The Access spectrum
Traditional, Carve-out, and Open as points of a continuous space
EXECUTIVE SUMMARY

The work described in this paper concerns a project undertaken in the past two years. The main purpose of the project was to develop a Decision Support System for operational design in outpatient clinics.

The study focused on the recently created Family Health Units, the first line of health care providers of the Portuguese National Health Service. These clinics have a team of up to ten general practice physicians, up to ten nurses, and up to seven administrative staff, providing primary care to a relatively stable population size, with panel sizes per doctor that may range between 1500 and 2000 users. Among the problems faced by these outpatient clinics is the definition of the adequate supply profile and the adequate access policies, ensuring good operational efficiency and performance.

We proposed to develop a discrete event simulator that would be amenable to be used by the medical staff, easily configurable to the specifics of each clinic, with the purpose of testing different supply profiles and access policies in order to quantitatively assess advantages and drawbacks of the tested alternatives.

The current simulator version is very close to the intended objectives and already allows testing a significant variety of alternative policies and policy parameters. In this paper, we use the present simulator version to investigate alternative access policies, given a fixed supply profile. The simulator emulates one of the FHU's we studied, using their own supply profiles and demand patterns. The data presented here quantitatively illustrates the effects of a small set of parameters and allows to relate them with macroscopic performance measures, as service level, backlog volume, no show rates, and substitution rates.

We believe that the set of numerical results presented, together with a formal framework and results, contributes to the debate concerning Advanced Access practices, shedding a numeric light on a debate that is seldom formally rigorous.

The Access spectrum
Traditional, Carve-out, and Open as points of a continuous space

February, 2012

Abstract

The Portuguese National Health Service is undergoing a significant structural change in terms of the way it is organizing its primary care. The old Local Medical Centers are being replaced by the Family Health Units, which are outpatient clinics composed of general practice physicians, nurses, and administrative staff, providing health care to a relatively stable population size. These new clinics have a higher degree of autonomy and pose some interesting management problems in terms of the definition of their supply profiles and access policies.

This has created an opportunity to conduct a comprehensive study on how these clinics should organize their operations. We developed a discrete event simulator that enables the teams to try different supply profiles and access policies, before committing their organizational structure. With such a tool we are able to provide a quantitative support to the debate concerning the management of operations in outpatient clinics.

In this paper we focus exclusively on access policies and argue that the adequate choice should take into account the specific dynamics of each clinic and how their incentives are defined. We claim that the alternative access policies should be seen as a continuum and not as competing and mutually exclusive alternatives.

Keywords: Health care management, Outpatient clinics, Access policies.

1 Introduction

Consider a primary care outpatient clinic dealing with the agenda management and patient access problems. On one hand a Supply Profile has to be defined and, once that is agreed upon, a set of procedures and scheduling practices is to be put in place in order to determine and condition how patients get their primary care needs satisfied. The definition of the set of procedures and scheduling practices, termed as the Access Policy, can be considered as a problem apart from the definition of the supply profile, although different supply profiles may condition some features of the access policy.

In this paper we propose to solely focus on access policies, assuming the supply profile as a given, with the purpose of contributing to the debate on the adequacy of different access policies. Although the supply profile conditions the performance of such systems, one can still evaluate different access policies in terms of their relative performance under a fixed supply profile. We claim that different access policies should not be seen as competing alternatives, but rather as different points on a continuous space of policies and that the adequacy of each of them is conditioned by context.

The literature on access policies traditionally considers the existence of three major models: Traditional Access, First Generation Open Access, and Second Generation Open Access, [7, 6]. Systems run under Traditional Access split demand into two major categories: urgent and non urgent. These systems attempt to schedule the urgent requests for the day they occur and push forward in time the non urgent requests. The First Generation Open Access systems reserve capacity for urgent requests ahead of time, what is known as Carve-out. In Second Generation Open Access there is no distinction between urgent and non urgent requests. These systems attempt to schedule all requests for the day they occur, unless the patient has no interest in being seen that day or the nature of the request imposes a later scheduling date.

There has been some debate over the two Open Access models, also referred as Advanced Access models, for outpatient clinics, where the second model is often presented as the alternative of choice. We feel that the debate has been lacking a formal framework and that it is, in many circumstances, based more on episodes or soft arguments, and some times serving specific agendas. We propose to offer a framework under which to discuss the alternatives, quantify their relative performances, and show

that they may be seen as extreme points of the same policy. Moreover, we developed a discrete event simulator for outpatient clinics where different features can be tested and competitive advantages and adequacy can be quantified.

This study focuses on Family Health Units that have been recently put in place in Portugal as one key component of the National Health Service, replacing the old Local Medical Centers. These units are primary care outpatient clinics with up to ten general practice physicians, each being the family doctor of families living in the vicinity of the clinic, and with panel sizes that may range between 1500 and 2000 patients. These units are bounded in size, being it the case that the maximum sized unit cannot have more than ten family doctors and family nurses, and seven administrative staff. One of the main differences relative to the old medical centers is the fact that these units have a higher degree of autonomy, a system of incentives, and are subject to regular evaluation, based on clinical and operational results, that drives and conditions funding and the achievable autonomy level.

The paper is organized as follows. Next section presents the motivation for this work and the formal framework being proposed. In light of that framework, Section 3 critically reviews the debate on the alternative access policies. Section 4 describes the discrete event simulator and presents a series of numerical results generated by it. The paper closes with the conclusions on Section 6.

2 The context and framework

Portugal has an almost forty years old National Health Service that provides health care to all citizens and is composed of several components from the Local Medical Centers up the the main Central Hospitals. The Local Medical Centers are the first line of prevention and health care. However, their past performance has been far from satisfactory as medical and administrative staff suffer from the basic drawback of public servants, which is the lack of a sense of ownership and lack of incentives to perform better.

To address this drawback, the Portuguese government has created the Family Health Units, FHU, which are expected to replace all Local Medical Centers in the next years. A FHU is created by proposal of a team of doctors, nurses, and administrative staff. Each FHU draws a contract with the Portuguese Health Ministry, establishing short and long term goals concerning primary care performance and results, as well as a basic set of compulsory services that have to be provided. As these goals are achieved the FHU will have access to higher funding and higher degree of autonomy on how to manage the funds and the facilities. Although all staff is composed of public servants, receiving their salary from the government, it is as if each FHU is a private clinic with a government contract to provide health care to a community.

These FHU's are also evaluated in terms of their operational performance, which raises the issue of how to dimension capacity and how to manage its utilization. This has created an opportunity to investigate alternative capacity profiles and access policies.

2.1 Services provided

Each person enrolled in a given FHU will have a doctor and a nurse assigned, which form a team that provides primary care services. Any two persons living together will have the same family doctor and nurse. Persons enrolled in a FHU are designated as users. Some FHU's may have different working hours and days, but the majority is open from 8:00 AM to 8:00 PM on weekdays and closed during weekends. We will be focusing on the services provided by doctors only, because the services exclusively provided by nurses are usually performed without a need of a prior scheduled date or appointment.

Any user can contact his/her doctor through phone or through a previously scheduled appointment, which can also be a house call. Appointments can be classified in terms of their type, as Women's Health or Children's Health, for instance. Each doctor provides a weekly plan specifying his/her availability throughout the week and what type of appointments are offered. A user in need of an urgent appointment is offered the possibility of being seen by another doctor of the FHU if the family doctor is not available. Doctor interchangeability can also apply to other types of appointments. Each FHU defines the types and/or situations under which a given request may have access to interchangeability. For instance, during a doctor's holidays some types, that otherwise would not, may have doctor interchangeability.

The supply profile of a FHU is the composition of the doctors' weekly plans and different types being offered. Each demand request may be initiated by a user, external demand, or by a doctor or nurse, internal demand, and asks for a specific type of appointment that needs to be fit into the available slots. The access policy is the set of mechanisms and rules that condition how a given request will be satisfied.

In Figure 1 we present the demand matrix. Demand is characterized by three vectors: Agent, Type, and Date. The Agent is who requests some service, which may be an internal request or an external one. Type specifies what contact is requested, being it direct, indirect, urgent appointment, house call, women's health, etc. The field Date characterizes the requested date, that can range from zero, meaning today, to some positive value like 180 days. So, any given request will occupy one cell of that matrix. It is assumed that a request is always made to the family doctor. Before moving on, we need to discuss the way dates may be produced. In many contexts a request for contact does not produce a crisp date, like asking for a direct contact in exactly ten days. In some cases the request may be for the earliest possible date, for next month, or within the next three weeks, as examples. Situations like this are well handled by humans but more difficult to model in a simulator. Therefore, we are assuming that every request for a contact will always carry a crisply defined date.

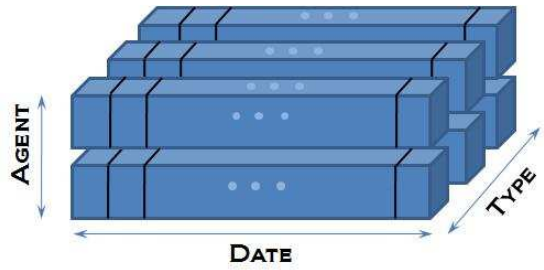

Figure 1: Demand matrix.

We should stress also that some of the cells of the demand matrix are empty. As an example, it does not make any sense that a request for an urgent direct contact is made to occur in exactly 21 days. One should expect all urgent requests to have the desired date equal to zero.

Each FHU will have a supply matrix that does not necessarily match the demand matrix cell by cell. For instance, the different types in the supply matrix may have lower cardinality than the different types in the demand matrix. The access policy is in charge of the mapping between the two matrices, and must ensure that each cell of the demand matrix will have a target set of cells in the supply matrix. For instance, a request for an appointment today may be satisfied today or in the next days. The complete supply profile includes also the definition of the total number of hours each doctor is available for each type of appointment and the nominal duration of every appointment type.

The definition of any specific supply matrix has implications on, and/or is conditioned by, the nature of the implementable access policies. Given that we want to focus on access policies, we will assume the supply matrix as a given and a constraint in terms of total hours being offered, but will introduce some variations to account for specific features of the access policies. As it will be seen in the following sections, the developed simulator may also be used to provide information concerning the relative quality of more radically changed supply matrices.

2.2 Formal framework

Every demand episode is a request for contact with the family doctor. We term it as **indirect contact** when the request is, or asks for, a phone call with the family doctor. When the request is for an appointment we term it as a **direct contact**. A request for direct contact requires the scheduling of such contact in the doctor's agenda. Every contact that takes place may originate the need of a future contact for a follow-up.

Some types of contacts generate a stream of future contacts, defined in the health legislation, like the follow-up plan during pregnancy. A contact is said with no priors when its request is not associated with a previous contact. A contact with prior is the one requested by a medical decision upon the conclusion of some other contact. Assuming there are no no shows, Figure 2 illustrates the total demand experienced by a FHU. Assuming d to represent the total no prior demand, the total volume of requests is given by $d + d_{\text{prior}}$, where d_{prior} represents the demand incurred due to medical decisions. The diagram of Figure 2 is overly simplified, but serves to illustrate one other fundamental issue concerning outpatient clinics. The existence of a positive feedback may induce instability when the medical decision for a follow-up contact has a high probability of occurring, in general, or a high probability of occurring for a specific type of contact. Instability should be understood here as a total volume of requests in hours being steadily above the total number of hours offered by the clinic, what would induce an increase in waiting time for appointments.

Figure 2: Total demand experienced by a FHU.

Let $d(n,k,s,t)$ represent the total demand for contacts on day n to occur on day $n+k$, with $k = 0,1,\ldots,K$, for doctor s, with $s = 1,2,\ldots,S$, of type t, with $t = 1,2,\ldots,T$. With no loss of generality, assume that $t = T$ represents the indirect contact type. S is the number of family doctors in the FHU, T represents the total number of possible types requested, and K represents some upper bound on how far ahead each request is made. We can say that

$$d(n,k,s,t) = d^0(n,k,s,t) + d^1(n,k,s,t), \tag{1}$$

where $d^0(.)$ represents the demand with no prior – external, and $d^1(.)$ represents demand due to a medical decision – internal, produced during some contact occurring on day n.

Let $y(m,n,s,j)$ be the number of direct contacts of type j scheduled for doctor s that at the end of day m are scheduled exactly for day n, with $n \geq m$, with $j = 1,2,\ldots,J$, and J the total number of types being offered by the FHU. In general, $T \geq J$. We define the mapping matrix R, $(T \times J)$, such that $\sum_{j=1}^{J} r_{t,j} = 1$ for every $t < T$, and $r_{t,j} = 1$ if requested type t is satisfied by a slot of the supply type j. $r_{t,j}$ is zero otherwise. Also, $r_{T,j} = 1, \forall j$.

The number of direct contacts of type j scheduled for doctor s at day n by the end of the day, is $x(n,s,j)$ and is such that $x(n,s,j) \geq y(n-1,n,s,j)$, with $j = 1,2,\ldots J$. The difference is due to the requests $d(n,0,u,t)$, $u = 1,2,\ldots,S$, that are successfully scheduled on the agenda of doctor s satisfying the requested date, with t being such that $R(t,j) = 1$.

Due to the fact that some users may fail to show to a previously scheduled direct contact, the total number of direct contacts of type j completed by doctor s during day n is $x^r(n,s,j) \leq x(n,s,j)$.

Indirect contacts are usually opportunistically satisfied by the doctors. That is, the pending requests for a phone call are served whenever the doctor has a free slot, either because there was no direct contact previously scheduled, because the user failed to show to a previously scheduled direct contact, or because the doctor has some free time for other reasons. Therefore, we may define as $x_T(n,s) \geq 0$ the total amount of indirect contacts that are pending for service at the end of day n. Thus, $x_T(n,s) = x_T(n-1,s) + d(n,0,s,T) - x_T^r(n,s)$, with $x_T^r(n,s)$ denoting the total number of indirect contacts satisfied during day n. The total number of contacts doctor s completes during day n is given by

$$z^r(s,n) = \sum_{j=1}^{J} x^r(n,s,j) + x_T^r(n,s). \tag{2}$$

Each one of the concluded contacts may generate a request for a new contact, as described earlier. Some of the medical decisions for a new contact may be for a different doctor. Such is typically the case when one direct contact is not conducted by the family doctor and the need for a follow-up is sent back to the family doctor. So, the medical decision will include a type, a date, and a doctor for the follow-up contact.

Let us define the concept of a single demand episode as one request $\delta_m(n,k,s,t)$, with $m = 1,2,\ldots$ identifying the arrival order. If the demand episode is a request for direct contact for which there is no interchangeability, it will generate a booking in the agenda that we may represent as $\xi_q(n+k+w,s,j)$, with $q = 1,2,\ldots$ representing some booking order and $w \in \mathbb{N}_0$ representing the excess waiting time. If the request is satisfied on the exact date for which it was requested, $w = 0$. Otherwise, there will be a waiting time of $k+w$ days. The first k days of the waiting is what the user was willing to wait or is what the doctor considered the proper waiting time for the follow-up to occur. We define as Type-1 service level the percent of direct contact requests that get $w = 0$.

Once a demand episode for direct contact is scheduled on some doctor's agenda, there will be some waiting time until the contact should take place. During such waiting time, the user may decide to call in to give up the booking, may call in to reschedule it, or will not show on the scheduled date with no warning. This later case implies a loss of capacity, while the first two cases may or may not imply a loss of capacity. We should expect that, for any of these three behaviors, their occurring probability to increase as $k+w$ increases. For instance,

$$Pr\{\text{No show}|k+w\} \leq Pr\{\text{No show}|k+w+1\}. \tag{3}$$

The first result we will establish concerns the stationary demand. We will simplify our model, aggregating all types and all doctors.

Theorem 1 Assume each external request occurring on day n will ask for date $n+k$ with probability $p_k^0 \geq 0$, with $k = 0, 1, \ldots, K^0$ and $\sum_{k=0}^{K^0} p_k^0 = 1$. Assume each internal request occurring on day n will ask for date $n+k$ with probability $p_k^1 \geq 0$, with $k = 0, 1, \ldots, K^1$ and $\sum_{k=0}^{K^0} p_k^0 = 1$. Assume that each contact will generate a request for a new contact with probability $0 \leq q \leq 1$, which is constant over time. The average long term number of total requests scheduled on each day is given by

$$\bar{x} = \frac{\bar{d^0}}{1 - q\bar{r}(K^0, K^1)}, \tag{4}$$

with $\bar{d^0}$ the average external demand, assumed to be stationary, and $\bar{r}(K^0, K^1)$ the average realization rate for the scheduled contacts.

Proof: With no loss of generality, assume that the capacity is such that all requests for contact are scheduled in the desired date and that all indirect contacts are also scheduled on the doctor's agenda, but their duration is negligible. Let $d(n) = d^0(n) + d^1(n)$ be the total demand that occurs on day n, $x(n)$ be the number of contacts scheduled to take place on day n, and $x^r(n)$ the number of contacts that actually take place on day n. Assume that

$$x^r(n) = r(n)x(n), \tag{5}$$

with $0 \leq r(n) \leq 1$ and $r(n)$ possibly a function of K^0 and K^1. Because each contact may generate a new request, the average number of contacts requested to be scheduled on day n, given the contacts that took place in the past, is

$$\begin{aligned}
\mathrm{E}[x(n)|x^r] &= \mathrm{E}[x(n)|x^r(n), x^r(n-1), \ldots x^r(n-K^1)] \\
&= \sum_{k=n-K^1}^{n} x^r(k) q p_{n-k}^1 + \sum_{k=n-K^0}^{n} d^0(k) p_{n-k}^0 \\
&= \sum_{k=n-K^1}^{n} x(k) r(k) q p_{n-k}^1 + \sum_{k=n-K^0}^{n} d^0(k) p_{n-k}^0
\end{aligned} \tag{6}$$

Therefore, the average number of contacts requested to be scheduled on day n are

$$\begin{aligned}
\mathrm{E}[x(n)] &= \mathrm{E}\left[\mathrm{E}[x(n)|x^r]\right] \\
&= \sum_{k=n-K^1}^{n} \mathrm{E}[x(k)] \bar{r}(K^0, K^1) q p_{n-k}^1 + \bar{d^0} \sum_{k=n-K^0}^{n} p_{n-k}^0
\end{aligned} \tag{7}$$

Taking the limit as $n \to \infty$ we obtain

$$\begin{aligned}
\bar{x} &= \lim_{n\to\infty} \mathrm{E}[x(n)] \\
&= \bar{x}\bar{r}(K^0, K^1) q \sum_{k=n-K^1}^{n} p_{n-k}^1 + \bar{d^0} \\
&= \bar{x}\bar{r}(K^0, K^1) q + \bar{d^0},
\end{aligned} \tag{8}$$

which yields the result.

Now assume that some contacts are scheduled after the requested date. The effect of that will be to change the long term individual values of p_k^0, p_k^1, K^0, and K^1. We can replace every request for contact with a new request for a different date, such that $w = 0$. As long as the upper bounds $M^0 = K^0 + \max\{w\}$ and $M^1 = K^1 + \max\{w\}$ remain finite, p_k^0 and p_k^1 will remain as probability distributions. If the long term value $\bar{x}$ is below the average capacity, $\max\{w\} < \infty$.

Q.E.D.

Equation 4 provides an easy way to determine if the clinic is stable. Assuming $\bar{r}(K^0, K^1) = 1$, i.e., all scheduled contacts will take place, we obtain an upper bound on the average number of contacts requested everyday as a function of the external demand and of the medical decision parameter, q. One needs to produce such a relation for each individual appointment type, when different appointment types require different nominal durations. For that, there is a need to model the medical decision with more detail.

Let there be a matrix $P = [Q, R]$, where Q is a $(T \times T)$ matrix and R is a $(T \times 1)$ vector. P is such that $\sum_{j=1}^{T} q_{t,j} + r_t = 1$, for all $t = 1, 2, \ldots, T$, with $0 \leq q_{t,j} \leq 1$ and $0 \leq r_t \leq 1$. We associate the following meaning to matrix P. For every line, t, r_t is the probability that a contact of type t will not generate a follow-up contact, whereas $q_{t,j}$ is the probability that contact type t generates a request for a contact of type j. From matrix P we may generate a new matrix Z, $(T+1) \times (T+1)$, as follows.

$$Z = \begin{bmatrix} Q & R \\ 0 & 1 \end{bmatrix}. \tag{9}$$

Matrix Z represents a finite state Markov chain with T transient states and one absorbing state. This type of chains is also known as terminating Markov chains. The absorbing state represents the non generation of a new contact. Matrix Q represents the transient states and how they jump from one another, while matrix R represents the probability of a transient state jump to the absorbing state. Assuming every doctor will use the same matrix Z, each contact will go through a series of jumps from type to type until it dies away.

Given Theorem 1, and assuming all scheduled contacts will take place, the stationary value of X, $(1 \times T)$, the average number of contacts that will be scheduled any day for each type, is given by

$$\begin{aligned} X &= D^0(I + Q + Q^2 + \ldots + Q^k + \ldots) \\ &= D^0(I - Q)^{-1} \end{aligned} \tag{10}$$

where I is the identity matrix with appropriate dimensions and D^0, $(1 \times T)$, is the expected daily external demand vector. Given the structure of Q, the inverse of $(I - Q)$ is known to exist, and matrix $(I - Q)^{-1}$ is called the fundamental matrix of the Markov chain Z, [8, 1, 2]. A simple inspection of Equation 10 allows to draw a parallel with Equation 4. Therefore, we may use the framework of Equation 10 to obtain the upper bounds on the average daily demand for every type of contact. Moreover, this can be done for each doctor, assuming different values for the expected external demand, D_s^0 and medical decision matrix, Q_s.

In any given day, n, there will be direct contacts scheduled in the agenda for the following days. We define backlog as

$$B(n) = \sum_{j=1}^{J} \sum_{s=1}^{S} \sum_{r=n+1}^{\infty} y(n, r, s, j) \tag{11}$$

At any day n, $B(n)$ will be constituted by two main components: good and bad backlog. The good backlog is defined as those bookings that were made in a future date because that was the desire of the agent who requested the booking. The bad backlog is defined as those bookings made in the future due to a lack of capacity in the present. For instance, an urgent direct contact requested for today that ends up in the agenda two days from now is bad backlog, whereas a request for a direct contact with $k = 60$ that gets $w = 0$ is good backlog. We will review these notions later.

Theorem 2 Consider a stationary clinic. If all scheduled contacts take place, i.e., no user ever misses an appointment, the average backlog, $B = \lim_{n \to \infty} \mathrm{E}[B(n)]$, is non decreasing with $\bar{K} = \sum_{k=0}^{K} k p_k$.

Proof: Let $d(n)$ be the total demand on day n. With no loss of generality, assume that the clinic has sufficient capacity to absorb all demand every day. That is, all requests for contacts get $w = 0$. Each request will ask day $n+k$ with probability $p_k \geq 0$, assumed independent of n and such that $\sum_{k=0}^{K} p_k = 1$.

Therefore, defining $y(n, m)$ as the total number of contacts that at the end of day n are scheduled to occur during day m, with $m > n$, it is the case that

$$\begin{aligned} \mathrm{E}[B(n)] &= \mathrm{E}\left[\sum_{k=1}^{K} y(n, n+k)\right] \\ &= \sum_{k=1}^{K}\sum_{i=k}^{K} p_i \mathrm{E}[d(n-i+k)] \end{aligned} \tag{12}$$

The above relation is valid because we assumed the capacity to always be enough for any demand variation. Assuming stationary external demand and stationary medical decisions for follow-up contacts, we take the limit as $n \to \infty$ to obtain

$$\begin{aligned} B &= \lim_{n\to\infty} \mathrm{E}[B(n)] \\ &= \lim_{n\to\infty} \sum_{k=1}^{K}\sum_{i=k}^{K} p_i \mathrm{E}[d(n-i+k)] \\ &= d(\infty) \sum_{k=1}^{K}\sum_{i=k}^{K} p_i \\ &= d(\infty) \sum_{k=1}^{K} k p_k \\ &= d(\infty)\bar{K} \end{aligned} \tag{13}$$

with $d(\infty) = \lim_{n\to\infty} \mathrm{E}[d(n)]$. Note that only as n grows will we get a stationary demand process, because the follow-up requests depend on the concluded contacts. Note that $d(\infty)$ is what we defined as $\bar{x}$ in Theorem 1.

Now, if the demand and supply variability are such that some requests get $w > 0$ this is equivalent to say that there is a different probability distribution for the values of $m = k + w$. That is, there is an alternative probability distribution, $p_m \geq 0$, for which the resulting excess waiting time is zero, $w = 0$. Assuming that the average long term daily demand is below the average daily capacity, it is the case that the maximum value follows the relation $K < K + \max\{w\} = M < \infty$. Replacing K by M in the above equations will suffice to establish the general result.

Q.E.D.

Suppose now that some users will waive a contact by letting the clinic know in advance (1), others will reschedule it (2), and others will fail to show with no previous warning (3). These three behaviors induce different consequences. When a contact is rescheduled, no change is produced on the backlog and on the expected contacts that will take place afterward. When a user gives up a scheduled appointment it reduces the backlog and the potential future follow-up appointments. A no show has no direct influence on the backlog size, only an indirect effect for loss of potential future follow-up appointments.

Therefore the global effect is that of a reduction on $d(\infty)$ and it is as if the demand episodes (1) and (3) never happened. Naturally, for larger values of K, episodes (1) and (3) will have a higher probability of occurring. Therefore, there may be situations where those effects cancel each other. However, from a practical point of view and from our experience, the spread of dates requested always supersedes the effect of an increase in no shows and waived bookings. The model assumed for the probability distributions on no shows and waived bookings as functions of $k + w$ produces very mild increases on these occurrences for larger spreads of the requested dates.

In any case, a direct consequence of Theorem 2 is that reducing the spread of dates and/or increasing the probability of asking for low values of k will reduce the average long term backlog.

This fact alone calls for the definition of an access discipline that enforces a low spread of dates. We define as **postponement** an access discipline that forces contacts to be scheduled closer to the desired date. For instance, suppose a given chronic pathology requires regular direct contacts every six months. One usual procedure is to schedule the next appointment on the same day the current appointment takes place, i.e., with $k = 180$. If a clinic enforces a 15 day postponement discipline, the user will be asked to schedule the next appointment only when there is a 15 days gap until the desired date. So, after the current appointment the user will go home and will wait five and a half months. After that waiting time,

the user will contact the clinic to schedule the appointment requested by the doctor. The postponement discipline will have two major consequences. The first, is a reduction of the average backlog as established by Theorem 2. The second, is a reduction of no shows for scheduled appointments. Naturally, some users may forget to call in at a later date. We argue that users who forget to schedule after the postponement waiting time, would likely forget to show for the appointment, had it been scheduled with $k = 180$. A user that forgets to schedule can be reminded after a couple of days. Without postponement, some clinics have a remind all policy, where an administrative will be calling all booked users a few days prior to their appointments. So, with postponement there is a reduction in the amount of reminder calls and a no show to schedule is far better than a no show to appointment, because it does not waste capacity. Also, it is naturally difficult for any person to know his/her availability on a date six months ahead into the future, which with postponement reduces potential rescheduling calls.

Another feature of a general access policy concerns mechanisms to deal with urgent requests, for users with an acute condition. There are basically two such mechanisms. The first is the doctor substitution. The second is **carve-out**. The FHU's have a doctor substitution practice and different such clinics offer different substitution procedures. Some have a doctor available to do them, others try to fit a request in an alternative doctor. As to the carve-out it consists on reserving a set of slots in every doctor's schedule that can only be booked for urgent requests on the day they occur. Carve-out may induce some capacity loss due to a lack of synchronism between demand and supply, but may also increase the likelihood of a user being seen by the family doctor and on the date requested. Note that having a doctor available for urgent requests that cannot be satisfied by the adequate family doctor is another way of implementing carve-out.

These two carve-out alternatives can be classified as **explicit carve-out**, given the fact that they produce daily slots which cannot be booked on earlier days. We define as **implicit carve-out** when some special days have special treatment. For example, some authors propose that some days after a doctor's return from holidays, the day following a national holiday, or even Mondays may be blocked from booking ahead of time by external demand requests, to free capacity for requests with $k = 0$ and/or to prevent those days to be excessively booked.

3 Review

Considering the discussion of the previous section, we propose to review the debate on alternative access policies. For the sake of simplicity let us assume $T = J = 1$, all appointments scheduled will take place, and that some triage procedure decides which requests are urgent and which are not.

What is considered the traditional access policy appears to be a case where the dates requested by the users are changed by the clinic according to the urgency evaluation. The determination of the urgency level does not change the volume of the external demand. It only changes the spread of requested dates. Therefore, if the clinic has sufficient long term capacity for the external demand combined with the internal demand, given by the medical decision matrix, the average requests for any single day are not affected, due to Theorem 1. According to Theorem 2, the only affected performance measure will be the backlog, that may be higher than if there were no artificial alterations of requested dates.

The first generation open access policy, which is the traditional policy with carve-out, degrades the performance of the non urgent requests because their w tends to be higher, due to a reduction on available capacity. The performance for the urgent requests should be better here, but the lack of synchronism between demand and supply will induce some long term loss of capacity. Loss of capacity is only critical when the systems are subjected to heavy loads, which according to some authors is not usually the case, [9, 3]. The numbers we have seen in the Portuguese FHU's studied point to yearly loads under 70%, which can hardly be considered as heavy load. In any case, the total average volume of appointments scheduled is not influenced by these features, and the only measurable consequence is a possible degradation of the backlog.

The second generation open access policy makes all, or almost all, requested dates equal to zero, which is equivalent to a first generation open access, with no carve-out, and a postponement of zero days. Since urgent and non urgent requests get the same treatment, one should expect a degradation for urgent requests and an improvement for non urgent. But the essential conclusion is that the daily average volume does not change and the backlog drops to negligible values.

If we consider that $T > 1$ and $J > 1$, then there are other aspects to take into account. Taking the second generation open access case, clinics under such policy are run like a queuing network. The higher the value of $J \leq T$ the less flexible is the network and there will be more capacity losses due to lack of synchronism between demand and supply. Ideally, one should want J to be equal to one. The same

holds true to the other access policies. Reducing the supply variety makes the best use of the available capacity.

However, in some circumstances it is not possible to have $J = 1$. In the case of the FHU's, some appointment types require a specifically equipped room, as is the case of a women's health appointment. To have a gynecological chair in every doctor's office is much too expensive, so each FHU only has one room equipped with such a chair. Therefore, instead of having the doctor moving from one room to the next in two consecutive appointments, the doctors tend to offer a series of continuous slots specifically for those appointments. If the demand for such appointments is not enough on a given day, that capacity is lost for all the remaining appointment types, except possibly for pending indirect contacts. The same holds true for house calls, which are booked on specifically set aside slots at the end of each doctor's office hours and on specific days, due to their low volume.

In summary, when there is supply variety special care has to be placed on determining the global volume of hours available for each type. One should note that, irrespective of the access policy in place, a mismatch between demand and supply is always bound to occur, because demand is stochastic in nature and each clinic does not offer the same volume of appointment hours along each day and from one day to the next. We will not spend much time on this particular problem, because we are assuming the supply profile to have been previously defined in our setting.

Based on the above discussion, there is nothing really intrinsically wrong with any of the three main access policies. What appears to be the case from published literature is that the clinics tend violate the access rules they define to solve specific occurrences generating very confusing and variability inducing behaviors. The analysis made in [7] regarding traditional and first generation access uses arguments that can be more associated with violation the the enforced rules, than with some intrinsically wrong features of the rules.

Also, in [4] a very strong criticism of the second generation open access policy is presented based on an experiment conducted in six clinics. However, a closer look to the arguments allows the understanding that most of the problems reported are due to lack of adherence to the principles, as pointed in [5]. Curiously, the arguments used in [5] could also be used in response to the criticisms made to other access policies in [7].

4 Simulator and numeric results

In order to provide quantitative information on the relative quality of alternative access policies we developed a discrete event simulator for general FHU's using iGrafx®. Figure 3 presents the architecture of the simulator. We use the database of any given FHU to determine all information concerning the supply profile, the external demand, the medical decision matrices, the daily and monthly seasonality patterns, etc. That information is fed into the simulator that produces a database where the agenda for the doctors is constructed and used during simulation. At the end of the simulation some output files are produced and the database contains raw data regarding the simulation run that needs to be collected and processed. The FHU's database is outside the developed application.

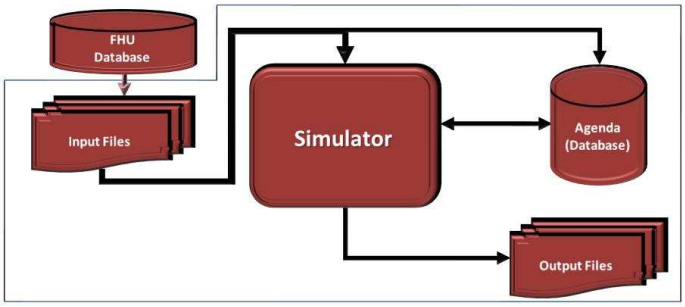

Figure 3: Simulator architecture.

Figure 4 displays a conceptual view of the simulator block, where we detail the standard flow for an external appointment request. Each demand episode generated as an external request goes to a block representing the front office and waits for an administrative. After, the request is scheduled according to the enforced rules, which are coded in. Upon getting a booking, the request will be sent to the waiting block, after which it moves on to the appointment. When the appointment is concluded, the request exits the system. The general diagram for the simulator is much more complex, because we need to

model cases where requests/users leave the waiting box without showing for the scheduled appointment, there are different paths for direct and indirect contact requests, some appointments require the user to go back to the front office, etc.

Figure 4: Standard flow of a demand episode.

The FHU under study is composed of eight family doctors and nurses. The value of T is 8 and the value of J is 4. Supply type 2 is for women's health appointments, type 3 for children's health appointments, type 4 for house calls, and type 1 for all the other appointments. Urgent appointments are scheduled on slots of type 1, except for the urgent house calls. So far, this is the minimum admissible value for T by the government contracts. Some other FHU's have higher values of T.

The substitution policy of the FHU applies only for urgent requests, house calls included. If the family doctor is not available or has no free slot, the system looks for an open slot on any of the doctors with office hours that same day. Only when there is no open slot, the search continues on the next days, always following the same sequence: family doctor first, an alternative doctor second.

The dates produced randomly are not subject to any administrative alteration. This implies, for instance, that the random generator may produce a date for a period the family doctor is out of the FHU, as is the case of the holidays. The scheduling of any direct contact is always made on the first available slot of a supply type compatible with the type requested. The requested dates range from 0 to 168 days, equivalent to six months, and the simulator uses different distributions for external and internal demand. Doctors tend to ask for wider date spreads.

4.1 Base configuration

The first set of simulations concerns the original FHU and their access rules. There is no clinic triage of urgent and non urgent requests. When a user asks for an urgent appointment, a slot will be given on the same day or the next available day following the substitution protocol. Demand is scheduled on a first come first serve order on the first available slot starting from the date requested by the agent.

The data presented here was obtained with 20 different runs of a two year period starting with an empty agenda. The first year is used as a warm up period. The average number of contacts (and standard deviation) is 39,335.7 (204.4), out of which 19,697.6 (160.1) are indirect contacts, and 6,311.5 (67,0) are urgent direct contacts. The fraction of urgent contacts that are conducted by a substituting doctor is 52.28%. In the left plot of Figure 5 we present the evolution of the average backlog of direct appointments calculated at the end of each month as produced by the 20 simulation runs. The simulation starts in the beginning of April 2006 and ends at the end of March 2008.

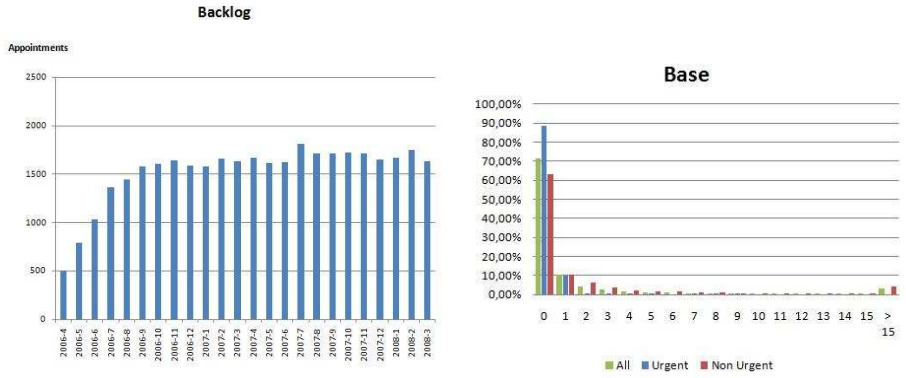

Figure 5: Backlog for the base configuration (left) and deviation between date requested and date obtained (right).

Since the initial agenda has zero backlog we see how the backlog grows in the first eight months until

it appears to reach a stationary behavior. Each bar in the plot corresponds to the average for that month of the 20 runs. The average backlog during the second year is 1,705.9 appointments with a standard deviations of 56.75. The 95% confidence interval is between 1,671.26 and 1,740.55 appointments. Note that the average backlog is about 8.7% of the annual demand for direct contacts.

The distribution of w is presented in the right plot of Figure 5. 88.38% of urgent calls are booked with $w = 0$ all thanks to the substitution mechanism in place. 98.52% of the urgent requests are scheduled with $w \leq 1$. The percent of direct appointments which get $w = 0$ is 71.27% and the non urgent contacts are satisfied on the requested date 63.16% of the time. 90.93% of non urgent contacts are satisfied with $w \leq 8$. There are some direct contacts which get $w \geq 15$, mainly due to the fact that some requests occur for a period when the family doctor is away on holidays.

With the no show model used, the base configuration produced 10.03% no shows to scheduled appointments. In the following section we present variations on the access policies and will compare the achieved performance with this base configuration.

5 Base configuration with postponement

The first variation consists on imposing a 14 day postponement for all direct contacts, external and internal. The simulator allows the definition of different postponement bounds for external and internal requests. The 20 simulation runs produced an average (and standard deviation) of 38,796.8 (183.3) total contacts during the second year, out of which 19,718.2 (150.3) are indirect contacts, and 6,291.2 (62.5) are urgent direct contacts. The substitution percent gets a slight reduction to 50.93%. The backlog of direct contacts changes dramatically as expected from Theorem 2, as displayed in Figure 6 – left plot, which has an average of 408.71 (61.35) which is slightly over 2% of the annual demand for direct contacts.

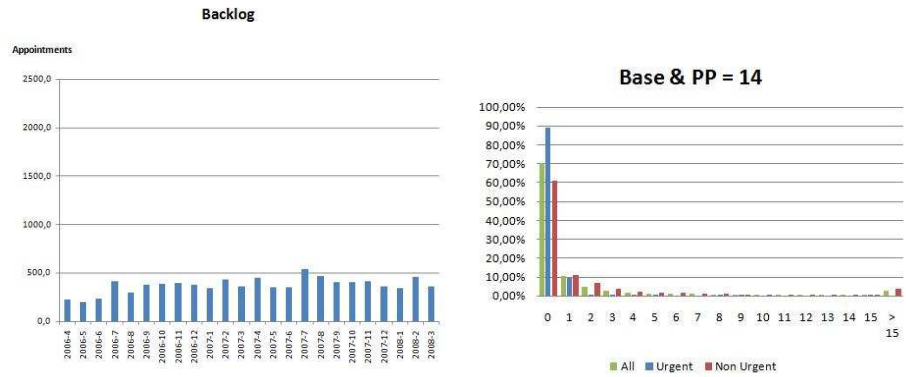

Figure 6: Backlog with a postponement of 14 days (left) and distribution of date deviations (right).

In the right plot of Figure 6 we present the distribution for w. The fraction of urgent requests satisfied with $w = 0$ is 89.31%, a 1% improvement over the base configuration. 98.61% of the urgent requests are scheduled with $w \leq 1$. The non urgent requests get $w = 0$ around 61.24% of the time, close to a 2% degradation relative to the base configuration. 90.74% of non urgent contacts are satisfied with $w \leq 8$. The set of all contacts gets $w = 0$ about 70.49% of the time. No shows drop to 9.17%.

5.1 Base configuration with carve-out

The second variation consists on introducing carve-out. The configuration tested is as follows. Each doctor has three slots at the end of his/her office hours specifically reserved for urgent requests, defining a new supply type, say 0. The total weekly supply is therefore $3 \times 5 \times 8 = 120$. This was done by converting some type 1 slots into type 0 slots. For the FHU under study this corresponds to 30 minutes per doctor per day, as the urgent direct appointments have a nominal duration of 10 minutes.

An incoming urgent request is firstly booked in the type 1 slots of the family doctor. Only if there is no regular slot available, the request will be booked on the carve-out slot. If this fails, the next attempt will be an alternative doctor on type 1 slots. If all these three fail, the request will move to the next day, where it will only have access to type 1 slots. Note that we do not use the carve-out slots for substitution, nor for "yesterday's" urgent requests. In summary, the carve-out slots are used as a last resort for the family doctor and only on the days the requests are generated. This also means that a booking in a

type 1 slot of the family doctor is usually better because it will take place sooner than the carve-out appointment.

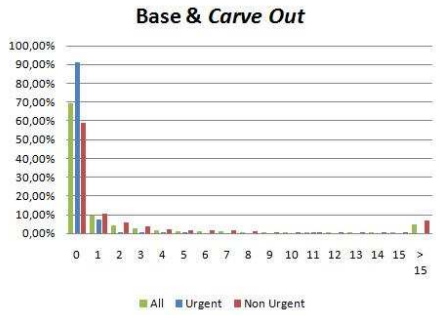

Figure 7: Distribution of dates deviation.

The total number of contacts produced in the second year (and standard deviation) were 39,323,8 (210.2), out of which 19,686.9 (151.4) were indirect contacts, and 6,333.5 (60.2) were urgent contacts. The fraction of substitution appointments drops to 42.66%, a close to 10% improvement over the base configuration. The number of no shows is similar to the base case, as should be expected, that is 10.18%. In Figure 7 we present the deviation of dates achieved. Now we get 90.95% of urgent requests satisfied the day they occur, which is over 2.5% more than in the base configuration. 98.64% of the urgent requests are scheduled with $w \leq 1$. Naturally there is a degradation on the non urgent requests, which drop to 58.94%, a little over a 4% reduction. 90.55% of non urgent direct contacts are satisfied with $w \leq 12$.

5.2 Base configuration with carve-out and a 14 day postponement

Finally, we combine the two mechanisms. The total number of contacts produced in the second year (and standard deviation) were 38,807,3 (165.3), out of which 19,094.2 (144.9) were indirect contacts, and 6,280.6 (105.1) were urgent contacts. The fraction of substitution appointments is 42.61%, in line with the previous variation. The percent of no shows is 9.29% in line with the second variation.

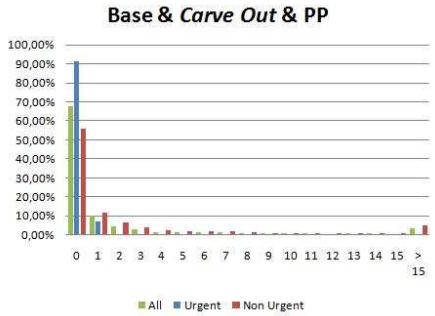

Figure 8: Distribution of dates deviation.

Figure 8 displays the achieved distribution for w. Now we get 91.57% of urgent calls with $w = 0$, a little over 3% improvement over the base configuration. 98.68% of the urgent requests are scheduled with $w \leq 1$. The non urgent requests that get $w = 0$ are 55,78% of the total such requests. 90.83% of non urgent contacts are satisfied with $w \leq 11$.

We also run 20 simulations for a postponement of 7 days and with the same carve-out levels. For that case, the average 6,307.3 urgent requests had a 42.64% substitution and the no show rate dropped to 8.26%. 91.92% of the urgent requests get $w = 0$. 98,73% of the urgent requests are scheduled with $w \leq 1$. 90.09% of the non urgent requests are scheduled with $w \leq 10$. The average backlog is 287.31 (61.25), which corresponds to a little over 1.5% of the annual average volume of direct contacts.

So we can conclude that only the carve-out mechanism moves some of the urgent appointments that were satisfied with $w = 1$ to become satisfied with $w = 0$ and increases the likelihood of a user being seen by the family doctor on an urgent appointment. The first introduction of carve-out pushed the 90% non urgent appointments to be satisfied within 12 days, but as the postponement parameter decreases, the

number of days to accommodate 90% of the non urgent requests recedes. Therefore, carve-out degrades the immediate satisfaction of the non urgent requests but the postponement may compensate the short term satisfaction.

We argue that a few extra days of waiting time for non urgent requests are not as harmful as it may be one extra day waiting time for an urgent request, and/or increased probability of not being seen by the family doctor.

6 Conclusions

We presented a discrete event simulator that is used to model outpatient clinics and serves the purpose of investigating the relative performance of different supply profiles and access policies. The simulator is sufficiently general to be applicable to any Family Health Unit and sufficiently flexible to accommodate a series of different access policies.

With this tool we conducted a series of experiments to understand the effects of some features of an access policy, given a fixed supply profile. The formal results we present imply that the debate on the adequacy of the traditional, carve-out, and open access policies is somewhat sterile. There is nothing structurally wrong with any of the three mechanisms, and the reports in the literature arguing against one to favor another are seldom based on extensive numerical studies. More frequently, the arguments are qualitative in nature and based on episodic experiences. The fair conclusion one should make out of those reports should emphasize the fact that people and organizations tend to violate organizational rules and methods. So, rather than criticizing the policies, a closer look should be placed on how the rules are broken on a daily basis, creating havoc on the organizations.

The most recent literature proposes the second generation open access as the adequate form of managing access to the doctor's agendas. It is centered on doing today's work today and argues that the urgency degree of appointment requests should not influence the access policy. Whereas there may be clinics where such principles are applicable, in the case of the Family Health Units the situation is diverse. In fact, one performance measure used in the FHU's evaluation procedures concerns the service level provided to urgent requests. If there were no doctor substitution and no carve-out, each urgent request would be treated the same way the non urgent requests are. Therefore, to improve their service level to acceptable values one would have to invest in increased capacity, which would likely render a very expensive system. Also, a FHU is stationary in nature, as the staff is not looking to increase the demand by enrolling more users. This fact alone makes these clinics relatively different from private clinics, where a pay per service environment may condition differently the managerial options and access policies.

From the numerical data presented here, we can draw a set of very interesting insights. Introducing a postponement discipline does not change the service level of any type of appointment. The performance measures directly affected by postponement are the no show rates and the backlog volume. Besides the no show rate, another feature influenced by low backlog volumes is the volume of potentially wasted capacity for bookings that will be waived at a later time. Bookings made ahead of time that may be waived, block the agenda for other bookings and, after being waived, may eventually not be put to use. When postponement is enforced, future no shows and waived bookings are replaced by open slots in the agenda, increasing the available capacity.

It is our understanding that all backlog is bad when its total volume is significant. A clinic with high backlog volumes cannot adjust easily to short term pattern changes or to unexpected events. If during the flu season there is a significant surge in urgent requests, it will be difficult to accommodate it without disturbing the bookings for a large number of people. Also, if a physician gets sick and has to miss work for a few days, the number of bookings that has to be rescheduled is much lower with low backlog volumes. Low backlog volumes are associated with agendas with more openings and with agendas where it is easier to introduce short term supply changes.

The introduction of doctor substitution is one way to improve the service level to urgent requests, irrespective of the presence of postponement. However, if one also wants to increase the likelihood of a user being seen by the family doctor, carve-out is the adequate option. The adequate combination of postponement and carve-out allows an increase of service level quality to urgent requests without excessively degrading service level quality for non urgent requests.

There are other issues that deserve a thorough analysis in general, like balancing the panel sizes of the physicians and carefully tailoring the supply volume for the individual types and physicians, for instance. As these concern more the problem of the demand profile they fall outside the scope of this work.

In any case, the developed simulator is a tool available to each FHU where many other features can

be tested before changes are implemented, providing also information on the transient behavior when changes are implemented.

References

[1] C. Grinstead and J. Snell, "Introduction to Probability: Second Revised Edition, AMS (1997)

[2] G. Latouche and V. Ramaswami, "Introduction to Matrix Analytic Methods in Stochastic Modelling", 1st edition, ASA SIAM (1999)

[3] M. Lee and K. Silvester, Case study to demonstrate the principles in the paper 'Reducing waiting times in the NHS: is lack of capacity the problem?', Clinician in Management, Vol. 12 (2004)

[4] A. Mehrotra and L. Markowitz and J. Ayanian, Implementing Open-Acess Scheduling of Visits in Primary Care Practices: A Cautionary Tale, Annals of Internal Medicine, Vol. 148 (2008)

[5] M. Murray, Evaluating Open Access: Problems with the Program or the Studies? Annals of Internal Medicine, Vol. 149 (2008)

[6] M. Murray and D. Berwick, Advanced Access: reducing waiting and delays in primary care, Journal of the American Medical Association, Vol. 289, No. 8 (2003)

[7] M. Murray and C. Tantau, Redefining Open Access to Primary Care, Management Care Quaterly, Vol. 7, No. 3 (1999)

[8] M. Neuts, "Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach", Dover Publications Inc. (1981)

[9] K. Silvester and R. Lendon and H. Bevan and R. Steyn and P Walley, Reducing waiting times in the NHS: is lack of capacity the problem?, Clinician in Management, Vol. 12 (2004)