MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220222

Simple Heuristics for Near-Optimal Appointment Scheduling in Primary Care

Prashant Meckoni[a], Hari Balasubramanian[a]

[a] *Department of Mechanical & Industrial Engineering, University of Massachusetts, Amherst, MA, USA,*

Abstract

In primary care allocating appointments to sequential requests can result in sub-optimal scheduling. Optimal scheduling requires hiring of consultants to analyze historical patterns. Many practices focus their resources on larger problems instead of optimizing appointment schedules. We simulate simple heuristics to compare their performance with optimal schedules uncovered using offline optimization models. We use uncapacitated appointment calendars for a nationally representative heterogeneous primary care panel to meet all patients' requests. The stochastic nature of appointment requests gives a distribution for daily appointments and for the uncovered optimal capacity. The First Minimum heuristic gives near-optimal schedules and can be easily implemented in small practices using pen-and-paper, without any investment in computer-systems.

Keywords:

Appointments and Schedules, Heuristics, Primary Health Care.

Introduction

Behind the popularity of appointment systems for primary care is the expectation to be seen by the doctor at a predetermined time and minimizing the inconvenience of waiting in the hall. In many developed countries computer-based appointment systems allow easy reservation of patient appointments, frequently using self-service portals. Appointment systems give an advance notice of demand which helps predict staff workload. Though self-service based reservations allow users to select from the available choices, the human scheduler or front-desk staff still provides a higher confidence when it comes to making intelligent decisions and understanding patients' needs [1]. Such computer-based appointment systems are unavailable in primary care in low and middle-income countries [2]. In such primary care providers that try to provide appointments, the scheduling staff makes decisions that must be aligned to the provider's objectives. These scheduling decisions are frequently sequential. They also eliminate accidental double-booking. Yet, the sequential nature makes it difficult to know if the decisions are optimal. Schedulers are often provided guidelines or policies to help in decision making to help in specific objectives. Examples of such objectives include: (i) reduce delay and waiting time to improve patient satisfaction, (ii) improve utilization to reduce staff idle time, (iii) increase revenue by increasing number of patients. Often, these objectives try to manage conflicting goals. The stochastic nature of appointment requests makes it difficult to balance such conflicting objectives: (i) improving utilization can increase delays, (ii) increasing number of patients see can reduce patient satisfaction, (iii) reducing wait time can lead increase in staff idle time. Balancing conflicting objectives is susceptible to sub-optimal scheduling.

Markov Decision Processes are frequently used in determining an optimal scheduling policy for patient appointments [3–6], which require consultants and computer systems for proper implementation. Two stage stochastic programming [7] can help to make a broad scheduling template which again requires analysis by experts for implementation. Thus, optimal scheduling is out of reach for most practices and patients. In this paper, we will evaluate simple heuristics that allows near-optimal scheduling. These heuristics are simple enough to be implemented by a front-desk staff and can be used with paper-based appointment systems typically appointment calendars. These methods, in our opinion are particularly attractive to practices that do not have resources to implement computer-based appointment systems. In this paper, we use a US nationally representative patient panel to simulate appointment requests. These appointment requests are allocated based on the heuristics to an uncapacitated appointment calendar. We then use the full knowledge of all the requests to determine the optimal appointment allocation to the uncapacitated calendar, using offline optimization to evaluate the performance of the heuristics. The use of an uncapacitated model helps to focus on need for minimizing capacity to meet all patient requests.

Opportunity for optimal decisions

The patient-scheduler interaction holds the key to unlock optimal decisions. The scheduler discusses patients' needs and schedules appointments at can best fit the available slots. If the patient is flexible in their needs, decisions can be made that allow for better scheduling. Patients with rigid needs may not have many choices and may either experience delays or the provider may need to accommodate them at the expense of overtime. However, the number of patients with rigid healthcare needs are small in proportion, while most patients are relatively healthy and somewhat flexible. This flexibility by most of the patients is the essence in improved scheduling. We use the annual number of appointment requests as a surrogate measure for patient health status which influences that patient's flexibility. From the schedulers' point of view, there is no way of being certain about the sequence of patients' requests and their corresponding appointment preferences. What may seem like the right decision for allocating the previous patient appointment request, may turn out to be a bad decision when we encounter the subsequent patients' conflicting request. Though human schedulers cannot compete with complex algorithms that can use forecasting, stochastic processes and advanced optimization methods to come up with optimal schedules, they can understand patients' needs better than algorithms. Schedulers need easy guidelines and heuristics to implement policies that can give near-optimal schedules.

Methodology

We consider a panel of patients represented by set S associated with a primary care. This panel is partitioned into J different

classes. Each class indexed by j is represented by the set S_j. For any two classes j and j' we have $S_j \cap S_{j'} = \emptyset$ and $S = \bigcup_j S_j$ The panel size is $n = |S|$ and a class j has $n_j = |S_j|$ patients. The probability of request for appointment for each patient of class S_j on any day is p_j. We assume that the requests for successive appointments are independent of each other. The probability of requesting an appointment on any day for patients in class j is p_j. We assume that the appointment requests follow the Bernoulli distribution. So the number of days for the next appointment follows the Geometric Distribution. We denote X_j as the random variable that determines the number of days after the previous appointment.

Let $q_{(k,i)}$ be the requested day for appointment number i for patient $k \in S_j$. Let $a_{(k,i)}$ be the corresponding allotted day for request $q_{(k,i)}$. We have the next appointment request given in equation 1.

$$q_{(k,i)} = a_{(k,i-1)} + X_j \text{ where } X_j \sim \text{Geo}(p_j)\ k \in S_j \qquad 1$$

The appointment can then be scheduled on a day given by constraints in expression 2.

$$q_{(k,i)} - \delta_j \le a_{(k,i)} \le q_{(k,i)} + \delta_j \qquad 2$$
$$l_{(k,i)} \le a_{(k,i)} \le u_{(k,i)}$$

Here δ_j is the flexibility that a patient can tolerate in the deviation of the allotted day of appointment corresponding to the request. We use $l_{(k,i)} := q_{(k,i)} - \delta_j$ and $u_{(k,i)} := q_{(k,i)} + \delta_j$ for the lower and upper limits for $a_{(k,i)}$.

An optimal allocation of all appointments will try to minimize the capacity required over a long horizon. This minimum capacity is the least number of appointment slots needed to accommodate all patient requests within their flexibility tolerances. Any slots more than this minimum capacity can be provided at the expense of lower utilization than necessary. Though we intend to uncover this minimum capacity, we will also understand the distribution of slots utilized every day. We can have this minimum capacity as a hard capacity limit that the practice can work on. Analysis of daily slots utilized will help in setting up with an operating capacity that is less than the hard capacity. Any demand more than the operating capacity and within the hard capacity limit can be addressed using overtime policies. This can help to increase utilization and retain flexibility for infrequent surges in demand.

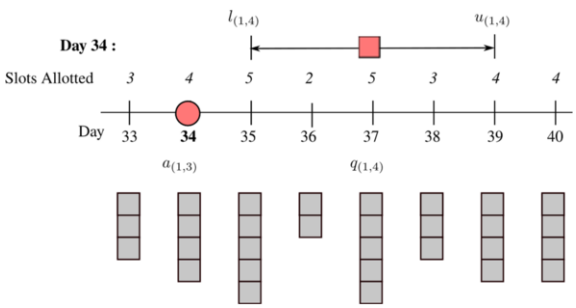

Figure 1 At the end of the third appointment of Patient-1 on day 34 shown by $a_{(1,3)}$, she requests for a next (fourth) appointment on day 37 denoted by $q_{(1,4)}$. She has a flexibility of 2 days, so her lower limit for the appointment is on day 35 denoted by $l_{(1,4)}$ and her upper limit is on day 39 denoted by $u_{(1,4)}$.

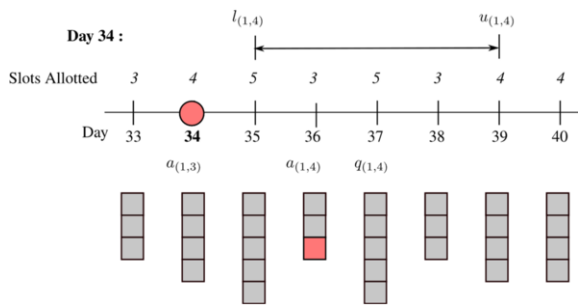

Figure 2 The scheduler looks up for the minimum slots allotted for other appointments on each of the days from day 35 to day 39. She finds that day 36 has the minimum number of slots allotted so far (two slots). This is the only day which has two slots. She allocates the request $q_{(1,4)} = 37$ on day $a_{(1,4)} = 36$. There are now three slots allotted on day 36.

Heuristics and Simulation

We simulate the appointment scheduling system for each patient in the panel. Each patient's appointment request is generated using equation 1. The appointment request is assumed to have come at the end of their previous appointment. All patients schedule their next appointment after their previous appointment. The request for appointment is only for a particular day. The appointment corresponding to $q_{k,i}$ is allotted such that it satisfies the constraints in expression 2. The simulation is initialized by randomly generating the first appointment request $q_{k,1}$ from equation 1 by assuming $a_{k,0} = 0$ for all patients.

We use three flexibility scenarios. The scenario "Hi" has the highest flexibility while the scenario "Lo" has the lowest flexibility for each patient class. The three scenarios that we will analyze are compared in the expression 3.

$$\delta_j^{Hi} \ge \delta_j^{Mid} \ge \delta_j^{Lo}\ \forall j \qquad 3$$
$$\delta_j \ge \delta_{j'} \forall j \ge j'$$

The decision for allocation of an appointment is based on the heuristic considered. The candidate heuristics do not require any complex calculations and thus can be easily replicated by a front desk staff. We have the following heuristics:

Uniform Random Any day in the interval $[l_{(k,i)}, u_{(k,i)}]$ is chosen with a same (uniform) probability. We consider this heuristic to see if random allocation can work better.

First Minimum The day with the minimum number of appointments in the interval $[l_{(k,i)}, u_{(k,i)}]$ is allotted. Ties are broken by selecting the earliest of such days. We consider this heuristic assuming we want to allocate most of the appointments as early

as possible, while balancing the number of appointments on all days.

Last Minimum The day with the minimum number of appointments in the interval $[l_{(k,i)}, u_{(k,i)}]$ is allotted. Ties are broken by selecting the latest of such days. This allows us to keep open earlier slots for frequently visiting patients that have less flexibility, while balancing the number of appointments on all days.

We can see an illustration of the first-minimum heuristic in Figure 1 where a patient makes an appointment request, and in Figure 2 where the scheduler allocates a slot to that request. The subsequent request for appointment is dependent on the previous appointment allotted as we see from equation 1. We record all the requests for appointments for optimal allocations.

At the end of each simulation, we analyze the number of appointments on each day. The number of appointments A_t on day t is in given equation 4.

$$A_t = \sum_{\substack{a_{(k,i)}=t, \\ k \in H}} 1 \qquad 4$$

The minimum capacity needed for allocation of each scenario is given in expression 5.

$$C^* = \max_t A_t \qquad 5$$

Offline Optimization

Offline optimization, in contrast to online optimization, considers all problem data available a priori. It is used to benchmark the online optimization or heuristics [8]. The requests are sequentially generated in the simulation runs. In offline optimization, all such appointment requests for all the patients corresponding to each simulation run are known before allocating appointments.

We propose an integer linear optimization problem to find the minimum capacity that is needed to allocate all the patient requests while satisfying the flexibility constraints. The integer linear optimization problem is given in problem 6.

$$\text{minimize } C \qquad 6$$

subject to

$$a_{(k,i),t} \leq C$$

$$\sum_{t=l_{(k,i)}}^{u_{(k,i)}} a_{(k,i),t} = 1 \; \forall (k,i)$$

$$C \geq 0$$

$$a_{(k,i),t} \geq 0, \text{Binary}$$

The tuple (k,i) represents the information related to ith appointment for patient k. The binary decision variable $a_{(k,i),t}$ indicates the appointment allotted related to request $q_{(k,i)}$ on day t in the interval $[l_{(k,i)}, u_{(k,i)}]$ as shown in the constraints of expression 2. Each appointment request is allotted exactly once. The sum of all appointments for each day t is restricted by the capacity C.

We can determine the daily sum of allotted appointments using equation 7 from the appointments allotted in the optimal solution.

$$A_t^* = \sum_{(k,i)} a_{(k,i),t} \quad \forall\, t \qquad 7$$

Experiment Evaluations

In the simulation we use a panel of 2000 patients. The panel constitutes of 20 classes of patients. This panel composition data has been taken from [9] which has in turn been summarized from MEPS [10]. We use equation 3 to construct ad-hoc flexibility for different patient classes for the simulation.

We evaluate each simulation and optimization run for 1250 days (5 years x 50 weeks /year x 5 days / week). Each flexibility-heuristic scenario is simulated 100 times with different random seeds. Thus, each scenario has 125,000 simulation days. With three heuristics and three flexibility tolerances, we evaluate nine scenarios. The simulation is run prior to the optimization to use the same appointment requests to preserve equivalence within each scenario. We capture the allotted appointments for further analysis.

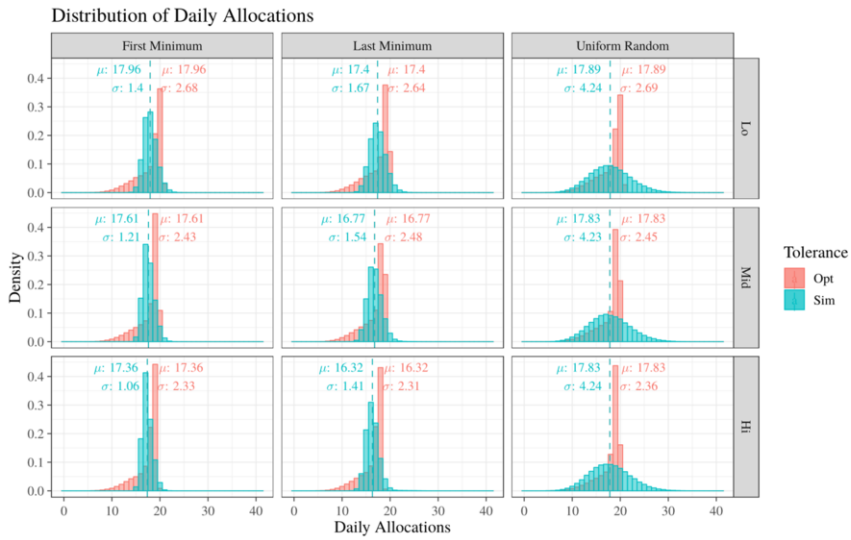

Figure 3 Histogram showing daily allotted appointments. The vertical axes represent the density, and the horizontal axes represent the number of allotted appointments each day. The heuristic's simulated daily allocation and the optimal daily allocation are shown in different colors for comparison. The mean of the daily allotted appointments is the same for the optimal allocation and the heuristic simulation. The heuristics have lower spread and thus lower variation than the optimal allocation. The optimal allocation has a negative skew while being capped at a lower value. We can also see a trend in the mean reducing as the flexibility increases from Lo to Mid to Hi. The "Last minimum" has lower mean and higher variability than the "First Minimum" heuristic.

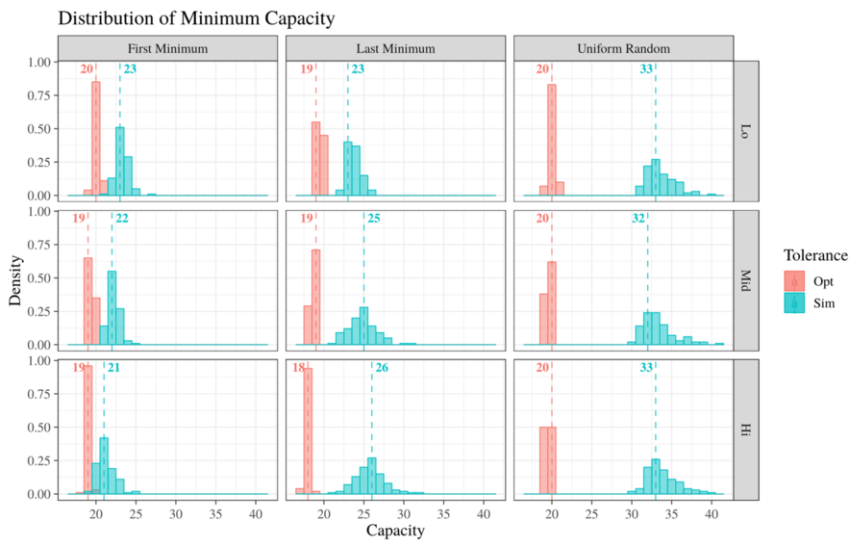

Figure 4 Histogram showing the distribution of the minimum capacity. The simulation and the offline optimization for each scenario are run 100 times to understand the distribution of the capacity needed. As the dashed lines indicate the mode of the distribution which is also noted in the chart. While the theoretical minimum capacity needed to accommodate all patients in the panel requesting for the exact same day is equal to the panel size, the probability of such an event is extremely low and has not been realized in the simulation.

Results

Daily allotted appointments

We can visually compare the daily appointments allotted in each heuristic-flexibility scenario in Figure 3. The optimal daily appointment allocated for each scenario have the same mean as the simulation. However, the simulation has a lower spread, since the optimal allocations are skewed to the left. So, an optimal allocation will have more days with less appointments than the different heuristics. The optimal allocation attempts to minimize the maximum number of appointments on each day while meeting all the appointment requests within their tolerance's flexibility. The optimal allocation does not optimize the variability, which is why the heuristics have a better variability than the optimal allocation. The First Minimum allocation heu-

ristic has a spread much closer to its corresponding optimal allocation, than the Last Minimum allocation heuristic. The spread for the "Uniform Random" heuristic is the worst, which implies that a random allocation within the flexibility tolerance of the patients is a bad idea.

As the flexibility increases from "Lo" to "Mid" to "Hi", the mean reduces for the "First Minimum" and for the "Last Minimum" heuristic. The flexibility scenarios also influence the spread. The spread for "Hi" flexibility is better than a "Lo" flexibility when we compare to their corresponding optimal allocations.

Capacity

Equation. 5 uses the simulation to get the minimum number of slots needed every day such that all appointments can be allotted. This can be considered as the minimum capacity to handle all requests. Theoretically, a practice would require the capacity at least equal to panel size since all patients can request for appointments on the same day. The probability of such an event will be extremely small, and thus such a large capacity would be practically unnecessary. Instead, the distribution of the discovered capacity may provider better insight. The histogram in Figure 4 compares the minimum required capacity of the simulations and the offline optimization. From all the flexibility scenarios, it is evident that capacity required for the "First Minimum" heuristic is the closest as compared to its corresponding optimal capacity and the "Last Minimum" is also a close contender.

Discussion

The use of scheduling heuristics may be a good-enough solution for many practices that have larger issues to spend their energies on. This is especially needed for small practices that have limited resources for any advanced scheduling. For practices that need more resources or those that need finer scheduling and capacity management, other optimal scheduling methods need to be explored. Online optimization may provide improvements over heuristics. The stochastic nature of appointment requests also beckons exploration of stochastic decision methods, especially dynamic programming and its approximations. The simplification of uniform flexibility for patients in the same class can be looked at closely, because it is this flexibility that, in the first place, allows near-optimal scheduling.

Conclusion

A simple solution that can be implemented immediately by appointment scheduling staff in primary care practices is to allocate each appointment request with the flexibility tolerances according to the first-minimum heuristic. Our solution does not require complex decision making. All it needs is accounting for the appointments allotted every day. The appointment calendar can be implemented using a simple paper diary. The First Minimum heuristic is near-optimal among the simple heuristic candidates.

References

[1] P. Zhao, I. Yoo, J. Lavoie, B.J. Lavoie, and E. Simoes, Web-Based Medical Appointment Systems: A Systematic Review, *J. Med. Internet Res.* **19** (2017) e6747. doi:10.2196/jmir.6747.

[2] S. Mendis, I. Al Bashir, L. Dissanayake, C. Varghese, I. Fadhil, E. Marhe, B. Sambo, F. Mehta, H. Elsayad, I. Sow, M. Algoe, H. Tennakoon, L.D. Truong, L.T.T. Lan, D. Huiuinato, N. Hewageegana, N.A.W. Fahal, G. Mebrhatu, G. Tshering, and O. Chestnov, Gaps in Capacity in Primary Care in Low-Resource Settings for Implementation of Essential Noncommunicable Disease Interventions, *Int. J. Hypertens.* **2012** (2012) e584041. doi:10.1155/2012/584041.

[3] Y. Gocgun, B.W. Bresnahan, A. Ghate, and M.L. Gunn, A Markov decision process approach to multi-category patient scheduling in a diagnostic facility, *Artif. Intell. Med.* **53** (2011) 73–81. doi:10.1016/j.artmed.2011.06.001.

[4] N. Liu, S. Ziya, and V.G. Kulkarni, Dynamic Scheduling of Outpatient Appointments Under Patient No-Shows and Cancellations, *Manuf. Serv. Oper. Manag.* **12** (2009) 347–364. doi:10.1287/msom.1090.0272.

[5] J. Patrick, A Markov decision model for determining optimal outpatient scheduling, *Health Care Manag. Sci.* **15** (2012) 91–102. doi:10.1007/s10729-011-9185-4.

[6] J. Wang, and R.Y.K. Fung, Adaptive dynamic programming algorithms for sequential appointment scheduling with patient preferences, *Artif. Intell. Med.* **63** (2015) 33–40. doi:10.1016/j.artmed.2014.12.002.

[7] S. Faridimehr, S. Venkatachalam, and R.B. Chinnam, Managing access to primary care clinics using scheduling templates, *Health Care Manag. Sci.* (2021). doi:10.1007/s10729-020-09535-z.

[8] P. Jaillet, and M.R. Wagner, Online Optimization - An Introduction, in: Risk Optim. Uncertain World, INFORMS, 2010: pp. 142–152. doi:10.1287/educ.1100.0072.

[9] M.C. Rossi, and H. Balasubramanian, Panel Size, Office Visits, and Care Coordination Events: A New Workload Estimation Methodology Based on Patient Longitudinal Event Histories, *MDM Policy Pract.* **3** (2018) 2381468318787188. doi:10.1177/2381468318787188.

[10] J.W. Cohen, S.B. Cohen, and J.S. Banthin, The Medical Expenditure Panel Survey: A National Information Resource to Support Healthcare Cost Research and Inform Policy and Practice, *Med. Care*. **47** (2009) S44–S50. doi:10.1097/MLR.0b013e3181a23e3a.

Address for correspondence

Hari Balasubramanian, hbalasub@umass.edu

University of Massachusetts Amherst

Mechanical and Industrial Engineering, 160 Governors Drive, Amherst, MA 01003