Evaluation of appointment scheduling rules: 
a multi-performance measures approach

Stefan Creemers
Pieter Colen
Marc Lambrecht

Abstract - Appointment scheduling rules are used to determine during which service session and at what time a customer is to receive service. Many different appointment scheduling rules have been devised and are being used in practice (e.g. in healthcare, legal services, administration and many other service and manufacturing industries). Which appointment scheduling rule is best however, is still an open question. In order to answer this question, we develop an analytical model to assess the performance (w.r.t. customer waiting time, server idle time and server overtime) of appointment scheduling rules in a wide variety of settings. More specifically, the model takes into account: (1) customer unpunctuality; (2) no-shows; (3) service interruptions; (4) delay of the service process. In addition, no restriction are imposed on the distributions used to capture the basic processes (i.e. the model is not limited to the use of exponential distributions). The model builds on matrix analytical methods and adopts an efficient algorithm (in terms of computational and memory requirements) to assess the performance of 314 appointment scheduling rules. Data envelopment analysis is used to compare the results of these appointment scheduling rules.

Keywords - Markov chain, appointment scheduling rules, data envelopment analysis

1 Introduction

Professionals in healthcare and other services (such as attorneys, faculty receiving students, tax accountants, consultants, barbers, automobile service centers, trailers at receiving bays and many others) face the problem of
allocating time windows to customers. This can be done by use of an appointment scheduling rule (ASR). Appointment scheduling rules (ASR) are used to determine during which service session and at what time a customer is to receive service. Hospitals (outpatient departments) and general practitioner’s offices are obvious examples of the application field of ASR. The growing importance of product service systems is another illustration: manufacturers involved in aftermarket services have to allocate technician time to service customers with a service contract (replacement of parts, scheduled maintenance, overhauls, performance based arrangements, . . . ). Although the literature on appointment scheduling rules is mainly focussed on healthcare, the research topic is generic and applicable in many sectors of the service economy.

Conducting scheduled appointments on time is becoming ever more important across service industries: timeliness of appointments is a key concern both for patients seeking treatment [6] and customers waiting for field service [1]. Patients/customers prefer short time windows during which they will get service. The customer waiting time consequently is a relevant performance measure. A second important objective of appointment policies has to do with the efficiency of the service. For private companies, the impetus to efficiency comes naturally. But also healthcare systems are under pressure to use their capacity effectively and efficiently. The doctors’ idle time and the doctors’ overtime (underutilization and overloads) are hereby important performance measures. The objective of this article is to identify appointment scheduling rules that minimize patients’ waiting time and minimize doctors’ idle time and overtime. This has to be done in an environment where both demand and supply characteristics are highly uncertain and subject to many sources of variability. Identifying the best appointment scheduling rule is the research question of this paper.

In the literature, many ASR are suggested. A good overview of these scheduling rules can be found in [7] and [3]. Let’s briefly introduce the most common ASR. First, we have the individual appointment scheduling rule in which each individual patient gets a scheduled arrival time taking into account the impact of the standard deviation of the service time (in this ASR it is allowed to assign more than one patient only at the start of a service session). A second type of ASR is the block appointment rule. Under this policy, we assign scheduled arrival times for blocks of patients. Third, we examine early-lateness (E-L), also known as variable arrival pace, appointment rules. Under such a policy we allow for an increasing arrival pace up
to a certain patient and afterwards the scheduled arrivals are characterized by a decreasing arrival pace (or vice versa). These rules will determine the planned (scheduled) arrival rate. The actual arrival time may of course be different from the planned arrival time. We, therefore, assign for each patient a probability of being too late or too early. We fix for each customer a probability of not showing up. Because of the no-show problem, i.e. patients not showing up for their appointment, the actual number of patient arrivals will be random even if the number of patients per session is fixed and predetermined. The performance of appointment scheduling rules is not only influenced by the arrival rate and service rate characteristics. Other types of outages during the service session are also important. We, therefore, will allow for delays at the start of a service session due to late arrival of the doctor or due to setup activities at the start of a session. We also allow preemptive and non-preemptive interrupts during the service session. All these extensions allow us to model real life appointment systems and identify ASR with a robust performance across different settings.

In order to answer our research question, we develop an analytical model that builds on matrix analytical methods [10]. Matrix analytical techniques (Markov Chains) allow us to deal with the complexity of the problem, while keeping the problem mathematically traceable. We use an efficient (in terms of computational and memory requirements) algorithm to assess the performance of appointment rules. The validity and accuracy of the model are supported by a simulation study. We use the model to assess the performance of a set of 314 ASR in an elaborate computational experiment. To compare the performance of these ASR (in terms of waiting time, idle time and overtime), we apply a data envelopment analysis (DEA) [4]. DEA is a well-established multi-dimensional performance evaluation tool: several performance dimensions have to be taken into account simultaneously in order to avoid selecting an ASR which performs very strongly in one dimension, e.g. waiting time, while performing very badly in others.

The contribution of this paper is threefold: (1) we develop a new analytical model to assess the performance of an appointment scheduling rule in very general settings; (2) we perform an elaborate computational experiment to assess the performance of a large number of ASR in a wide variety of settings; (3) we use DEA to identify the best ASR.

This paper is organized as follows. Section 2 provides a general description of the model. Due to conciseness concerns we do not report the mathematical formulations in this conference proceeding. The results of the computational
experiment are discussed in Section 3. We conclude in section 4.

2 Model Description

In this section we briefly discuss the properties of the procedure we use. The analytical details of our approach are omitted due to conciseness concerns. First, we define a Markov chain based model that can be efficiently solved by use of matrix analytical models. The model itself provides us with performance measures concerning the service execution process.

With respect to ASR, comprehensive comparisons of various rules are available, e.g. [8, 9]. In this article, we will focus only on static ASR, i.e. decisions are made prior to the start of the service session. The Markov chain based model has the following properties:

- Customers are served by a single server.
- Customers have i.i.d. service time distributions.
- All customers that arrive during the service session are served.
- Service is provided even if only a single customer is to be served.
- Customers that arrive early (i.e. prior to their scheduled arrival time) receive service if the server is idle. Remark that this implies the possibility of overtaking other customers.
- The arrival process can differ for each individual customer.
- Optimization occurs over a limited set of ASR.

Contrary to other existing models, we allow for an individual characterization of the arrival process for each patient/customer. In addition, computational performance and model accuracy (and hence practical applicability) of our model significantly exceed the capabilities of comparable models in the literature on ASR. The performance measures of interest (calculated by means of the Markov chain based model) are: (1) the expected waiting time of a customer; (2) the expected amount of overtime a server performs; (3) the expected amount of time a server resides in an idle state. Our model may be used to obtain these performance measures for any given schedule of customers (i.e. the outcome of any given ASR or scheduling procedure).
this article however, we limit ourselves to the comparison of the performance of schedules generated by a set of 314 ASR.

Most ASR may be classified in terms of:

- $A_i$, the scheduled arrival time of customer $i$,
- $\mu^{-1}$, the mean service time requirement of a customer,
- $\sigma_i$, the standard deviation of the service time requirement of customer $i$,
- $N$, the number of customers that require scheduling.

We implemented a set of 314 ASR and use an analytical model to perform an extensive computational experiment in which the performance of these rules is assessed w.r.t. three performance measures in a wide variety of settings. The adopted set of ASR is an extension of the set of 50 ASR selected in [8, 9]. These ASR are common in daily practice or have been shown to yield good and robust results [8, 9].

The ASR may be summarized as variations of: (1) the individual appointment rule; (2) the block appointment rule; (3) early-lateness rules (hereafter referred to as the EL rules).

The individual appointment rule schedules the arrival times of customers as follows:

$$
A_i = \begin{cases} 
a \mu^{-1} & \forall i : i < l, \\
A_{i-1} + \mu^{-1} + h \sigma_i & \forall i : l \leq i < N. 
\end{cases}
$$

Where: (1) $a$ is a multiplier to delay the start of the first arriving customers; (2) $l$ denotes the number of customers scheduled for arrival at the start of a service session; (3) $h$ is a multiplier used to adjust the impact of $\sigma_i$.

We implemented 91 variants of the individual appointment rule by allowing parameters $a$, $l$ and $h$ to vary over set $\{0, 0.3, 0.5\}$, set $L = \{1, 2, 3, 4, 5\}$ and set $H = \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ respectively.

The block appointment rule may be summarized as follows:

$$
\begin{align*}
A_i &= 0 & \forall i : i < b, \\
A_{nb} &= A_{nb} & \forall n : 1 \leq n < \frac{N}{b}, \\
A_{nb+i} &= A_{nb} & \forall i : 1 \leq i < b.
\end{align*}
$$
Where $b$ denotes the block size (i.e. the number of customers assigned to arrive at a single time instance). Varying parameters $b$ and $h$ over set $(\mathbf{L}\setminus\{1\})$ and set $\mathbf{H}$ respectively, we obtain 28 ASR.

The EL rules speed up and/or slow down the pace of scheduled arrivals using correction factors $r_1$ and $r_2$. The computation of scheduled arrival times is performed in two steps. First, all scheduled arrival times are initialized using the individual appointment rule where $(l = 1)$ and $(h = 0)$. Next, a correction is applied to speed up and/or slow down the pace of scheduled customer arrivals.

Initialization:
\[
A_0 = 0, \\
A_i = A_{i-1} + \mu^{-1} \quad \forall i : 1 \leq i < N.
\]

Correction:
\[
A_i = A_i - r_1(z - i)h\sigma_i \quad \forall i : 1 \leq i \leq z, \\
A_i = A_i - r_2(z - i)h\sigma_i \quad \forall i : z < i < N.
\]

(3)

Where: (1) $r_1$ and $r_2$ are correction factors used to speed up or slow down the succession of scheduled arrivals; (2) $z$ is any multiple of 5 smaller than $N$. Parameter $r_1$ controls the arrival pace of the first $z$ customers; the arrival pace of these customers increases as $r_1$ increases. Conversely, parameter $r_2$ controls the arrival pace of those customers that are scheduled to arrive after customer $z$; the arrival pace of these customers decreases as $r_2$ increases. When varying parameter $h$ over set $(\mathbf{H}\setminus\{0\})$ and parameters $r_1$ and $r_2$ over the set $\{0, 1, 2\}$ (where $(r_1 + r_2) > 0$), we obtain 39 times times $\lfloor \frac{N-1}{5} \rfloor$ ASR.

A summary of the ASR may be found in table 2.

We determine the performance of the ASR in a wide range of environmental settings. In the most simple environment, all customers arrive punctually at their assigned appointment time. Complexity is introduced in the form of so-called “environmental variables”. An extensive overview of such environmental variables is provided in [3]. We will take into account environmental settings that differ in the following parameters

- The number of patients/customers scheduled for each service session ($N$) can differ. We let $N$ take a value 10, 20 or 30.

- The service time may vary between customers and is uncertain, i.e. the squared coefficient of variation of the service time ($SCV(service)$) is larger than zero ($SCV(service) \in \{0; 2; 0.5; 1\}$).
Rule \( A_i = ia\mu^{-1}, \quad \forall i : i < l, \)
\[ A_i = A_{i-1} + \mu^{-1} + h\sigma_i, \quad \forall i : l, \]
Rule no. \( 1 - 7, 8 - 14, 15 - 21, 22 - 28, 29 - 35, \)
Conditions \( l = 1, 2, 3, 4, 5 \land a = 0 \land h \in \mathbb{H}, \)
Rule no. \( 36 - 42, 43 - 49, 50 - 56, 57 - 63, \)
Conditions \( l = 2, 3, 4, 5 \land a = 0.3 \land h \in \mathbb{H}, \)
Rule no. \( 64 - 70, 71 - 77, 78 - 84, 85 - 91, \)
Conditions \( l = 2, 3, 4, 5 \land a = 0.5 \land h \in \mathbb{H}, \)

Rule \( A_i = 0, \quad \forall i : i < b, \)
\[ A_{nb} = A_{(n-1)b} + b\mu^{-1} + h\sqrt{\sigma_{nb}}, \quad \forall n : 1 \leq n < \frac{N}{b}, \]
\[ A_{nb+i} = A_{nb}, \quad \forall i : 1 \leq i < b, \]
Rule no. \( 92 - 98, 99 - 105, 106 - 112, 113 - 119, \)
Conditions \( b = 2, 3, 4, 5 \land h \in \mathbb{H}, \)

Rule initialize \( A_0 = 0, \quad A_i = A_{i-1} + \mu^{-1}, \quad \forall i : 1 \leq i < N, \) then
\[ A_i = A_i - r_1(z - i)h\sigma_i, \quad \forall i : 1 \leq i \leq z, \]
\[ A_i = A_i - r_2(z - i)h\sigma_i, \quad \forall i : z < i < N, \]
Rule no. \( 120 - 125, 159 - 164, 198 - 203, 237 - 242, 276 - 281, \)
Conditions \( z = 5, 10, 15, 20, 25 \land r_1 = 0 \land r_2 = 1 \land h \in (\mathbb{H} \setminus \{0\}), \)
Rule no. \( 126 - 128, 165 - 167, 204 - 206, 243 - 245, 282 - 284, \)
Conditions \( z = 5, 10, 15, 20, 25 \land r_1 = 0 \land r_2 = 2 \land h \in \{0.2, 0.25, 0.3\}, \)
Rule no. \( 129 - 134, 168 - 173, 207 - 212, 246 - 251, 285 - 290, \)
Conditions \( z = 5, 10, 15, 20, 25 \land r_1 = 1 \land r_2 = 0 \land h \in (\mathbb{H} \setminus \{0\}), \)
Rule no. \( 135 - 140, 174 - 179, 213 - 218, 252 - 257, 291 - 296, \)
Conditions \( z = 5, 10, 15, 20, 25 \land r_1 = 1 \land r_2 = 1 \land h \in (\mathbb{H} \setminus \{0\}), \)
Rule no. \( 141 - 146, 180 - 185, 219 - 224, 258 - 263, 297 - 302, \)
Conditions \( z = 5, 10, 15, 20, 25 \land r_1 = 1 \land r_2 = 2 \land h \in (\mathbb{H} \setminus \{0\}), \)
Rule no. \( 147 - 149, 186 - 188, 225 - 227, 264 - 266, 303 - 305, \)
Conditions \( z = 5, 10, 15, 20, 25 \land r_1 = 2 \land r_2 = 0 \land h \in \{0.2, 0.25, 0.3\}, \)
Rule no. \( 150 - 155, 189 - 194, 228 - 233, 267 - 272, 306 - 311, \)
Conditions \( z = 5, 10, 15, 20, 25 \land r_1 = 2 \land r_2 = 1 \land h \in (\mathbb{H} \setminus \{0\}), \)
Rule no. \( 156 - 158, 195 - 197, 234 - 236, 273 - 275, 312 - 314, \)
Conditions \( z = 5, 10, 15, 20, 25 \land r_1 = 2 \land r_2 = 2 \land h \in \{0.2, 0.25, 0.3\}, \)

Table 1: Summary of the different appointment scheduling rules
Patients’/ customers’ unpunctuality can take different forms. They can arrive on time but may arrive too early or too late or may The probabilities of a customer showing up too early or too late \( (P(\text{early}), P(\text{late})) \) are varied between zero and 0.1 \( (P(\text{early}), P(\text{late}) \in \{0, 1\}) \). Moreover, we also take into account the amount of time a customer arrives early or late \( (SCV(\text{early}), SCV(\text{late}) \in \{0, 5; 1\}) \).

- A last environmental variable is the probability of no-shows. Indeed patients/customers may fail to show up for their appointment. The probability of no-shows is varied between three levels \( P(\text{No show}) \in \{0; 0.1; 0.2\} \).

- The time needed to perform the service is uncertain as such service time can vary from customer to customer \( (SCV(\text{service})) \)

By combining the different values for the environmental parameters, we obtain 243 environmental settings in which we evaluate the performance of the ASR.

3 Results

After obtaining the relevant performance measures by use of the analytical Markov Chain model, for all 314 ASR under the different environmental settings, we perform a data envelopment analysis to get a performance ranking of all the ASR. This allows us to select those ASR that work best under a wide range of settings and performance criteria. In this section, we present our key results.

We will first discuss the performance of the ASR across all environments. Next, we will select nine ASR to discuss in more detail with respect to the influence of the environmental parameters and the incorporation of subjective information in the evaluation process. Although our method is suitable to select the best ASR for any given setting and preference structure, we will focus on general insights with respect to the classes of ASR: individual ASR, block ASR and early-lateness (E-L) ASR (see section 1).

Thanks to our multi-performance and multi-environment approach, we are able to identify those ASR that have a strong and robust performance across all environments and performance measures. The performance of a
ASR_i is measured by use of a composite indicator (CI).

\[ CI_i = \frac{1}{\sum_{k=1}^{3} w_{ik} x_{ik}} \]  

(4)

The composite indicator (4) of an ASR_i is calculated as the inverse of a weighted sum of the performance measures, viz. the patient waiting time, the server overtime and idle time. We take the inverse because in our setting we want to minimize the performance measures. The weights w_{ik} are determined by means of the data envelopment analysis. A higher CI_i value identifies a stronger performing ASR. Next to the CI scores which we will average across environmental settings, we also measure the sensitivity of the CI score to different choices in the weight sets (w_{ik}). For this purpose we use the maverick index, which is a well-known measure in the data envelopment literature in order to gauge the sensitivity of the performance to the selection of a particular set of weights (w_{ik}) [5].

Table 2 provides an overview of the best performing ASR based on the average performance across all environments: we report the average CI values across the different environmental settings. In the table we also indicate to which type of ASR, individual, block or E-L ASR, the specific ASR belongs. The type of ASR is indicated by a value ”1” in the suitable column. It is striking that the six best performing ASR are all individual ASR with the common characteristics of zero delay for the first arrival, an initial number of patients of 3 or more and with a very small to zero adjustment for service time variance. This observation is encouraging for practitioners as these rules are simple to implement. At the downside, the maverick index (the last column of table 2) indicates that the efficiency scores of these simple ASR are quite sensitive to the weight selection. The lower sensitivity to the weight selection can be a reason to opt for an E-L rule, which are firmly established in the top 15 as from rank seven.

Block ASR are clearly the less attractive type of ASR. The best performing block ASR (ASR 92, 93, 94) have a block size of two with a small or zero adaption for service time variance. The dominance within the block ASR of rules with a block size of two, i.e. customers arrive in groups of two at a time, corresponds with the conclusion of earlier research [2, 8]. Lastly, we note that the simple Bailey-Welch rule with two initial arrivals followed by one-by-one arrivals (ASR 8) performs rather well both in terms of the efficiency score and weight sensitivity, respectively 99.81% and 0.0959.
From these results, we can conclude that individual ASR, with conscientiously chosen settings, have the best performance across environments taking into account server overtime, idle time as well as customer waiting time. E-L ASR, however, seem to be more robust with respect to the value judgement, i.e. importance weighting, between the three performance measures: overtime, idle and waiting time. The performance of block ASR is dismal and they should be avoided because better alternatives clearly exist.

<table>
<thead>
<tr>
<th>Rank</th>
<th>ASR</th>
<th>CI (%)</th>
<th>Individual ASR?</th>
<th>Block ASR?</th>
<th>E-L ASR?</th>
<th>Weight Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>99.998</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.195624</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>99.9976</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.313921</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>99.9964</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.389075</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>99.9954</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.284273</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
<td>99.9941</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.412453</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>99.9838</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.163625</td>
</tr>
<tr>
<td>7</td>
<td>273</td>
<td>99.9724</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.066703</td>
</tr>
<tr>
<td>8</td>
<td>162</td>
<td>99.965</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.171877</td>
</tr>
<tr>
<td>9</td>
<td>256</td>
<td>99.9453</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.070306</td>
</tr>
<tr>
<td>10</td>
<td>298</td>
<td>99.9184</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.064002</td>
</tr>
<tr>
<td>11</td>
<td>31</td>
<td>99.906</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.364031</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>99.8875</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.340854</td>
</tr>
<tr>
<td>13</td>
<td>321</td>
<td>99.8871</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.063741</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>99.8838</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.076893</td>
</tr>
<tr>
<td>15</td>
<td>209</td>
<td>99.872</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.095337</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
<td>99.8136</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.09592</td>
</tr>
</tbody>
</table>

Table 2: Ranking of ASR based on average efficiency across environments

Based on the results discussed above and prior research [8, 9], we select nine ASR which are interesting to discuss in more detail. Table 3 provides an overview of the selected ASR and their characteristics. First, we will illustrate the effect of the different environmental parameters. Second, the impact of subjective information concerning the importance of the different performance measures on the selection of an ASR is demonstrated. None of the selected individual ASR prescribes a delay for the first arriving patient, this choice was made given the inferior performance of ASR with a delay of the first arriving patient. The selected individual ASR differ in the number of patients arriving at the start of the session and ASR 7 is the only individual ASR with an adjustment of the arriving times to account for the variance in service times. The two selected block ASR both have a block size of two and no or very small adjustments of the arriving times to compensate for the service times variance. Lastly, two E-L rules are selected which both
dictate that after the first ten patients of the session, the following patients should be scheduled with a postponement relative to the as-soon-as-possible schedule. Contrary to the other selected ASR, the adjustment for the service time variance is substantial for the selected E-L rules.

<table>
<thead>
<tr>
<th>ASR</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Individual ASR, one patient at session start, max. (h=0.3) adjustment for service time variance</td>
</tr>
<tr>
<td>8</td>
<td>Individual ASR, two patients at session start, no adjustment for service time variance. (= Simple B-W rule)</td>
</tr>
<tr>
<td>15</td>
<td>Individual ASR, three patients at session start, no adjustment for service time variance</td>
</tr>
<tr>
<td>22</td>
<td>Individual ASR, four patients at session start, no adjustment for service time variance</td>
</tr>
<tr>
<td>29</td>
<td>Individual ASR, five patients at session start, no adjustment for service time variance</td>
</tr>
<tr>
<td>92</td>
<td>Block ASR, block size of two patients, no adjustment for service time variance</td>
</tr>
<tr>
<td>93</td>
<td>Block ASR, block size of two patients, small (h=0.05) adjustment for service time variance</td>
</tr>
<tr>
<td>161</td>
<td>E-L ASR, postpone arrivals after first 10 patients (z=10), strong (h=0.25) adjustment for service time var.</td>
</tr>
<tr>
<td>162</td>
<td>E-L ASR, postpone arrivals after first 10 patients (z=10), max. (h=0.3) adjustment for service time var.</td>
</tr>
</tbody>
</table>

Table 3: Overview of the selected ASR

For the nine selected ASR, we now look at the impact of changes in the environmental parameters: the number of patients in a service session (N), SCV of the service times (SCV(service)), SCV of both the early and late arrivals (SCV(early), SCV(late)), probability of early and late arrivals (P(early), P(late)) and the probability of a no-show (P(noshow)). We will change these parameters one by one compared to a base case scenario which has minimal uncertainty: N = 10, SCV(service) = 0.2, SCV(early) = SCV(late) = 0 and P(early) = P(late) = P(noshow) = 0. When the number of patients within a service session increases the patient waiting time increases (figure 1). At the same time, also the idle time and overtime increase, albeit not always proportionally. The effect of an increasing number of patients is very similar for the idle time and the overtime such that we only add the graph 2 showing the effect on the idle time. As could be expected, the E-L rules, ASR 161 and 162, perform strongly in terms of waiting time. Both the individual and block ASR, incur higher patient waiting times than the E-L rules. The exception is ASR 7 which is an individual ASR, but performs very strongly in terms of waiting time. However, the server idle and overtime explodes under this ASR. In general, the evaluation based on server oriented measures is fully opposite compared to the evaluation based on the patient waiting time: the individual ASR and the block ASR have
low idle time (and overtime) for all levels of $N$. Moreover, we note that the impact of the number of patients can be significant, in some cases even higher than proportional to the increase of $N$ which seems to favor smaller service sessions.

Figure 1: Impact of $N$ on patient waiting time

![Figure 1](image1.png)

Figure 2: Impact of $N$ on server idle time

Whereas our base case environment is characterized by a minimal uncertainty in service times ($SCV(service) = 0,2$) and no uncertainty in the patients' arrival process, we now inflict more uncertainty. An increase in the $SCV(service)$ has an unfavorable impact on all performance measures (figures 3 and 4). Once again the impact on idle and waiting time is very similar. Based on patient waiting time, the E-L rules (and the individual ASR 7) should be preferred. However, we observe that exactly these ASR
are most influenced by an increase of the service time variability. If the server oriented measures (idle and overtime) are key, the most suitable ASR are individual ASR, viz. ASR 8, 15, 22 and 29.

![Figure 3: Impact of SCV(service) on patient waiting time](image)

![Figure 4: Impact of SCV(service) on server idle time](image)

The probability of patient arriving too early or too late and the SCV of the deviation time only have minor influences on the performance of the ASR. This could explain why these environmental settings are not studied in the existing literature.

As discussed by many authors the probability that patients do not show up at their appointment is a major source of variability in real life healthcare systems. Patients not showing up cause the total waiting time for the patients to decrease (figure 7). Moreover, no-shows imply additional server idle
time (figure 5), while the server overtime decreases (figure 6). We bring the reader’s attention to the fact that while the server idle time increase more or less proportional to the no-show probability (figure 5), the impact on the amount of overtime decreases as the probability increases (figure 6). This shows how detrimental high no-show rates can be. Moreover, the impact of patients not showing up is substantial whatever ASR is used, indicating the importance of avoiding no-shows. This can explain the recently introduced "open-access" systems where no long term appointments are made anymore but where patients/customers are invited to call in the morning in order to arrange an appointment that same day [11]. Hence, an open-access systems minimizes the problem of no-shows, by having only recently booked appointments but it incurs higher variability in the workload.

Figure 5: Impact of $P(\text{noshow})$ on server idle time

Figure 6: Impact of $P(\text{noshow})$ on server overtime
Figure 7: Impact of $P(noshow)$ on patient waiting time

Given the results above it is clear that the impact of environmental variable depends on the ASR. In general, we can state that the probability of no-shows has the highest impact, the number of patients in a session and the service time variance have a significant impact, while the impact of patients’ punctuality ($P(late), P(early), SCV(late), SCV(early)$) is limited. Moreover, from the figures above it is clear that E-L rules (and ASR 7) are to be preferred w.r.t. patient waiting time, while individual and block ASR (8, 15, 22, 29, 161 and 162) outperform in server idle and overtime. Strikingly, is the relative balanced performance of the simple B-W rule, i.e. ASR 8.

At this point, we have established which ASR work well across environments and specifically in the base case environment. Moreover, we have determined the impact of the different environmental parameters. However, the hospital management team may want to stress one (or more) of the performance measures. Our approach allows for the incorporation of such subjective information about the importance of the different performance measures. We will now discuss how the suitability of the ASR changes when either the patient waiting time or the server idle time is pivotal for scheduling purposes.

If patient waiting time is pivotal, we can impose on the DEA that the weight associated to the average weighting time should be higher than the weight allocated to either the idle time and overtime measures. From these results (not reported in this paper), it is clear that the E-L ASR dominate the performance ranking when the patient waiting time is the most important criteria. The top 4 ranked ASR, when waiting time is highly important, are all E-L ASR. Interestingly, these are exactly the same ASR that are the
best performing E-L ASR without any weight restrictions (ranked 7th to 10th in table 2). In sharp contrast, the top 6 individual ASR in table 2 tumble down in the ranking. The strong performance of the E-L rules across all environments when patient waiting time is considered most important reinforces our earlier observation that these rules perform strongly w.r.t. waiting time. Similar as without weight constraints, the block ASR perform badly: ASR 94 and ASR 93 are ranked 107 and 108, respectively. Within the group of block ASR, the ASR with block size of 2 again have the strongest performance. Lastly, we note that the simple B-W ASR performs less well when patient waiting time is important, decreasing from rank 23 to 37 out of 314.

Albeit the benefits of waiting time minimization, the expensive equipment (and people) used in healthcare systems together with the high amount of work may favor the use of ASR that minimize idle time. Consequently, we incorporated the additional constraint that the weight allocated to the idle time should be the highest, i.e. idle time is the most important performance criteria. With this constraint on the weight selection, the individual ASR return to the top rankings, pushing back the E-L ASR. All of the top six ranked individual ASR from table 2 re-appear in the top 15 and make up the four best ranked ASR when idle time is the most important evaluation criteria. Even more consistency, can be found in the best performing E-L ASR: the ASR 273, 162, 256, 298, 321 and 209 are part of the top 15 ranked ASR across all environments in the case without weight restrictions, as well as with waiting time or with idle time as the most important performance measure. This let us to belief that these ASR are especially suitable for a wide range of environments and operating conditions. In line with the previous observations, the block ASR perform badly: the best ranked block ASR (ASR 93) is ranked 72th.

4 Conclusion

In our research we have developed an analytical procedure based on a Markov chain and data envelopment analysis in order to evaluate different appointment scheduling rules (ASR) and identify the best performing ASR. An ASR prescribes when patients/customers have to come in for service. ASR are widely used across service industries including hospitals (health care), car maintenance (after sales) and work counseling (social services).
First, we used a Markov Chain model to efficiently calculate the performance measures for a large group of ASR. Specifically, we evaluate 314 ASR based on three performance measures, viz. patient waiting time, server idle time and overtime, in a wide range of environmental settings. The environmental settings differ in the number of patient/customer per session, service time variability, patient unpunctuality and the probability of patient no-shows. The Markov model builds on matrix analytical methods and adopts an efficient algorithm (in terms of computational and memory requirements). The model offers us the possibility to efficiently calculate the performance measures for all the ASR in each of the different environmental settings.

Second, we needed to evaluate the large set of ASR across the different environments based on several performance measures. For this challenging task, we applied a non-parametric multi-performance evaluation technique called data envelopment analysis (DEA). The DEA results in a ranking of the ASR based on their performance taking into account all three performance measures.

The procedure we developed enables the comparison of individual ASR in specific environmental settings such that e.g. a general practitioner can select the best ASR for his specific situation. Nevertheless, we focus in this work on the suitability of the different types of ASR (individual, block and early-lateness ASR) across all settings. In general, we can conclude that individual ASR have a superior performance across all environments. However, E-L rules hardly underperform and often have a performance that is less depending on the value judgement about the importance of the different performance measures. Furthermore, E-L rules outperform individual ASR if patient/customer waiting time is the most important performance criteria. We were able to identify a group of six E-L rules that appear in the top 15 of best performing ASR for all three tested value judgment situations: when there is no value judgement about the importance of the different performance measures, when patient waiting time is the most important criteria and when server idle time is the main criteria. The performance of block ASR is dismal and the use of such rules should be avoided as better alternatives exist. The probability of no-shows among the patients/customers has the most severe impact on the service performance. Also the service time variability and the number of patients during one sessions have large influences on the service performance. Patients’ unpunctuality, however, has only minor influence on the performance measures.
References


