
(Option) Look ahead to reschedule appointments earlier
%capacity used on day $t + h$ < 100(1-r)% ?
Yes
Reschedule at most 100p% of later appointments to day $t + h$ while respecting the daily capacity.
No

Figure 1. Design logic of the dynamic scheduling algorithm

= 310.24. Proportion of demand by category = 0.3, 0.03 and 0.67 for m = 1, 2 and 3, respectively.

- Current year: Demand = 18,292 (2015-16), standard deviation (from annual demand in 2013-16) = 364.66. Proportion of demand by category = 0.29, 0.02 and 0.69 for *m* = 1, 2 and 3, respectively.

Priority weights (or penalty cost)

- $P_1^- = 10^8$, $P_2^- = 10^6$, $P_1^+ = 10^3$, $P_2^+ = 10^2$, $P_3^+ = 1$

The demand-related parameters will be used to simulate $n = 30$ instances, each for a one-year period. Appointments made beyond this period ($t > D$) will also be recorded for performance evaluation. Results of the dynamic scheduling algorithm and model R' will also be compared with the reported access time statistics in 2015-16 [9]. Table 2 and Table 3 give the summary results of access time performances and the level of achievement of the targets, respectively.

With the adopted ordered set of priority weights, the dynamic scheduling algorithm can achieve similar or shorter access times in the urgent and semi-urgent categories ($m = 1$ and 2, respectively) at the expense of worse performance in the non-urgent category ($m = 3$). Nevertheless, the extreme performance of the largest 10% access time in this category is much reduced (66.6 vs 112 weeks). The average number of rescheduled appointments per day is 5.4 (with standard deviation of 0.62). The retrospective model R' produces the lower bound for the results of the dynamic scheduling algorithm. The worse performance of access times (25th and 50th percentile) in the non-urgent category is expected to be due to the ordered set of priority weights resembling a non-preemptive goal programming model allowing almost no tradeoff between different goals.

In terms of computational time, the dynamic scheduling algorithm takes an average of less than 6 CPU seconds per (one-year) instance on an Intel(R) Xeon(R) CPU E31270, 3.4GHz processor. To verify optimality in the deterministic problem, model R' could take over 4 CPU hours with the initial solution given by scheduling the categories in decreasing order of urgency and FCFS for patients in the same category. All algorithms have been coded in Microsoft Visual Basic .NET 2010 version.

Table 2. Performance of access times for all categories

Access time (weeks)	$m = 1$			$m = 2$			$m = 3$		
Percentile	25th	50th	90th	25th	50th	90th	25th	50th	90th
Dynamic sch. alg.	0	0.4	1.2	3.0	3.0	3.0	30.9	49.2	66.6
Retrospective model R'	0	0	0	0	0	0	28.5	37.6	42.4
Reported statistics [9]	< 1	< 1	1	3	6	7	11	15	112

Table 3. Fulfillment of access time targets for the urgent and semi-urgent categories

Percentage of patients with access time achieving target	**$m = 1$ ($T_1 = 2$ weeks)**	**$m = 2$ ($T_2 = 8$ weeks)**
Dynamic sch. alg.	96.7%	100%
Retrospective model R'	100%	100%
Reported statistics [9]	> 90%	> 90%

Discussion and conclusion

The reported statistics in Table 3 showed that not all urgent categories can meet their access time targets. Current performance assessment is based on the median access time of the two most urgent categories and both can achieve their respective target. However, the tradeoff is the long access time of the non-urgent category (routine new cases) with no imposed target time but large backlog accumulated over the years. This is reflected in the extreme case of 112 weeks, indicated by the 90th percentile of access time in the non-urgent category. The limitation of this study includes the simulation of only one previous (the most recent) year's demand to examine the impact on the current year's performance. The patient arrivals are uniformly generated in the 1-year period, as there is no information on the fluctuation of demand within the year. By contrast, the additional sessions offered by individual hospitals in reducing the access time of non-urgent new cases have improved the performance not reflected in the routine operations. Hence, absolute comparison with the reported statistics is not possible here.

The contribution of this work shows the potential to increase the percentage of urgent new cases achieving their respective target access time. The long access time of non-urgent new cases can also be reduced. Rescheduling appointments to make better use of anticipated idle capacity can significantly affect the access time performance. Hence, resources could be allocated to allow more rescheduling, if considered. Future research effort could focus on reducing the disparity in access times between the semi-urgent (m = 2) and non-urgent (m = 3) categories. The algorithm parameters (Table 1) could be made adaptive to respond to the dynamic environment. Multiple specialist outpatient clinics with data available [8-10] could also be used to test the algorithms developed.

Acknowledgments

The author thanks the postgraduate students from the City University of Hong Kong for their contribution during the project course. This work is supported by the Research Grant Council of Hong Kong (Project no. T32-102/14-N).

References

1. Hospital Authority (2017). Waiting Time for Stable New Case Booking for Specialist Out-patient Services. Retrieved 19 August 2017 from http://ha.org.hk/haho/ho/sopc/dw_wait_ls_eng_txt.pdf.
2. Legislative Council, Panel on Health Services, Cross-cluster referral arrangement for public specialist outpatient services of the Hospital Authority, 20 April 2015. LC Paper No. CB(2)1237/14-15(05). Retrieved 28 October 2016 from http://www.legco.gov.hk/yr14-15/english/panels/hs/papers/hs20150420cb2-1237-5-e.pdf .
3. Balasubramanian, H., S. Biehl, L. Dai, and A. Muriel. (2014). Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments. *Health Care Management Science*, 17, 31–48.
4. Lu, Y., X. Xie, and Z. Jiang. (2017). Dynamic appointment scheduling with wait-dependent abandonment. *European Journal of Operational Research*, accepted.
5. Patrick, J., M. L. Puterman, and M. Queyranne. (2008). Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research*, 56, 1507–1525.
6. Vermeulen, I. B., S. M. Bohte, S. G. Elkhuizen, H. Lameris, P. J. M. Bakker, and H. L. Poutré. (2009). Adaptive resource allocation for efficient patient scheduling. *Artificial Intelligence in Medicine*, 46, 67-80.
7. Castro, E., and S. Petrovic. (2012). Combined mathematical programming and heuristics for a radiotherapy pre-treatment scheduling problem. *Journal of Scheduling*, 15, 333-346.
8. Replies to initial written questions raised by Finance Committee Members in examining the Estimates of Expenditure 2014-15 (Reply Serial No.: FHB(H)064). Retrieved 1 October 2017 from http://www.legco.gov.hk/yr13-14/english/fc/fc/w_q/fhb-h-e.pdf .
9. Replies to initial written questions raised by Finance Committee Members in examining the Estimates of Expenditure 2016-17 (Reply Serial No.: FHB(H)089). Retrieved 1 October 2017 from http://www.legco.gov.hk/yr15-16/english/fc/fc/w_q/fhb-h-e.pdf .
10. Replies to initial written questions raised by Finance Committee Members in examining the Estimates of Expenditure 2017-18 (Reply Serial No.: FHB(H)055). Retrieved 1 October 2017 from http://www.legco.gov.hk/yr16-17/english/fc/fc/w_q/fhb-h-e.pdf .