International Journal of Business and Economic Development
Vol. 2, No. 3; September, 2021. pp 1-15
ISSN 2692-5591 (print), 2692-5583 (online)

Published by International Center for Promoting Knowledge

Integrated Capacity and Appointment Scheduling Strategy for Access Time Management

Carrie Ka Yuk Lin
Associate Professor
Department of Management Sciences
College of Business, City University of Hong Kong
Tat Chee Avenue, Kowloon Tong, Hong Kong

Abstract

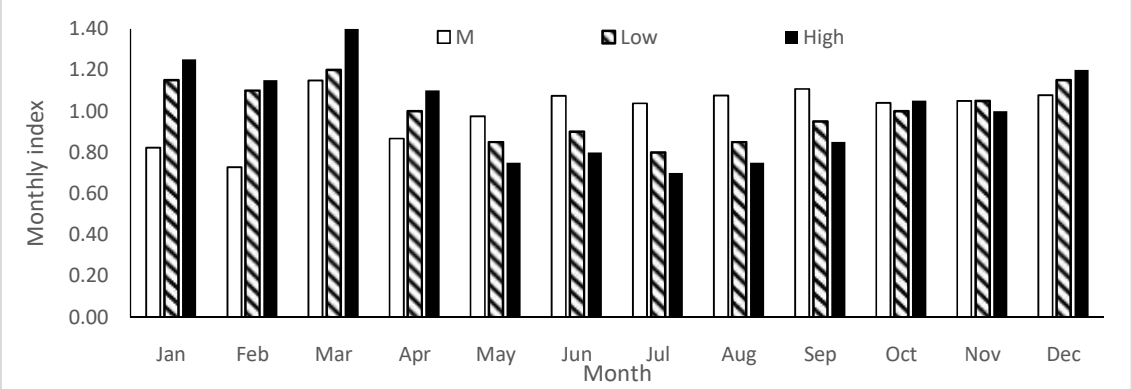
A common strategic decision problem faced by service providers is seasonal demand and limited resources. This paper proposes an integrated scheduling strategy of service capacity and patient appointments over a multi-day planning horizon. Demand is characterized by priority classes with seasonal variation. Short-term capacity is allowed to vary while the total capacity is kept constant, or within certain limits. System performances are measured from both demand and provider perspectives, including access time to an appointment, capacity utilization and completion time. Combining optimization approaches of integer programming and scheduling, an optimal capacity schedule and appointments with the smallest total and range of access times for each priority class is developed. Experiments are designed using published data on annual demand, capacity, costs in public out-patient clinics, and seasonality data in literature. Results indicate higher capacity utilization, shorter completion time converted into relative cost savings, and significant reduction in non-urgent patient access times.

Keywords: Capacity scheduling, Appointment scheduling, Seasonal demand, Integer programming

1. *Introduction*

Aggregate planning is known to be a strategy for macro-level planning that balances capacity and demand with the aim of minimizing the total costs in a period. In general, it includes the planning of production levels, manpower levels, inventory, demand backordering, based on the demand forecasts. It helps minimize shortcomings of hierarchical planning and short-term strategies which may result in large variation in capacity and resources required and subsequently incurs higher system cost (Nahmias and Olsen, 2015). The level strategy and chase strategy in aggregate planning illustrate two classical approaches to adjust the capacity to match with (changing) demand. The level strategy provides a constant level of capacity and is easy to manage. The chase strategy aims at adjusting capacity level to meet with changing demand by use of various options, such as overtime, part-time labour, demand backordering. The two strategies have been combined for use in practice according to the needs of the organization. In service environment, complex services like clinical services typically involve multiple resource types and patient classes. This paper is motivated by the appointment systems in public specialist out-patient clinics with large patient volume and multi-priority classes. Health care professionals often overwork and are understaffed in public healthcare systems (Ma, 2019; Lasater *et al*., 2020). Job control by management or staff could alleviate workload and reduce stress (Portoghese *et al*., 2014). Demand pattern for specialist services varies according to clinic type. In a public

medical clinic (Wong, 2012), there was no specific monthly demand variation except after Lunar New Year (around March). In other public and private clinics, larger variations have been observed over different months and weekdays of the year (Cayirli *et al*., 2019). Fig. 1 shows their monthly demand variations based on a standard monthly index of 1. Larger demand variations would require additional effort in staff scheduling or recruitment.



M: Public medical clinic (Wong, 2012); Low, High: Public and private clinics (Cayirli *et al*., 2019)

Fig.1. (Standardized) Monthly demand variations in out-patient clinics

Some countries have regulations or guidelines on staff working time to prevent occupational burnout. The European Union imposes a maximum of 48-hour workweek including overtime, averaged over a period of up to 4 months (European Union, 2020). In Hong Kong, a registered nurse in the Hospital Authority is required to work 44 hours per week on shifts and 6 days a week. Another flexible alternative for services experiencing seasonal demand is the annualized hours contracts which have been adopted in UK since 1980s. These allow an employee to work an agreed number of hours over a year, rather than on monthly or weekly basis. Apart from seasonal demand, it is found that organizations operating on a 24-hour basis and using a high level of overtime will be suitable for implementing annualized hours (Ryan and Wallace, 2019). Case studies of two Irish organizations, one perceived successful and the other failure, were analyzed. Structural factors, e.g., workplace partnership, play an important role facilitating successful implementation.

During the COVID-19 pandemic in 2020, many industries need to reduce their service hours. Some medical services have been postponed and appointments rescheduled. This paper investigates a flexible strategy of simultaneous capacity and appointment scheduling for medium-term planning. A one-year horizon is adopted in this paper to match with the available data from annual reports (Secretary for Food and Health, 2020). While past research on annualized hours had focused on workforce scheduling and costs from the management perspective in demonstrating its effectiveness (Hung, 1999; Van der Veen *et al*., 2015), this paper contributes to exploring the impact of simultaneous scheduling of capacity and appointments from the patient perspective, specifically, in terms of patient access times in multi-priority classes. The problem complexity includes large (annual) volume of demand with seasonal variation and capacity typically handled by public clinics. In this paper, both demand and capacity are measured by attendance figures (number of appointments) of new cases in which the access time statistics have been reported publicly (Secretary for Food and Health, 2018, 2019 and 2020).

2. Literature review

In the literature on matching supply and demand in health care, similar out-patient appointment systems with both first-visit and re-visit patients have been analyzed (Nguyen *etal*., 2018; Aslani *et al.,* 2020). In both studies, the objective is to determine the maximum required capacity over a medium-term planning horizon. Nguyen *et al*. (2018) modelled the uncertainty in arrivals of first-visit patients by chance constraint approach while Aslani *et al*. (2020) applied robust optimization. In both approaches, the stochastic capacity planning problem is converted to a deterministic equivalent form and solved as a linear optimization problem. The current paper is different with the objective of improving access times of multiple (three) categories of new

case (first-visit) patients by scheduling a given total capacity (allocated appointments) strategically over the planning horizon. Observing that the annual reported new case patients were exceeded by the capacity allocated in past years, no assumption of total supply matching demand is made, nor the existence of a static probability distribution of arrivals. The seasonality and uncertainty in arrivals are considered by demand scenarios created from past data. In reality, the access time to an appointment for a (low-priority) new case patient is prolonged (in terms of years) under inadequate capacity in some public clinics. However, follow-up (re-visit) patients will not be considered here due to lack of detailed information. Seasonality of demand created additional challenges especially in capacity-constrained clinics. Cayirli *et al.* (2019) analyzed an integrated problem of capacity allocation and patient scheduling for macro and micro level planning, respectively. The capacity allocation problem at the macro level determines the daily number of appointment slots to be reserved for walk-ins (and the rest for scheduled patients) during a one-year period, accounting for seasonal variation of demand. At the micro level, patients are scheduled to specific times on the appointment day.By use of an integrated simulation model, the macro level applied heuristic rules to allocate capacity (between walk-in/scheduled patients), providing input to the micro level appointment scheduling problem. One assumption is the daily capacity is fixed while allocation between walk-in and scheduled patients could vary in response to seasonal demand. The current paper relaxes the assumption of fixed daily capacity to total capacity over the planning horizon and investigates the impact on the system performances with seasonal (monthly and weekday) varying demand. Stochastic versions of the capacity allocation problem have been modelled in different ways. Deglise-Hawkinson *et al.* (2018) analyzed a network of specialist services and modelled demand seasonality by five independent demand distributions, one for each day of a 5-day workweek. Queueing network approximation (linearized by mixed integer programming) is applied to determine the maximum number of patients of a class to be admitted on a day to a department in the service network. If the assigned workload to a department exceeds capacity, it is served through overtime instead of overflowing to the following day. The current paper reduces such overflow by varying the daily capacity while balancing the total capacity (or within some allowable limit) in the multi-day planning horizon. Leeftink *et al*. (2019) analyzed an intraday appointment planning problem in an out-patient clinic treating multi-disciplinary cancer patients. Assuming fixed capacity of clinician type and same service duration as input parameters, the objective is to determine the right number of appointment slots to reserve for multi-disciplinary patients and other slots for scheduling regular patients that would minimize the expected (weighted) sum of patient waiting time, clinician overtime and idle time. The stochastic nature of patient routing is handled by the sample average approximation method.

The current paper shares several similar characteristics in that the objective is to mitigate undesirable customer (patient) experiences, expressed by a cost function comprising job waiting cost (patient access time weighted) by priority class; unit capacity required for each job (appointment) and large state space vector (depending on problem size) due to high volume of annual demand and capacity. Past research has pointed out that the choice of the access rules at the macro level (affecting patient access times) may be determined independently of the scheduling rule used at the micro level(Cayirli*et al*., 2019). Hence, the current paper will simply focus at the macro level.In contrast with previous studies, the deterministic version of problem is solved by integer programming and its relaxation, while the uncertain demand is modelled by demand scenarios created using seasonality information from past research. Furthermore, the daily capacities are included as decision variables (with the total capacity in the planning horizon kept constant) and the appointments are scheduled optimally by a simple and fair scheduling rule. When the demand is not fulfilled by the capacity in the same horizon, the completion time is simply extended to the period beyond the planning horizon.

From the provider perspectives, annualized hours' scheme has been introduced in some European and Scandinavian countries in the 1970s (Gall, 1996). Van der Veen *et al*. (2015) proposed a tactical planning model in allowing various types of employee contracts for a multi-skilled workforce to be used under annualized hours. The problem in finding a combination of cost-efficient contracts is formulated by a mixed

integer programming model, resulting in possible cost savings of 5.2% in a case study of an emergency department.

In operations scheduling research, Cappanera *et al.* (2019) studied a simultaneous scheduling problem of scanners and staff resources for patients requiring magnetic resonance imaging examinations. An online and an offline scheduling approach are designed and compared with three demand patterns. The offline approach is formulated as a mixed integer program and implemented using a rolling horizon of 52 weeks. It was found to offer more effective, equitable schedules when capacity is tight as compared with demand. The scheduling problem of jobs with release dates on identical parallel machines to minimize total weighted completion time (denoted as $P|r_j|\sum_j w_j C_j$) was formulated by several mixed integer programs (Kramer *et al*., 2020),one of which is a set-covering model with a branch-and-price algorithm that can solve instances of up to 200 jobs on 10 machines. Other specific problems closely related to the current paper include scheduling unit-time jobs on identical parallel machines with the objective of minimizing maximum job lateness or sum of weighted job tardiness. For objectives that can be expressed as non-decreasing functions of job completion times and weights, such as weighted completion time ($\sum_j w_j C_j$), necessary and sufficient conditions have been derived for efficiently checking a given solution for optimality (Brucker and Shakhlevich, 2016).

The current paper is related to the recent work of the author on dynamic scheduling of multi-priority requests with both service level and weighted total access time objectives (Lin, 2019) by considering capacity adjustment as decisions, in addition to appointment scheduling. The deterministic problem that can be solved here to optimality efficiently provides a basis for the dynamic problem in future.

3. Problem description

The current problem is applicable to service providers that can vary their short-term capacity (e.g., daily work hours) to a certain extent while maintaining the total capacity constant (or within an agreed limit) in a longer time period, such as to comply with contractual agreement or government regulations. This offers flexibility in varying capacity in response to demand changes and allows employees to have stable income. Another source of motivation came from public clinics providing specialist out-patient services to patients referred from general practitioners. The variation in the daily number of arrivals may depend on the month and day of the week. Typically, a patient (i) arriving at the clinic on a day (r_i) is triaged into one of the (M) priority classes ($m_i \in \{1,\ldots, M\}$). Each patient is scheduled an appointment on the arrival day or a future day. A patient-centric performance measure is the access time to an appointment, defined by the time elapsed between the arrival day of request and the appointment day scheduled. Both demand and service capacity in this problem are expressed in terms of number of appointments (requested and offered, respectively). The service capacity has a normal level (Q_t) on each day (t) and can vary within certain limits$[Q_t - \Delta_t^-, Q_t + \Delta_t^+]$, depending on factors from the supply, demand side or both. The sum of daily capacities over the period should be kept at some given total (Q_{sum}), or not more than a maximum limit ($Qsum + \rho$). This paper considers scheduling the capacity and appointments for a given set of (n) patients in a planning horizon of D days, with an estimated upper limit ($T \geq D$) of completion time of all n appointments. The objective is to minimize the total patient access time, weighted by the priority class. The provider-centric measures of capacity utilization during the D-day period and the completion time of all n appointments are also evaluated. Improvements over a base capacity strategy are converted into estimated cost savings. Naturally, the D-day period (e.g., a year) could be decomposed into multiple shorter adjustment periods (e.g., weekly, monthly), in which the sum of capacities within each adjustment period is to be balanced. For simplicity, this paper considers a single adjustment period with the full planning horizon of D (= 247) work days.

3.1. Assumptions

The model assumptions and the rationale are outlined as follows:

A.1 The deterministic simultaneous scheduling problem of demand with given total supply is considered. The solution can serve as benchmark for the dynamic or stochastic problem.

A.2 The dynamics on the scheduled appointment day, such as specific patient appointment times, on-site waiting time, patient and staff behavior, will not be considered. This paper aims at the macro level planning.

A.3 Other dynamic or stochastic factors, such as rescheduling or cancellation of a given appointment, due to request changes, is beyond the present scope. Realistic factors, like backlog of patient appointments from previous periods can be simply treated as given initial condition in the system, and hence will not be considered.

A.4 Each appointment requires unit capacity, regardless of priority class. In the government reported data (Secretary for Food and Health, 2018, 2019 and 2020), the annual service completions of new case patients are simply expressed in terms of attendance figures, with no differentiation among the 3 priority classes (urgent, semi-urgent and routine).

A.5 No rejection of appointment request. It is assumed that the general practitioner's referral letter and the triage at the clinic have screened the suitability of the patient for receiving treatment in the specialist clinic.

3.2. Problem formulation

The problem is formulated by an integer programming model, followed by the solution approach (section 4). Notice that there could be more than one way to formulate the same problem. The following choice of approach is due to the solution quality guaranteed and computational efficiency.

Decision variables

x_{it} = 1 if patient i has a scheduled appointment on day t; 0 otherwise (i = 1,…, n; t = 1,…, T)

q_t = revised capacity level on day t= 1,…, D

Integer Programming Model

Minimize $\boldsymbol{Z} = \sum_{i=1}^{n} \sum_{t=r_i}^{T} w_i \cdot (t - r_i) \cdot x_{it}$ (1)

subject to: $\sum_{t=r_i}^{T} x_{it} = 1,\ i = 1,\ldots, n$ (2)

$$\sum_{\{i|t \geq r_i\}} x_{it} \leq \begin{cases} q_t, & t = 1, \ldots, D \\ Q_t, & t = D+1, \ldots, T \end{cases} \quad (3)$$

$Q_t - \Delta_t^- \leq q_t \leq Q_t + \Delta_t^+,\quad t = 1,\ldots, D$ (4)

$\sum_{t=1}^{D} q_t = Q_{sum}$ (or $\sum_{t=1}^{D} q_t \leq Q_{sum} + \rho$) (5)

$x_{it} \in \{0, 1\}$, i = 1,…, n, t = 1,…, T, $q_t \geq 0$ and integer, t = 1,…, D (6)

The objective of the problem is to minimize the total patient access time to the service. The access time is the time elapsed between a patient's arrival day (r_i) and the scheduled appointment day (t).This objective is expressed as a weighted function in (1),where the weights (w) of importance are parameters reflecting the relative urgency in treating patients of different priority classes. These user-defined weights can be interpreted as penalties on access time delay with larger values associated with higher priority patients.The assignment of a patient to an appointment is formulated in constraint (2). The scheduled appointments on any day during the planning horizon (t = 1,…, D) is to be met by the revised daily capacity (q_t) in constraint (3). Beyond the planning horizon(t = D+1,…, T), scheduled appointments should match the given daily capacity (Q_t).Constraint (4) allows the capacity during the planning period to be varied within certain limits in response to the demand trend. The sum of daily capacity in the planning horizon is to be maintained constant (Q_{sum}) as formulated in constraint (5). In some situations, certain amount of increase/decrease in work (ρ) is allowed and specified in advance. In the experiments (section 5), the constant total in constraint (5) is adopted while the alternative can also be solved by a relaxation of the integer program (section 4). Lastly, the appointment scheduling and capacity adjustment decisions are declared in constraint (6).

4. Methodology

Three capacity allocation strategies are designed and compared in each test instance. They are named as Regular, Prorated and Proposed (optimal).

4.1. Proposed (optimal) strategy

A two-stage approach is proposed in solving the problem formulated (section 3.2). The first stage solves a relaxation version of the integrated model to provide an initial set of optimal decisions. The second stage adopts the optimal capacity decisions as input and improves the appointment decisions to a more equitable schedule for patients in each priority class.

4.1.1. Stage I Capacity allocation

The first stage determines the optimal capacity levels $\{q_t, t = 1,\ldots, D\}$ by relaxing the integer requirements of certain decision variables in the model (section 3.2). These include all the assignment variables, except the lowest priority class, i.e., $\{0 \leq x_{it} \leq 1 \mid m_i \neq M, t = r_i, \ldots, T\}$. The resulting relaxation model is a mixed integer program (MIP) which can be solved easily. In the experiments (section 5), the capacity levels and its adjustment parameters, $\{Q_t, \ \Delta_t^+ \ , \Delta_t^-, t = 1, \ldots, D\}$, adopt integer values as they represent number of appointments. The revised capacity decisions are also relaxed, i.e., $\{q_t \geq 0, t = 1,\ldots, D\}$. Results show that the optimal values of all decision variables are all integers satisfying the requirement of the original model. However, the range of patient access times in the lowest priority class could be large due to existence of many multiple optimal solutions. To obtain a more equitable and balanced set of access times in each priority class while maintaining the same minimum total access time, the second stage is applied which is computationally efficient (of polynomial time complexity).

4.1.2. Stage II Appointment scheduling

With the optimal revised daily capacity from the first stage as input, the earliest start schedule in Brucker and Shakhlevich (2016) for optimally scheduling unit time jobs with different release times is applied. One of their problems in minimizing the total weighted job completion time $(P \mid r_j, \ p_j{=}1 \mid \sum w_j C_j)$ is equivalent to the current stage. A simple optimal scheduling rule is to group all n patients by their priority classes. Then sort patients in each priority class in non-decreasing order of arrival days. At each decision point (day), schedule an available patient from the highest to the lowest priority class until all patients are scheduled. This simple procedure is efficient (of time complexity O ($nlogn$)) and offers fair treatment of patients in the same priority class as they are first-come-first-served.

4.2. Alternative capacity strategies

The Regular and Prorated strategy are designed to mimic the level and chase strategy in aggregate planning, respectively. The Regular strategy represents the given daily normal capacity level $\{Q_t, t = 1,\ldots, T\}$ with no capacity adjustment option. (Section 5 describes the details of simulating the daily capacities.) The Prorated strategy is designed to allocate the daily capacity in a way that the rate of *cumulative* capacity matches the *cumulative* demand rate. As the total capacity (Q_{sum}) is kept constant (or within an agreed allowance change ofρ) in the planning horizon, the Prorated strategy is based on a *cumulative* chase concept while preserving the total capacity and staying within the daily limits. Table 1 shows the algorithm of determining the daily capacities (denoted as $q_t^{prorata}$, $t = 1,\ldots, D$). On any day when the cumulative capacity exceeds the given total (Q_{sum}), the excess will be distributed to previous days. (After the D-day planning horizon, the daily capacity in the Prorated strategy will be set at the given level.)Once the daily capacities are decided in the Regular or Prorated strategy, the appointments are scheduled optimally by the same rule in the Proposed strategy (section 4.1.2: Stage II).

Table 1

Algorithm of the Prorated strategy in determining daily capacities $\{q_t^{prorata}, t = 1, \dots, D\}$

Let n_t = number of patients arriving on day t (= 1,…, D), where $\sum_{t=1}^{D} n_t = n$

Algorithm

initialize: **set** *cumulative demand* = 0, *cumulative capacity* = 0.

for t = 1 to D:

cumulative demand=*cumulative demand* + n_t

q =*min* $\{max\{Q_t - \Delta_t^-, \frac{cumulative\ demand}{n} \cdot Q_{sum} - cumulative\ capacity\}, Q_t + \Delta_t^+\}$

$q_t^{prorata} = q$

if q + *cumulative capacity*>Q_{sum} **then**

    excess = q + *cumulative capacity*$-Q_{sum}$

    $t' = t$

    while *excess* >0 d**o**

        $t' = t' - 1$

        δ=$min\ \{excess, q_{t'}^{prorata} - (Q_{t'} - \Delta_{t'}^-)\}$

        $q_{t'}^{prorata}$=$q_{t'}^{prorata} - \delta$, $excess$=$excess - \delta$

        cumulative capacity =*cumulative capacity*$-\delta$

    end while

else if $t = D$ and q + cumulative capacity <Q_{sum} **then**

    slack = Q_{sum}–(q+cumulative capacity)

    $t' = t$

    while *slack* >0 **do**

        if $q_{t'}^{prorata} + 1 < Q_{t'} + \Delta_{t'}^+$ **then**

            $q_{t'}^{prorata}$=$q_{t'}^{prorata} + 1$, *slack* =*slack* – 1

            If $t' \neq D$ Then *cumulative capacity*=*cumulative capacity* + 1

        end if

        t' = (t' *mod* D) + 1

    end while

end if

cumulative capacity=$q_t^{prorata}$+ *cumulative capacity*

end for

Return $\{q_t^{prorata}, t = 1, \dots, D\}$

5. Computational experiments

5.1. Data and information sources

The experiments have been designed using the annual data reported from specialist out-patient clinics in public hospitals (Secretary for Food and Health, 2018, 2019 and 2020) and seasonality data in past research (Cayirli *et al*., 2019; Wong, 2012). The daily demand and normal capacity levels are simulated (section 5.2).

The experiments will focus on new cases of specialist out-patients as data on annual capacity, demand, access time percentiles and relevant costs are available by specialty. The data reported in the 2017-18 period has been selected for the experiments as they reflected more recent performance than the 2015-16 data used in the author's recent work (Lin, 2019). These include the reported service capacity (attendance figure of new cases from all priority classes) and demand of new cases by priority class (Secretary for Food and Health, 2019). Three specialties, Medicine, Orthopaedics & Traumatology and Psychiatry, and their busiest hospital cluster are selected due to their relatively large access times compared to the same specialty in other hospital clusters. Psychiatry specialty has a ratio of demand to capacity below one, in contrast with the other two specialties. However, its demand growth is anticipated to increase in future. With regard to the reported performances, the higher priority classes have their 90th percentile access times closed to the corresponding target time but in the last priority class (routine cases), half of the cases had access time exceeding a year (52 weeks). Table 2 presents their relevant data used as parameters in the current experiments.

Table 2
Selected out-patient clinics and reported data in 2017-18 (Secretary for Food and Health, 2019)

Specialty (hospital cluster)	Demand (Number of new cases by priority class)	Capacity (Attendance of new cases)	Ratio of demand to capacity	Access times percentiles in weeks (25th, 50th, 90th)			Average cost per attendance of new case (HK$)	Median salary per month of doctors in hospital cluster (HK$)
				Priority 1 (Urgent)	Priority 2 (Semi-urgent)	Priority 3 (Routine)		
MED[1] (KEC)	22,214 (=1,865+5,016+15,333)	18,468	1.203	(<1, 1, 2)	(4, 6, 8)	(21, 87, 104)	2,120	111,730
ORT[2] (KEC)	16,688 (=3,642+3,941+9,105)	12,465	1.339	(<1, 1, 1)	(5, 7, 8)	(21, 108, 117)	1,030	111,730
PSY[3] (HKWC)	3,701 (=365+884+2,452)	3,838	0.964	(<1, 1, 2)	(2, 3, 7)	(23, 63, 118)	1,440	111,730

[1]MED: Medicine (Kowloon East cluster) [2]ORT: Orthopaedics & Traumatology (Kowloon East cluster) [3]PSY: Psychiatry (Hong Kong West cluster)

For long-term planning horizon (1 year), demand uncertainty could be high. For future demand estimation (as input parameter to the Proposed and Prorated strategies), a simple linear trend model has been tested on the annual total new cases in the previous 5 years' data (2012-13, …, 2016-17).The coefficient of determination (R^2) is found to be 79%, 97% and 65% for the three selected clinics. Subsequently, the projection to year 2017-18 is found to have an absolute percentage error of 6%, 10% and 37%, respectively. More accurate forecasts are often associated with clinics of larger demand (e.g., Medicine, Orthopaedics & Traumatology). Similar trend forecasting can also be applied to priority classes with large demand.

5.2. Design of demand scenarios and test instances

To compare the three capacity strategies in each selected clinic, 6 demand scenarios are created, resulting in a total of 18 test instances. The 6 demand scenarios per clinic represent combinations from 3 sets of monthly variations (Fig. 1) and 2 sets of weekday variations (Fig. 2). The monthly variations labelled as "High" and "Low" are adopted from Cayirli *et al*. (2019). Both experienced low demand periods in summer (May to September) and high demand from December onwards with peak in March. The data from a local medical clinic (Wong, 2012) also experienced peak demand in March, but with less monthly variation. The lowest demand was observed in February, possibly due to the Lunar New Year effect. These three sets of monthly variations will be labelled as "H", "L" and "M" in the demand scenarios. The two sets of high and low weekday variation from Cayirli *et al*. (2019) plotted in Fig. 2 indicated that Monday is the busiest and Wednesday has the lightest demand. (The average weekday index is standardized to 1.) They are labelled as "H" and "L" for the high and low weekday variation, respectively.

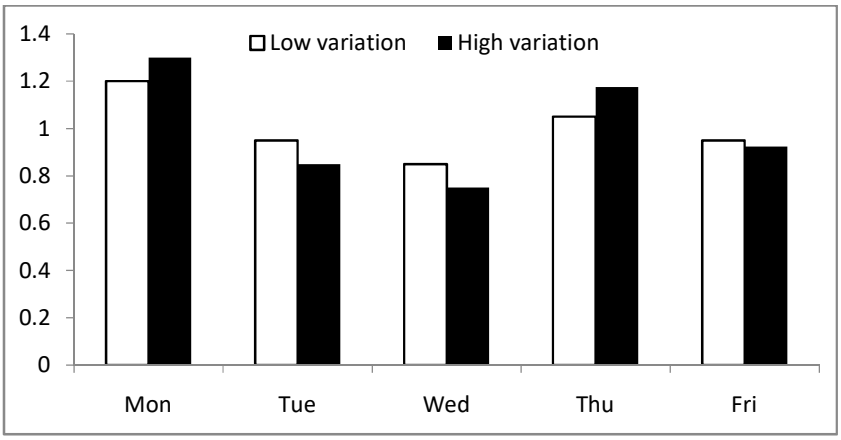


Fig. 2. Weekday demand variation in out-patient clinics (Cayirli *et al*., 2019)

5.2.1. Simulated parameters

From the data on annual demand and service capacity (Table 2), the daily demand and capacities are simulated for the planning horizon of D days. In year 2017-18, D = 247 workdays in Hong Kong. July is set as the first month in the experiments to be consistent with the reporting period (Secretary for Food and Health, 2019).

- Demand

In the demand scenarios, a multiplicative relationship is assumed between the monthly and weekday effects. This means weekday j in month i would have an average demand of $\frac{N}{D} \cdot s_i \cdot d_j$, where N is the reported annual number of new cases, s_i and d_j is the index of month i and weekday j, respectively. The daily demand is randomly generated while ensuring three conditions to hold: (i) the total over D days is N (ii) the relative ratios of the 12 monthly total demand match the monthly indices $\{s_i, i = 1,\ldots, 12\}$ (iii) the relative ratios of the 5 weekdays' demand in each month match the weekday indices $\{d_j, j = 1,\ldots, 5\}$. To quantify the priority between different classes, the weights of importance representing penalty on access time delay (section 3.2) are assigned subjectively with values of 1000, 100 and 1 per day for priority class 1, 2 and 3, respectively. Note that the Regular and Prorated capacity strategies do not require such weights in determining the capacity decisions.

- Capacity

The normal level daily capacities $\{Q_t, t = 1,\ldots, T\}$ are first generated. These will serve as inputs to the three capacity strategies. The average daily capacity, $\bar{Q}$, is first calculated by dividing the reported (annual) capacity, Q_{sum}, by D days (in the 1-year planning horizon). Assuming the range of capacity adjustment allowed is 10%, the daily capacity (Q_t) is randomly generated from the interval $[||\bar{Q} - 0.05\bar{Q}||, ||\bar{Q} + 0.05\bar{Q}||]$, where $||x||$ is the nearest integer to the real number x. To avoid large day-to-day variation, the sum of daily capacities in every half-month period is maintained constant at the average level of $||\bar{Q} \cdot D/24||$.

5.3. Results

The three strategies have been coded in Microsoft Visual Basic .NET 2015 with IBM ILOG CPLEX 12.6.3 incorporated as the solver for the Proposed strategy (section 4.1.1: Stage I). All experiments were performed on an Intel® Core™ i7-6700 CPU @ 3.40 GHz processor with an installed RAM of 8 GB (7.89 GB usable). Despite the large total patient volume handled in a year (Table 2), the computational time shown in Table 3 for all three capacity strategies is within 5 minutes which can facilitate rerunning and re-scheduling, whenever necessary. In terms of time complexity, the three strategies sorted in non-decreasing order are Regular, Prorated and Proposed.

Table 3

Computational time statistics (CPU seconds) of capacity strategies in the three clinics

Capacity strategy	MED		ORT		PSY	
	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
Regular	166	24	151	56	40	6
Prorated	163	6	125	11	53	19
Proposed	261	41	153	28	12	3

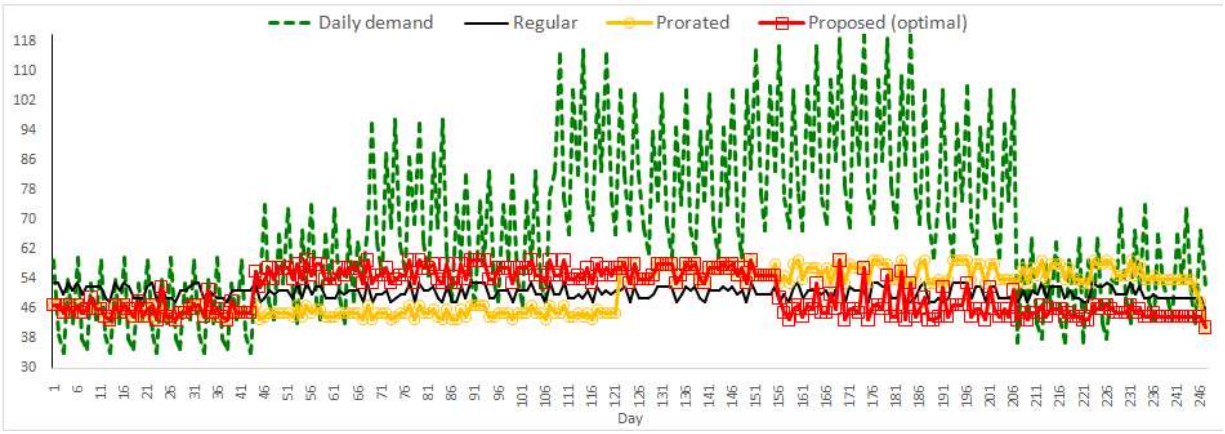


Fig. 3. Capacity strategies (Regular, Prorated and Proposed) in scenario (ORT clinic, High monthly variation, High weekday variation)

Fig. 3 highlights the differences among the three capacity strategies for the Orthopaedics & Traumatology clinic in the scenario of high variation in both monthly and weekday demand. (This clinic has the largest ratio of annual demand to capacity among the 3 clinics.) The largest gap between demand and supply occurred during December to April (day 108 to 206 in Fig. 3). This period happens to match with that of high monthly indices (Fig. 1) from which the daily demand is generated. The Regular strategy with no seasonal adjustment is considered as the base strategy for comparison with others. An interesting observation is made in the Proposed (optimal) strategy that it does not follow the same demand trend. While high demand occurs during December to March (day 108 to 187), the Proposed (optimal) strategy allocates capacity to its upper limits earlier, from September to end of January (day 45 to 148), exceeding the Regular capacity level. Thereafter, its capacity level falls to the lowest limits (and below the Regular capacity level) even in the peak demand month of March (day 167 to 187). This characteristic could be explained as demand trend is increasing, allocating capacity as early and as much as possible could reduce the cumulative backlog. (This is important for clinics with demand exceeding planned capacity.) As a comparison, the Prorated strategy, based on the updated cumulative demand and capacity information without looking ahead into future demand trend, allocates capacity slowly when demand increases gradually from September to December (day 45 to 126). This time lag of information causes it to allocate more than the Regular strategy from March (day 167) onwards as demand starts to drop. Similar observations were made in the (MED, H, H) scenario for the Medicine clinic which has the second largest ratio of demand to capacity. In the Psychiatry clinic where demand is less than the capacity available, the Proposed (optimal) strategy traces the demand trend more closely, within the daily allowable limits, while the Prorated strategy allocates capacity in a similar pattern as in Fig. 3. In brief, the main difference between the Prorated strategy and the Proposed (optimal) strategy is the former operates reactively to cumulative demand recorded while the latter operates proactively in using the demand information in the full period for capacity allocation.

5.3.1. Patient-centric performances

Figures 4 to 9 display the patient access time statistics for demand scenarios of the 3 clinics, typically those with high/medium monthly variation and high weekday variation. The highest priority (urgent) patients have no access time delay in all three capacity strategies. For the second highest priority (semi-urgent) patients, Orthopaedics & Traumatology (with the largest ratio of demand to capacity) is the only clinic recorded some access delay. In most scenarios, the proposed (optimal) strategy gives the best or the same access time performances. Although the objective targets at minimizing the weighted sum of access times, the proposed optimal strategy also excels in the percentiles of minimum, median and maximum access time. The greatest benefit is associated with the lowest priority (routine) class which has the largest patient volume. Hence, the reduction in the total access times of this class is significant, as shown in Figures 5, 7 and 9.

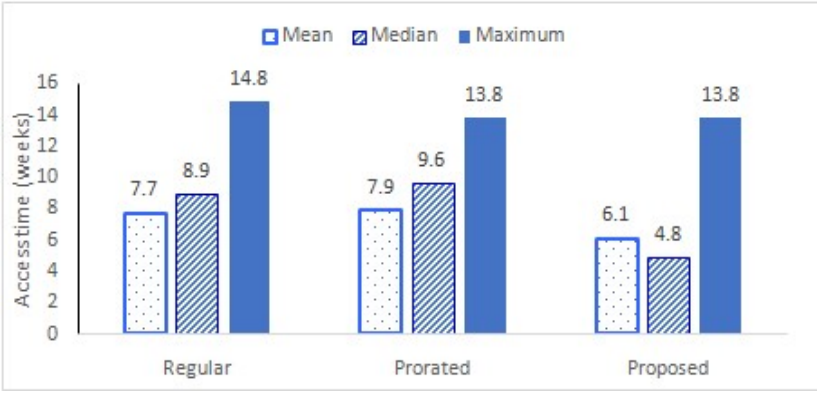


Fig. 4. Access time statistics for priority 3 patients in scenario (MED, H, H)

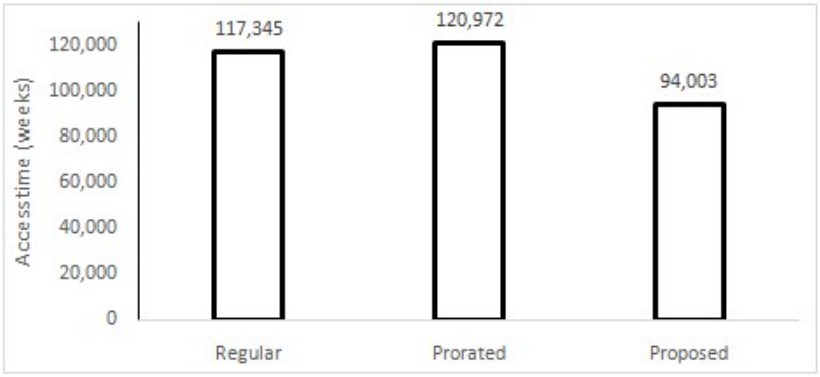


Fig. 5. Total access times for priority 3 patients in scenario (MED, H, H)

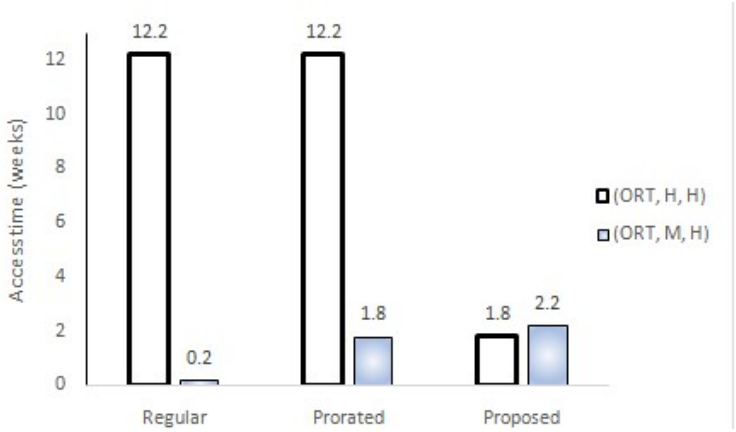


Fig. 6. Total access times for priority 2 patients in scenario (ORT, H/M, H)

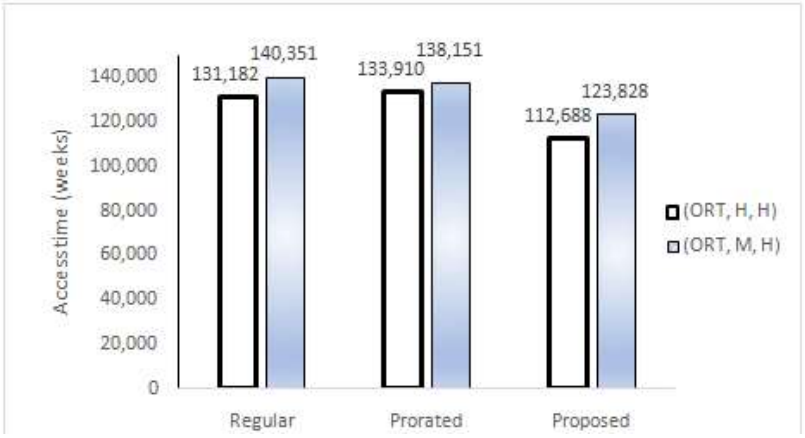


Fig. 7. Total access times for priority3 patients in scenario (ORT, H/M, H)

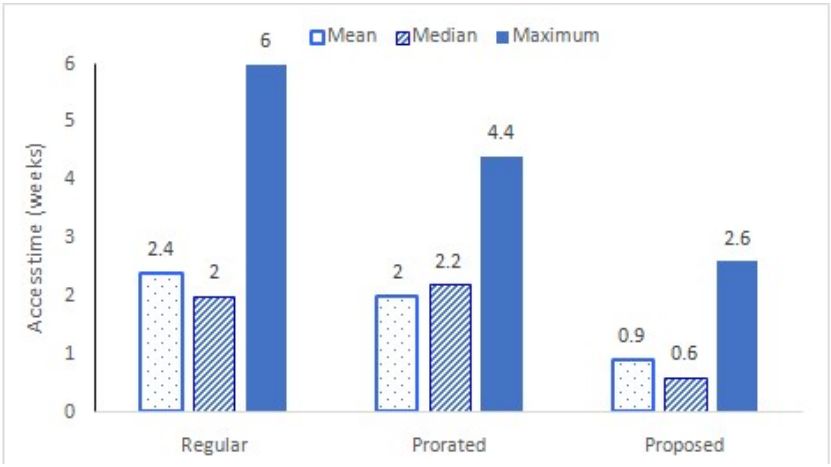


Fig. 8. Access time statistics for priority 3 patients in scenario (PSY, H, H)

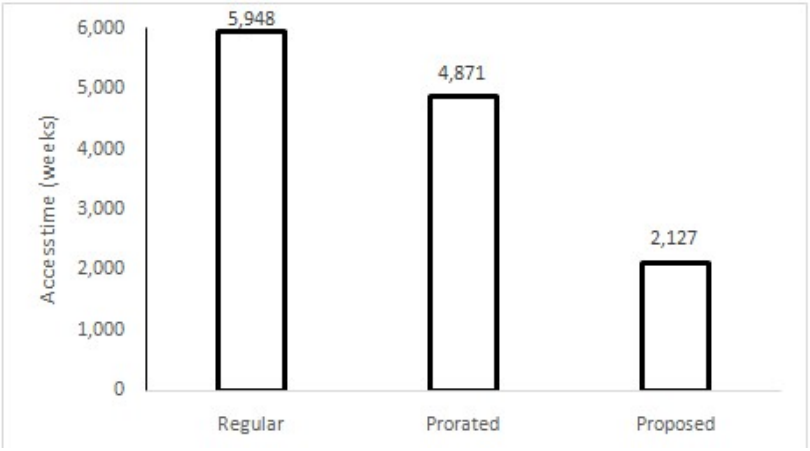


Fig. 9. Total access times for priority 3 patients in scenario (PSY, H, H)

5.3.2. Provider-centric measures

Two measures from the providers' perspective are evaluated: capacity utilization in the 247-day planning horizon and additional days required to complete all appointments. Table 4 shows the relative comparison among the three capacity strategies.

Table 4
Comparison of capacity strategies on provider-centric measures

Scenario (Clinic, monthly variation, weekday variation)	Capacity utilization (%) in 247-day planning horizon			Additional time (days) beyond 247-day planning horizon		
	Regular	Prorated or Proposed	Cost savings in improved utilization (HK$)	Regular	Prorated or Proposed	Cost savings per doctor (HK$)
(MED, H, H)	96.5	98.5	783,043.20	59	54	27,410.89
(MED, M, H)	100	100	0	51	51	0
(MED, L, H)	98.5	100	587,282.40	54	51	16,284.53
(MED, H, L)	96.6	98.5	743,891.04	59	51	43,425.43
(MED, M, L)	100	100	0	51	51	0
(MED, L L)	98.6	100	0	54	51	16,284.53
(ORT, H, H)	98.2	99.9	218,262.15	88	84	21,712.71
(ORT, M, H)	100	100	0	84	84	0
(ORT, L, H)	100	100	0	84	84	0
(ORT, H, L)	98.2	100	231,101.10	88	84	21,712.71
(ORT, M, L)	100	100	0	84	84	0
(ORT, L, L)	100	100	0	84	84	0
(PSY, H, H)	91.4	95.4	221,068.80	12	3	48,853.60
(PSY, M, H)	93.9	96.4	138,168	6	1	27,140.89
(PSY, L, H)	95.8	96.2	22,106.88	2	1	5,428.18
(PSY, H, L)	91.5	95.4	215,542.08	12	3	48,853.60
(PSY, M, L)	95.9	96.4	27,633.60	2	1	5,428.18
(PSY, L, L)	93.9	96.3	132,641.28	9	1	43,425.43

The proposed strategy and Prorated strategy recorded the same results in each scenario, and both perform better or no worse than the Regular strategy. The improvement in each measure over the Regular strategy is converted into cost savings based on the relevant cost data reported (Secretary for Food and Health, 2020).The average cost per attendance of a new case patient in each clinic (Table 2) enables the additional capacity utilized in the Prorated and Proposed strategies over the Regular strategy to be converted into cost saving. Similarly, the reported monthly median salary of doctors (Table 2) is projected to 12 months in the year (247 days). This allows the completion time (days) improved over the Regular strategy to be calculated as cost saving per doctor. Note that these cost savings could be over-estimates as the unused capacity (appointment time slots) in out-patient service could be used to serve other purpose in the hospital. Conversely, they could be under-estimates as public out-patient clinics typically consist of multiple doctors and other clinical staff. To summarize, while the service provider's workload remains much the same at a high level (Table 4), cost savings could result when integrating the capacity and appointment scheduling to better match supply with demand.

6. Conclusions and future research

This paper proposed a simultaneous scheduling strategy for capacity and appointments which minimizes the sum of patient access times, weighted by priority class. It has common applications in sectors which allow short-term capacity to vary while maintaining the total capacity constant (or within an agreed limit)in a planning horizon. The range of access times in each priority class is also minimized by applying a simple optimal scheduling rule (Brucker &Shakhlevich, 2016). Accordingly, this implies other objective functions in non-decreasing order of job (appointment) completion times and weights could apply the proposed strategy for simultaneous scheduling. Improvement in capacity utilization and completion time can be translated into cost savings expected to be significant for professional services under limited manpower supply.

References

Aslani, N., Kuzgunkaya, O., Vidyarthi, N., &Terekhov, D. (2020). A robust optimization model for tactical capacity planning in an outpatient setting. Health Care Managemen Science, 24, 26–40.

Brucker, P., &Shakhlevich, N.V. (2016). Necessary and sufficient optimality conditions for scheduling unit time jobs on identical parallel machines. Journal of Scheduling,19, 659-685.

Cappanera, P., Visintin, F., Banditori, C., &Di Feo, D.(2019). Evaluating the long-term effects of appointment scheduling policies in a magnetic resonance imaging setting. Flexible Services and Manufacturing Journal, 31, 212-254.

Cayirli, T., Dursun, P., &Gunes, E.D. (2019). An integrated analysis of capacity allocation and patient scheduling in presence of seasonal walk-ins. Flexible Services and Manufacturing Journal,31, 524-561.

Deglise-Hawkinson, J., Helm, J.F., Huschka, T., Kaufman, D.L., &Van Oyen, M.P. (2018). A capacity allocation planning model for integrated care and access management. Production and Operations Management,27, 2270-2290.

European Union, (2020), Working hours. [Online] Available: https://europa.eu/youreurope/business/human-resources/working-hours-holiday-leave/working-hours/ (May 4, 2020)

Gall. G. (1996). All year round: the growth of annual hours in Britain. Personnel Review, 25, 35-52.

Hung, R. (1999). Scheduling a workforce under annualized hours. International Journal of Production Research, 37, 2419-2427.

Kramer, A. Dell'Amico, M. Feillet, D., &Iori, M. (2020). Scheduling jobs with release dates on identical parallel machines by minimizing the total weighted completion time. Computers & Operations Research,123, article 105018.

Lasatre, K. B., Aiken, L.H., Sloane, D. M., French, R., Reneau, M. B., Alexander, M., & McHugh, M. D. (2020). Chronic hospital nurse understaffing meets COVID-19:an observational study. BMJ Quality & Safety, 0, 1-9.

Leeftink, A.G., Vliegen, L.M.H., &Hans, E.W. (2019). Stochastic integer programming for multi-disciplinary outpatient clinic planning. Health Care Management Science, 22, 53-67.

Lin, C.K.Y. (2019). Dynamic appointment scheduling with forecasting and priority-specific access time service level standards. Computers & Industrial Engineering, 135, 970-986.

Ma, Joanne, (2019), Overworked HK doctors and nurses protest against being overworked and understaffed. South China Morning Post. [Online] Available: https://www.scmp.com/yp/discover/news/hong-kong/article/3060283/overworked-hk-doctors-and-nurses-protest-against-being (January 21, 2019)

Nahmias, S., & Olsen, T.L. (2015). Production and operations analysis: strategy, quality, analytics, application.(7th ed.). Illinois: Waveland Press, Inc., (Chapter 3).

Nguyen, T. B. T., Sivakumar, A. I.,& Graves, S. C. (2018). Capacity planning with demand uncertainty for outpatient clinics. European Journal of Operational Research, 267, 338-348.

Portoghese, I., Galletta, M., Coppola, R.C., Finco, G., &Campagna, M. (2014). Burnout and workload among health care workers: the moderating role of job control. Safety and Health at Work,5, 152-157.

Ryan, L., &Wallace, J. (2019). Mutual gains success and failure: two case studies of annual hours in Ireland. The Irish Journal of Management, 38, 26-37.

Secretary for Food and Health, (2018), Replies to initial questions raised by Finance Committee Members in examining the estimates of Expenditure 2018-19. [Online] Available: https://www.legco.gov.hk/yr17-18/english/fc/fc/w_q/fhb-h-se.pdf(2018)

Secretary for Food and Health, (2019), Replies to initial questions raised by Finance Committee Members in examining the estimates of Expenditure 2019-20. [Online] Available:https://www.legco.gov.hk/yr18-19/english/fc/fc/w_q/fhb-h-e.pdf (2019)

Secretary for Food and Health, (2020), Replies to initial questions raised by Finance Committee Members in examining the estimates of Expenditure 2020-21. [Online] Available: https://www.legco.gov.hk/yr19-20/english/fc/fc/w_q/fhb-h-e.pdf (2020)

Van der Veen, E., Hans, E. W., Veltman, B., Berrevoets, L. M., &Berden, H. J. J. M. (2015). A case study of cost-efficient staffing under annualized hours. *Health Care Management Science,* 18, 279-288.

Wong, Ming-yan, Sharon,(2012), Management of access to Hong Kong public specialist out-patient services. The University of Hong Kong. [Master thesis online] Available: http://hub.hku.hk/handle/10722/179938(2012)