Appointment Scheduling with No-Shows and Overbooking: A Multi-Server Model

Michael Pinedo
Department of IOMS
Stern School of Business
New York University
`mpinedo@stern.nyu.edu`

Christos Zacharias
Department of Management Science
School of Business Administration
University of Miami
`czacharias@bus.miami.edu`

High demand for medical appointments

- Patients experience difficulties in accessing medical care.

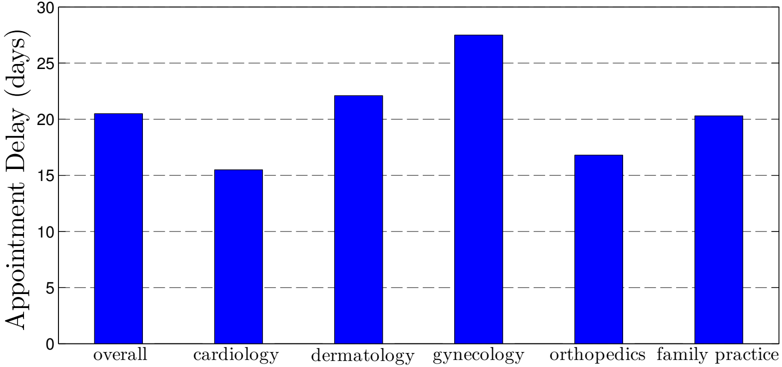


Figure: Merrit et al. (2011): Appointment delay for 1162 medical offices in 15 US metropolitan areas

High demand for medical appointments

BBC

News | Sport | Weather | Capital | Future | Sho

Home | US & Canada | Latin America | UK | Africa | Asia | Europe | Mid-East | Business | Health | Video

28 May 2014 Last updated at 16:13 ET

Share

US veterans waited 115 days for care

"significant delays in access to care negatively impacted the quality of care at this medical facility."

"VA guidelines say veterans should be seen within 14 days of their desired date for an appointment."

The New York Times

Why So Many V.A. Delays? Too Few Doctors, for Starters.

HUFF POST

"Cooking the books" at VA hospitals has exploded into public view since allegations arose that up to 40 patients may have died at the Phoenix VA hospital while awaiting care.

No-Shows

- Patient no-shows
    - **26% in Dermatology**, Perio and Niemeier (2011)
    - **21% in Psychotherapy**, Defife et al. (2010)
    - **30% in Obstetrics and Gynecology**, Dreiher et al. (2008)
    - **31% for MRI screening**, Green and Savin (2008)
    - **15%-51% in Mental Health**, Galucci et al. (2005)
    - ...
- Unattended appointments
    - Clinic under-utilization
    - Limit the access to other patients

Appointment Overbooking

No-Shows
$+$
High Demand $\Longrightarrow$ **Appointment Overbooking** $\Longrightarrow$ **Clinic Overcrowding**

Appointment Scheduling

1. How many patients to schedule?
2. How to allocate appointment slots throughout working day?
3. What is the optimal sequencing of heterogeneous patients?

Static & Sequential Scheduling

- **Static (Offline) Scheduling:** The set of customers to be scheduled and their characteristics are known in advance.
    - Kaandorp and Koole (2007)
    - Hassin and Mendel (2008)
    - Klassen and Yoogalingam (2009)
    - Robinson and Chen (2010)
    - Begen and Queyranne (2011)
    - Cayirli et al. (2011)
    - LaGanga and Lawrence (2012)...
- **Sequential (Online) Scheduling:** Requests for appointment come in gradually over time.
    - Muthuraman and Lawley (2008)
    - Zeng et al. (2010)
    - Liu et al. (2010)
    - LaGanga and Lawrence (2012)...

Overview of the Problem

- Patients are heterogeneous
  - Different no-show probabilities
  - Different weights
- Objective: minimize the weighted sum of
  1. Patients' waiting times
  2. Doctor's idle time
  3. Doctor's overtime
- Static Scheduling
- Sequential Scheduling

Contributions

Contributions

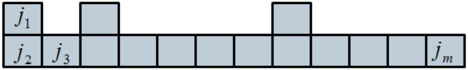
- Sensible practice of appointment overbooking can significantly improve the operational performance.
- Patient heterogeneity affects optimal schedule and should be taken under consideration.
- New sequencing rule is introduced.
- Heuristic solution is proposed for the online problem.

Roadmap

1. Static Scheduling
   a. Heterogeneous Patients
   b. Homogeneous Patients
2. Sequential Scheduling
3. Conclusion

Single Server Model

- Single server.
- n time slots available.
- m customers are scheduled to arrive, $m \geq n$.
- One time slot of service.
- Customer j will show up with probability $r_j = 1 - q_j$ exactly at the beginning of the time slot she was assigned.



Single Server Model

- **Three costs:**
  1. **Waiting cost:** w_j per time slot patient j has to wait.
  2. **Idle time cost:** c_I per time slot of idle server.
  3. **Overtime cost:** c_o per overtime slot.

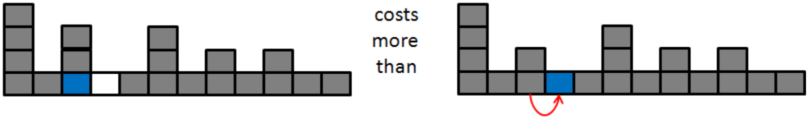
- **Objective:**

$$\min_{s} E[W(s) + I(s) + O(s)]$$

Structural Properties

Lemma

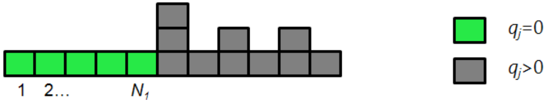
Each time slot has at least one customer assigned to it.



Structural Properties

Lemma

There exists an optimal schedule that assigns all customers from the set $C_1 = \{j : q_j = 0\}$ to the first $N_1 = |C_1|$ time slots.



Structural Properties

- Consider for now the problem where

$$c_I, c_O >> w_j \text{ for all } j.$$

- primary objective

Minimization of doctor's idle + overtime cost

$$\min_s E[I(s) + O(s)]$$

- secondary objective

Minimization of patients' waiting cost

$$\min_{s \in \mathcal{A}} E[W(s)]$$

$$\text{where } \mathcal{A} = \left\{ s' : s' = \operatorname*{argmin}_s E[I(s) + O(s)] \right\}$$

Sequencing Rules

Proposition

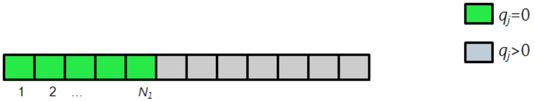
The appointment schedule within class $\mathcal{A}$ that minimizes expected waiting cost has the structure:

(a) Customers in set $\mathbf{C_1}$ are assigned to slots $\mathbf{1, 2, ..., N_1}$, no overbooking.

(b) $m - n + 1$ customers are assigned to slot $N_1 + 1$, prioritized in decreasing order of w_j.

(c) Remaining $n - N_1 - 1$ customers are assigned to slots $N_1 + 1, N_1 + 2, ..., n$ in increasing order of $z_j = \dfrac{w_j(1 - q_j)}{q_j}$.

(d) Customer with lowest weight assigned to slot $N_1 + 1$ has z index lower than that of customer assigned to $N_1 + 2$.



Sequencing Rules

Proposition

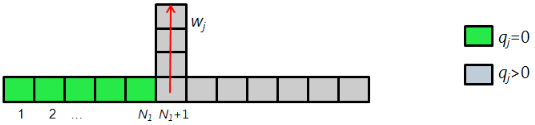
The appointment schedule within class $\mathcal{A}$ that minimizes expected waiting cost has the structure:

(a) Customers in set C_1 are assigned to slots $1, 2, ..., N_1$, no overbooking.

(b) $\mathbf{m - n + 1}$ customers are assigned to slot $\mathbf{N_1 + 1}$, prioritized in decreasing order of $\mathbf{w_j}$.

(c) Remaining $n - N_1 - 1$ customers are assigned to slots $N_1 + 1, N_1 + 2, ..., n$ in increasing order of $z_j = \dfrac{w_j(1 - q_j)}{q_j}$.

(d) Customer with lowest weight assigned to slot $N_1 + 1$ has z index lower than that of customer assigned to $N_1 + 2$.



Sequencing Rules

Proposition

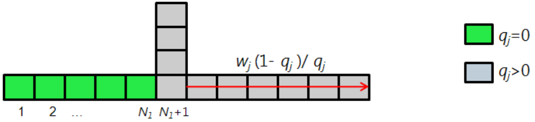
The appointment schedule within class $\mathcal{A}$ that minimizes expected waiting cost has the structure:

(a) Customers in set C_1 are assigned to slots $1, 2, ..., N_1$, no overbooking.

(b) $m - n + 1$ customers are assigned to slot $N_1 + 1$, prioritized in decreasing order of w_j.

(c) Remaining $\mathbf{n - N_1 - 1}$ customers are assigned to slots $\mathbf{N_1 + 1, N_1 + 2, ..., n}$ in increasing order of $\mathbf{z_j = \frac{w_j(1-q_j)}{q_j}}$.

(d) Customer with lowest weight assigned to slot $N_1 + 1$ has z index lower than that of customer assigned to $N_1 + 2$.



Sequencing Rules

Proposition

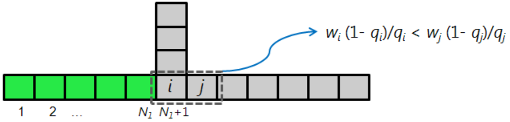
The appointment schedule within class $\mathcal{A}$ that minimizes expected waiting cost has the structure:

(a) Customers in set C_1 are assigned to slots $1, 2, ..., N_1$, no overbooking.

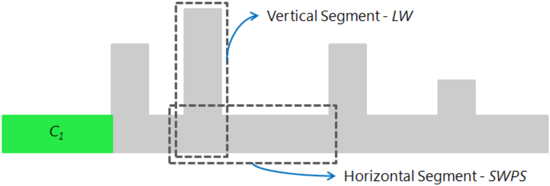
(b) $m - n + 1$ customers are assigned to slot $N_1 + 1$, prioritized in decreasing order of w_j.

(c) Remaining $n - N_1 - 1$ customers are assigned to slots $N_1 + 1, N_1 + 2, ..., n$ in increasing order of $z_j = \dfrac{w_j(1 - q_j)}{q_j}$.

(d) Customer with lowest weight assigned to slot $\mathbf{N_1 + 1}$ has z index lower than that of customer assigned to $\mathbf{N_1 + 2}$.



Sequencing Rules

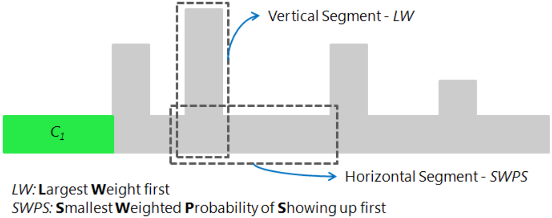


Sequencing Rules

Proposition

In an optimal schedule:

- The customers in any vertical segment of the optimal schedule are ordered in decreasing order of w_j.
- The customers in any horizontal segment of the optimal schedule are scheduled in increasing order of $z_j = \dfrac{w_j(1-q_j)}{q_j}$.



Sequencing Rules

Proposition

In an optimal schedule:

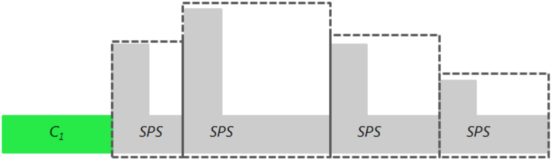
- The customers in any vertical segment of the optimal schedule are ordered in decreasing order of w_j.
- The customers in any horizontal segment of the optimal schedule are scheduled in increasing order of $z_j = \dfrac{w_j(1-q_j)}{q_j}$.

LW: **L**argest **W**eight first
SWPS: **S**mallest **W**eighted **P**robability of **S**howing up first

Sequencing Rules

Corollary

If $w_j = w$ for all $j = 1, 2, ..., m$, then all customers in a vertical segment and in the immediately following horizontal segment have to be scheduled in decreasing order of q_j.



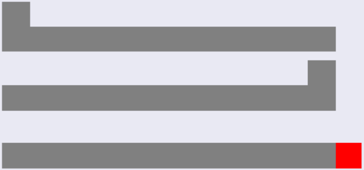
SPS: **S**mallest **P**robability of **S**howing up first

Offline Scheduling-Homogeneous Customers

- $w_j = w \; \forall j = 1, 2, ..., m.$
- $q_j = q \; \forall j = 1, 2, ..., m.$

Proposition

If $m = n + 1$ then the optimal schedule is

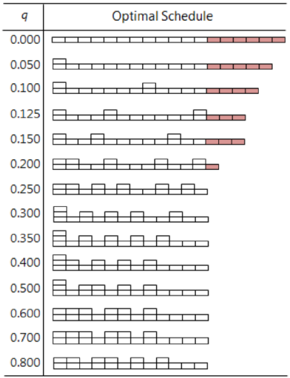


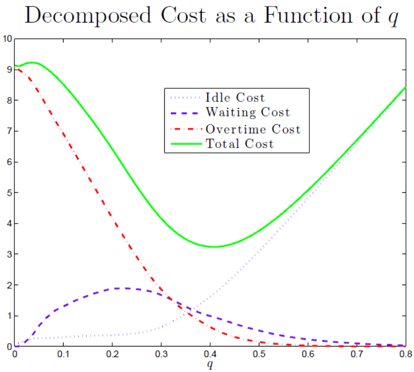
if $\frac{w(1-q)}{q} \leq c_I + c_0$

if $c_I + c_0 \leq \frac{w(1-q)}{q} \leq c_I + c_0(1 + \frac{1}{1-q})$

if $c_I + c_0(1 + \frac{1}{1-q}) \leq \frac{w(1-q)}{q}$

Numerical Experiments: Fixed m





- $n = 12$, $m = 16$.

Numerical Experiments: m subject to optimization

Regime	No-Show rate = 20%	No-Show rate = 30%
w=0.10		
w=0.15		
w=0.20		

Figure: Optimal Schedules: $n = 16$, $c_o = 1.5$, $c_I = 1$

Optimal Overbooking level y

- Poisson regression
- $x = (n, w, q, 1)$
- Model: $E[y|x] = e^{\beta^T x}$, for some $\beta \in \mathbb{R}^4$.

	Coefficient	Standard Error	z	$P > \|z\|$	95% Confidence Interval
n	0.15	0.02	9.68	0.00	[0.12,0.18]
w	-2.70	0.20	-13.60	0.00	[-3.09,-2.31]
q	8.36	0.28	29.44	0.00	[7.80,8.91]
1	-3.26	0.18	-17.72	0.00	[-3.62,-2.90]
$X^2_{\text{LR}} = 1588.91$			$P(\chi^2(3) > 1588.91) = 0.0000$		

Table: Poisson Regression

$$y = e^{\beta_1 n + \beta_2 w + \beta_3 q + \beta_4},$$

where $\beta = (0.15, -2.70, 8.36, -3.26)$.

Deterministic Vs Lognormal Service Times

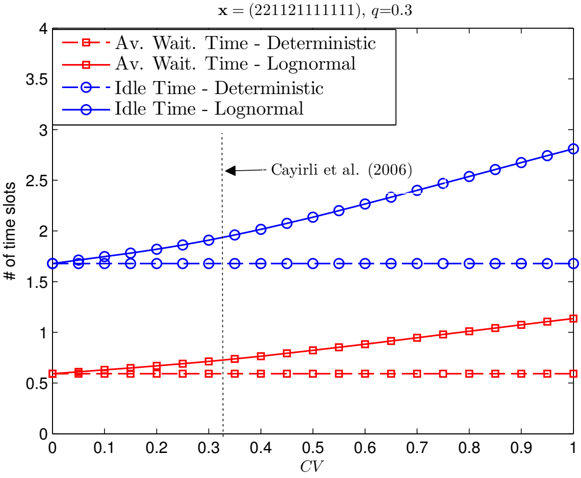


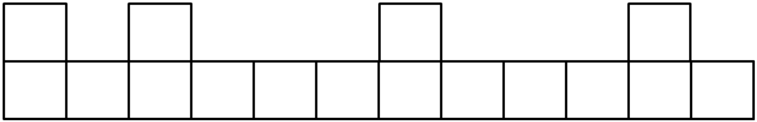
Figure: Deterministic Vs Lognormal Service Times

Deterministic Vs Lognormal Service Times

		$q = 0.2$	$q = 0.3$	$q = 0.4$
$w = 0.01$	$\mathbf{x}^*$	3 1 1 1 1 1 1 1 1 1 1 1	3 2 1 2 1 1 1 1 1 1 1 1	4 2 1 2 1 2 1 1 1 1 1 1
	$\mathbf{x}_l^*$	3 1 1 1 1 1 1 1 1 1 1 1	3 2 1 2 1 1 1 1 1 1 1 1	5 2 1 2 1 1 1 1 1 1 1 1
$w = 0.05$	$\mathbf{x}^*$	2 1 2 1 1 1 1 1 1 1 1 1	2 2 1 1 2 1 1 1 1 1 1 1	3 2 1 2 1 2 1 2 1 1 1 1
	$\mathbf{x}_l^*$	2 2 1 1 1 1 1 1 1 1 1 1	3 1 1 2 1 1 1 1 1 1 1 1	4 1 2 1 2 1 1 1 1 1 1 1
$w = 0.10$	$\mathbf{x}^*$	2 1 1 1 1 1 1 1 1 1 1 1	2 1 2 1 1 2 1 1 1 1 1 1	3 1 2 1 2 1 1 2 1 1 1 1
	$\mathbf{x}_l^*$	2 1 1 1 1 1 1 1 1 1 1 1	2 2 1 1 2 1 1 1 1 1 1 1	3 2 1 2 1 2 1 1 1 1 1 1
$w = 0.15$	$\mathbf{x}^*$	2 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 2 1 1 1 1 1 1 1	2 2 1 1 2 1 1 2 1 1 1 1
	$\mathbf{x}_l^*$	2 1 1 1 1 1 1 1 1 1 1 1	2 1 1 2 1 1 1 1 1 1 1 1	3 1 2 1 1 2 1 1 1 1 1 1
$w = 0.20$	$\mathbf{x}^*$	2 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 2 1 1 1 1 1 1 1	2 1 2 1 2 1 1 2 1 1 1 1
	$\mathbf{x}_l^*$	2 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 2 1 1 1 1 1 1 1	3 1 1 2 1 1 2 1 1 1 1 1
$w = 0.25$	$\mathbf{x}^*$	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 2 1 1 1 1 1 1	2 1 2 1 2 1 1 2 1 1 1 1
	$\mathbf{x}_l^*$	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1 1 1	2 2 1 1 1 1 2 1 1 1 1 1
$w = 0.30$	$\mathbf{x}^*$	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1 1 1	2 1 1 2 1 1 2 1 1 1 1 1
	$\mathbf{x}_l^*$	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1 1 1	2 1 2 1 1 1 2 1 1 1 1 1
$w = 0.40$	$\mathbf{x}^*$	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1 1 1	2 1 1 2 1 1 1 2 1 1 1 1
	$\mathbf{x}_l^*$	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 2 1 1 1 1 1 1
$w = 0.50$	$\mathbf{x}^*$	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 2 1 1 1 1 1 1
	$\mathbf{x}_l^*$	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1 1 1
$w = 0.60$	$\mathbf{x}^*$	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1 1 1
	$\mathbf{x}_l^*$	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1 1 1
$w = 0.70$	$\mathbf{x}^*$	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1
	$\mathbf{x}_l^*$	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1

Table: Deterministic vs Lognormal Service Times

Sequential Scheduling



Online Scheduling-Proposed Heuristic

Phase I

- The design of the scheduling framework
- Depends of the clinic's and patients' characteristics.

Phase II

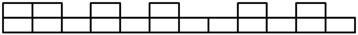
- Schedule Generation
- Gradual appointment booking based on the framework established in Phase I.

Phase I - Design of the schedule

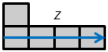
Phase I

Step 1: Determine target number of customers to overbook.

Step 2: Determine specific time slots for overbooking.



Step 3: Determine appropriate ranges of index values for each slot according to *SWPS*.

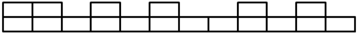


Phase I - Design of the schedule

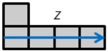
Phase I

Step 1: Determine target number of customers to overbook.

Step 2: Determine specific time slots for overbooking.



Step 3: Determine appropriate ranges of index values for each slot according to *SWPS*.



$$\text{Schedule patient } j \text{ at } \begin{cases} 1 & \text{if} \quad 0.00 \leq F_Z(z_j) \leq 0.25 \\ 2 & \text{if} \quad 0.25 < F_Z(z_j) \leq 0.50 \\ 3 & \text{if} \quad 0.50 < F_Z(z_j) \leq 0.75 \\ 4 & \text{if} \quad 0.75 < F_Z(z_j) \leq 1.00 \end{cases}$$

Proposed Heuristic: Example

- $n = 12$
- average weight $\bar{w} = 0.2$ and average no show probability $\bar{q} = 0.3$.
- $W = \begin{cases} w_L & \text{w.p.} \quad 0.5 \\ w_H & \text{w.p.} \quad 0.5 \end{cases}$ and $Q = \begin{cases} q_L & \text{w.p.} \quad 0.5 \\ q_H & \text{w.p.} \quad 0.5 \end{cases}$

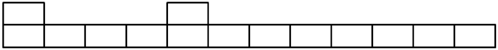
Proposed Heuristic: Example

- $n = 12$
- average weight $\bar{w} = 0.2$ and average no show probability $\bar{q} = 0.3$.
- $W = \begin{cases} w_L & \text{w.p.} \quad 0.5 \\ w_H & \text{w.p.} \quad 0.5 \end{cases}$ and $Q = \begin{cases} q_L & \text{w.p.} \quad 0.5 \\ q_H & \text{w.p.} \quad 0.5 \end{cases}$
- Therefore there are four types of customers. Type (i, j) customer corresponds to one with weight w_i and no-show probability $q_j, \quad i, j \in \{L, H\}$.
- $Z = W(1 - Q)/Q$ and $z_{ij} = w_i(1 - q_j)/q_j$
- $z_{LH} \leq z_{HH}, \;\; z_{LL} \leq z_{HL}$
- Assuming that W and Q are independent, Z takes each one of these values with probability $\dfrac{1}{4}$.

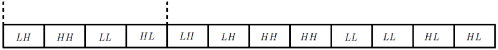
Proposed Heuristic: Example

Phase I

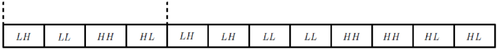
- Overbook $y = [e^{\beta_1 n + \beta_2 \bar{w} + \beta_3 \bar{q} + \beta_4}] = 2$ appointments.
- The scheduling framework is the optimal schedule for the homogeneous customers problem with $m = 14$, $w = 0.2$ and $q = 0.3$.



- We intend to fill the Horizontal segments according to the *SWPS* rule.



Step 3 of Phase I, $z_{HH} \leq z_{LL}$



Step 3 of Phase I, $z_{LL} \leq z_{HH}$

Proposed Heuristic: Example

Phase II

Proposed Schedule	**Baseline Schedule**
Assign the first n customers following the structure of Phase I	Assign the first n customers randomly (equivalent to uniform patient preferences)
Assign the last $m - n$ requests to the overbooking slots specified in Phase I	Assign the last $m - n$ requests to the overbooking slots specified in Phase I

We don't reject customers. If there is no proper slot available, then schedule to an adjacent slot.

Proposed Heuristic: Example

- We simulate 100,000 samples.
- Each sample consists of 14 consecutive requests for an appointment drown from Z.
- We evaluate the proposed schedules compared to the baseline schedules.
- $\Delta_w = w_H - w_L$, $\Delta_q = q_H - q_L$.

	% decrease in cost										
Δ_w \ Δ_q	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.00	0.00	1.34	2.81	4.16	5.35	6.16	8.11	9.40	10.03	10.99	14.52
0.05	0.86	1.35	2.64	4.08	4.56	5.33	5.19	8.00	8.66	9.10	12.25
0.10	1.42	2.27	2.02	3.19	4.14	4.41	6.44	8.22	5.27	9.49	11.66
0.15	2.01	2.37	3.20	3.75	3.84	6.39	6.04	7.66	9.29	7.22	11.91
0.20	2.62	3.88	3.53	4.35	7.45	5.92	8.41	5.99	6.98	11.05	8.08
0.25	3.40	4.19	4.39	5.34	4.99	6.72	8.38	9.76	10.80	11.71	10.67
0.30	4.66	4.51	5.61	6.97	6.65	7.53	7.04	6.50	9.22	11.55	13.79
0.35	4.95	6.42	6.51	6.87	6.00	4.86	7.55	9.13	8.81	9.62	13.71

Table: Performance of the Proposed Heuristic

Summary

- Overbooking model for scheduling arrivals under no-shows.
- Heterogeneous customers.
- Offline and Sequential Scheduling are considered.

Conclusions

- A sensible practice of appointment overbooking can effectively address the no-show phenomenon.
- No-show rates and patient heterogeneity affect the optimal schedule and should be taken under consideration.
- Front-loaded schedules with repeating patterns.
- Structural properties and a priority rule are introduced for the offline problem.
- A heuristic solution is developed for the sequential problem.

Limitations

In order to focus on no-shows we assumed

- Deterministic service times:
  - Have shown to perform very well (click here).
- Punctuality of arriving customers:
  - Blanco White and Pike (1964) compare patient's waiting times when customers are punctual or unpunctual and are shown not to differ significantly.

References

- Begen, M.A., M. Queyranne. 2011. Appointment scheduling with discrete random durations. *Math. of Oper. Res.*, **36**(2):240-257.
- Chakraborty, S., K. Muthuraman, M. Lawley. 2010. Sequential clinical scheduling with general service times and no-show patients. *IIE Transactions*. **42**(1): 1-13.
- Dreiher, J., M. Froimovici, Y. Bibi, D.A. Vardy, A. Cicurel, A.D. Cohen. 2008. Nonattendance in Obstetrics and Gynecology Patients. *Gynecologic and Obstetric Investigations*. **66**: 40-43.
- Green, L., S. Savin. 2008. Reducing delays for medical appointments: A queueing model. *Operations Research*. **56**(6): 1526-1538.
- Hassin, R., S. Mendel. 2008. Scheduling arrivals to queues: A single server model with no-shows. *Management Science*. **54**(3): 565-572.
- Kaandorp, G. C., G. Koole. 2007. Optimal outpatient appointment scheduling. *Health Care Management Science*. **10**(3): 217-229.
- LaGanga, L.R., S.R. Lawrence. 2007. Clinic overbooking to improve patient access and provider productivity. *Decision Sciences*. **38**(2): 251-276.
- LaGanga, L.R., S.R. Lawrence. 2012. Appointment overbooking in health care clinics to improve patient service and clinic Performance. *Production and Operations Management*. To appear.
- Muthuraman, K., M. A. Lawley. 2008. A Stochastic Overbooking Model for Outpatient Clinical Scheduling with No-Shows. *IIE Transactions*. **40**(9): 820-837.
- Robinson, L. W., R. R. Chen. 2010. A comparison of traditional and open-access policies for appointment scheduling. *MSOM*. **12**(2): 330-346.
- Robinson, L. W., R. R. Chen. 2011. Estimating the implied value of the customer's waiting time. *MSOM*. **13**(1): 53-57.
- Rust, C.T., N.H. Gallups, W.S. Clark, D.S. Jones, W.D. Wilcox. 1995. Patient appointment failures in pediatric resident continuity clinics. *Archives of Pediatrics & Adolescent Medicine*, **149**(6): 693-695.
- Zeng, B., A. Turkcan, J. Lin, M. Lawley. 2010. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Annals of Operations Research*. **178**: 121-144.