Improving Open Access Policy for Scheduling Outpatient Appointments

Balaraman Rajan
California State University East Bay
balaraman.rajan@csueastbay.edu

Abraham Seidmann
University of Rochester
avi.seidmann@simon.rochester.edu

Abstract
We present empirical and analytical results of our research that focuses on improving the Open Access (same day appointment) models for scheduling patients in a family practice setup. We show the value of dynamically incorporating the effect of anticipated behavior of patients on the expected load of the clinic. We find that many policies outperform OAP but when delay costs are introduced OAP is still robust when there is enough capacity at the clinic. We show that when both the capacity of the clinic is less than the demand and when the delay costs are significant, a simple forward looking policy does better than other current policies.

1. Introduction

In 2007, there were 994.3 million ambulatory patient visits to office based non-federally employed physicians [1]. Appointment system helps in capacity planning which in turn should ideally lead to better service to customers as well as better management of demand uncertainty. But, in reality, there is a huge problem of backlogs at most primary care centers [2]. The backlogs lead to two subsequent problems. One is that when a patient calls for an appointment, often there are no immediate appointments available and hence care is delayed. Second, patients end up scheduling their appointments months in advance to secure a slot with their physician. This in turn leads to uncertainty over patient availability on the day of their appointment and some patients either miss or cancel their appointment.

Because of inefficient appointment systems, no-shows and cancellations can be very significant in medical practices, up to 30% or more of all appointments [3]. These result in wasted physician capacity and can sometimes lead to lower utilization. Standard queuing models also ignore these aspects and do not incorporate them always. Administrators try to solve this by overbooking which has its own issues [4]. As a result, the appointment system results in unpredictable workload patterns. Even financially, it is reported that no-shows can be very expensive, £600 million a year for NHS [5]. Reminders and educating the patient have only solved the problem partially. Patient wait times for an appointment get longer with excessive backlogs. And this leads to increased no-shows, resulting in a vicious cycle [6]. Current literature has continued to investigate this trade-off to find an optimal appointment scheduling system.

Open Access Policy (OAP) tried to ameliorate the situation [7]. Very briefly, it provides same day appointment for all patients. It was found to dramatically reduce waiting time to get an appointment. It also increased the satisfaction levels for patients, physicians, and staff. There was less wasted capacity for physicians and higher levels of continuity of care. It also resulted in less support staff needed. But it was not universally effective. The performance of OAP was dependent on the ratio of capacity and demand. More importantly, it necessitated the need for physicians and nurses to work overtime when required. So, there were concerns on how the providers could handle such heavy loads.

Broadly, there are three major decisions in the case of patient scheduling system. One is the booking horizon – how many days in advance should a clinic accept or schedule patient visits. Second is the number of open slots – how many slots should the clinic keep open for emergency or walk-in appointments. Third is the ratio of appointment slots – how many slots for outpatient and inpatient categories. In this work, we would focus on two key decisions – how many patients to schedule each day and which patients to schedule each day.

Robinson and Chen [9] find optimal and detailed appointment slots/day by considering three different costs: Direct waiting cost (after arrival at the clinic) for patients, Overtime costs for the physician and Idle time costs for the physicians. They also consider same day or next day policy and find what is best when the system is at capacity or, otherwise. But they do not consider no-shows as a function of appointment lead time.
Liu et al [8] find MLE estimates for no-show and cancellation probabilities. They formulate a multiple day appointment problem as a Markov Decision Process (MDP) and consider cost of scheduling with overtime cost and revenue for seeing the patient. They compare OAP with many other policies and prove that for a linear reward function, the Two Day Probabilistic Policy (TPSP) is in fact reduced to the Next Day Policy (NDP).

2. Modeling Details

In this work, we focus on two key decisions – how many patients to schedule each day and which patients to schedule each day. Similar to Liu et al [8], we make the following assumptions. All patients call in advance to book appointments. The arrival rate of patient appointments follows a Poisson distribution. Patients are flexible for the day of appointment, that is patients always accept the appointment dates given to them by the scheduling algorithm. Service times are constant and patients will either: Show-up on the day of their appointment or, Cancel before their appointment day or, Not show up on the day of their appointment.

An efficient appointment system faces the following challenges. The appointment system should ideally balance the needs of the patients calling for care, and the physicians’ office economics. It should also incorporate the stochastic behavior of patients which includes their random arrival pattern, cancellations and no-shows. The proposed scheduling policy should also be easy to understand and implement at the clinical office. In other words, the algorithm we employ should limit the curse of dimensionality and should have tractability.

We have the following parameters in our model:

- **T** – Appointment Horizon
- **(i,j)** – Index for a patient who called i days before from today and has been given an appointment j days later from today
- **a** – Probability that a patient who is currently of type (i,j) will show up for the appointment
- **b** – Probability that a patient who called i days before from today would not have canceled the appointment j days later from today
- **X** – Number of patients of type (i,j) at the beginning of the day
- **Y** – Patient scheduling decision on day t. Represents how many patients are scheduled j days from now(t)
- **M** – Capacity of the clinic (number of patients that can be seen in a day)

Based on Binomial distribution we can then calculate the actual number of patients scheduled for each day and seen each day. For example, given **X** the number of patients on schedule the next day will be given by:

\[
X_{ij}^{t+1} = \begin{cases} 
B(Y_{t+1}, \beta_{01}) & \text{if } i = 1, j = 0, ..., T - 1 \\
B(X_{i-1, j+1}, \beta_{-1, 1}) & \text{if } 2 \leq j \leq T - i 
\end{cases}
\]

where **B(n,p)** represents the Binomial random variable with parameters **n** and **p**. The number of patients seen on day t will then be given by:

\[
U_t = \begin{cases} 
B(Y_0, a_{00}) & \text{if } i = 0 \\
B(X_{i0}, a_{i0}) & \text{if } i = 1, 2, ..., T 
\end{cases}
\]

We consider the clinic to receive one unit of revenue from seeing each patient, but no-shows or cancellations do not yield any revenue. The revenue can be thought of as net revenue after adjusting for any marginal cost of seeing each patient.

The indirect cost for the clinic is based on the actual number of patients scheduled for that particular day. Thus, while a no-show results in an indirect cost, since it was originally on schedule at the beginning of the day, it does not yield any revenue. For example, support staff and other costs depend on the schedule. We also include an overtime cost which occurs if the system is over loaded because of the schedule. So, our cost function takes the form

\[
w(x) = \begin{cases} 
K + h_1x & \text{if } x \leq M \\
K + h_1M + h_2(x - M) & \text{if } x > M 
\end{cases}
\]

where **K** is the daily fixed cost for the clinic, **h_1** is the cost per patient during regular time and **h_2** is the cost per patient during overtime.

We also consider a delay cost based on the original motive behind OAP. Patients call the clinic for a reason which may be an annual checkup, for immunization (flu or Tetanus) or it could for a painful sore throat with high fever. Obviously, not all of them carry the same urgency. For some patients there is a significant cost in delaying the patient’s appointment. The clinic’s reputation is also at stake while giving appointments based on need. Hence the delay cost is imputed on the clinic (loss of reputation). Besides third party payors put pressure on family medicine clinics to reduce the use of urgent care clinics and ER during regular office hours for clinics. We consider the delay function to take the simple linear form of one unit per day of delayed care.

Hence in our model we consider four different costs: a fixed daily cost, cost of scheduling per patient, cost of overtime per patient if the schedule exceeds the capacity and the delay costs based on the actual appointment slots given to the patients. The total cost is then given by:

\[
h_1 \left( Y_0 + \sum_{i=1}^{T} X_{i0} \right) + h_3 \sum_{j=1}^{T} j Y_j , \quad \text{if } Y_0 + \sum_{i=1}^{T} X_{i0} \leq M
\]
where \( h_3 \) is the cost of delaying a patient appointment by 1 day.

A scheduling policy should be such that it maximizes the “profit” (net revenue-costs) for the clinic. However, solving the MDP problem suffers from the curse of dimensionality and hence we resort to simple heuristics policies.

We consider three heuristic policies that are widely being followed: Threshold policy (TP), Balanced Scheduling Policy (BSP) and Randomized Scheduling Policy (RSP) [8]. Threshold Policy (TP) considers a threshold for the number of patients to be seen each day and schedules each new patient for the earliest day with less than the fixed threshold. If no such day exists in the horizon (all the days equal or exceed the threshold), then patients are scheduled on the day with the fewest appointments. Balanced Scheduling Policy (BSP) spreads the appointments uniformly over the appointment horizon. Randomized Scheduling Policy (RSP) schedules each new patient on a particular day in the horizon based on some pre-determined probability distribution.

Liu et al [8] prove that if the revenue and cost functions are linear, then the optimal Two-day Probabilistic Static Policy (OTPSP), which optimally schedules the arriving patients over the same day and the next day, actually turns out to be a Next Day Policy (NDP), that is, schedule all the patients the next day of their arrival day.

### 3. Results and Insights from the base case

In our simulations below we assume that the appointment arrivals are Poisson with mean arrival of patients as 50 per unit time. We then consider several cases where the capacity of the clinic (\( M \)) is smaller and greater than the expected demand per day. We allow \( a_{ij} \) and \( \beta_{ij} \) to take the form described in Liu et al [8], take \( T = 15 \), \( h_1 = 0.5 \) and \( h_2 = 0.95 \). We also vary \( h_3 \) from 0 to 0.5. Each simulation was run for 11 batches with each batch consisting of 200 days and the first batch was used as a warm-up.

The following numerical results show that the optimal policy is dependent on the magnitude of the delay cost. Table 1 presents the percentage improvement of the specified policy over OAP. Numbers in red indicate that the values are negative, that is, the policy does worse than OAP.

<table>
<thead>
<tr>
<th>( h_1 )</th>
<th>NDP</th>
<th>TP</th>
<th>BSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.11</td>
<td>108.17</td>
<td>104.18</td>
</tr>
<tr>
<td>0.2</td>
<td>1.80</td>
<td>144.96</td>
<td>139.89</td>
</tr>
<tr>
<td>0.5</td>
<td>3.70</td>
<td>296.87</td>
<td>286.24</td>
</tr>
<tr>
<td>0</td>
<td>1.69</td>
<td>88.40</td>
<td>90.83</td>
</tr>
<tr>
<td>0.2</td>
<td>0.78</td>
<td>119.79</td>
<td>121.60</td>
</tr>
<tr>
<td>0.5</td>
<td>1.59</td>
<td>238.27</td>
<td>243.98</td>
</tr>
<tr>
<td>0</td>
<td>3.48</td>
<td>4.88</td>
<td>89.34</td>
</tr>
<tr>
<td>0.2</td>
<td>3.25</td>
<td>5.06</td>
<td>117.34</td>
</tr>
<tr>
<td>0.5</td>
<td>2.93</td>
<td>6.21</td>
<td>225.92</td>
</tr>
<tr>
<td>0</td>
<td>5.34</td>
<td>2.43</td>
<td>89.60</td>
</tr>
<tr>
<td>0.2</td>
<td>4.78</td>
<td>2.24</td>
<td>116.57</td>
</tr>
<tr>
<td>0.5</td>
<td>5.23</td>
<td>2.54</td>
<td>220.60</td>
</tr>
</tbody>
</table>

Table 1: Percentage improvement over OAP with \( h_3 = 0.05 \).

The following numerical results show that the Next Day policy performs well only when the delay costs are negligible. Otherwise, the Threshold Policy does better when the capacity of the clinic is close to the expected demand (Fig 1) unless the delay cost is significantly high.

![Figure 1: Comparison of policies when Capacity = Expected Demand (M = 50)](image)

But when the capacity of the clinic is less than the expected demand, OAP outperforms the other policies if the delay cost is significant (Fig 2). Again only when the delay costs are low enough, the Next Day policy performs better than the other two policies.
The Next Day Policy does better only when the capacity is more than the expected demand and the delay costs are small enough. Otherwise, the Threshold Policy is still better than other policies (Fig 3).

We then also address the case when delay costs are heterogeneous (Flu immunization vs Sore throat) and propose a Forward-Looking Heuristic Policy. We use a simple cut-off policy to check whether scheduling the patient the next day or later will be efficient and schedule accordingly.

In our numerical experiments, the Forward Looking Policy seems to work well especially when capacity is less than the expected demand (Fig 4).

4. Summary and Conclusion

In this work we explore two key decisions typically faced by a family medicine clinic – how many patients to schedule each day and which patients to schedule each day. We find that many policies outperform the recently popular Open Access Policy (OAP) or same day appointments. But when a cost for delaying patient appointments is introduced, OAP is still robust especially when there is enough capacity at the clinic.

But when both the capacity of the clinic is less than the demand placed on the clinic and when delay costs are significant, a simple forward looking policy does better than other currently followed heuristic policies. More investigation is needed to check the robustness of this policy under different settings.

5. References