Providing Timely Access to Care: What is the Right Patient Panel Size?

Linda V. Green, PhD

Sergei Savin, PhD

Columbia Business School

Mark Murray, MD, MPA

Mark Murray & Associates
Abstract

**Background:** Timely access to care is a key component of high quality healthcare. Yet, delays for appointments are prevalent, resulting in patient dissatisfaction, higher cost and possible adverse clinical consequences. A “just-in-time” approach to patient scheduling, called *advanced access*, has been demonstrated to be effective in reducing delays in multiple clinical settings. The fundamental goal of advanced access is to offer most patients appointments on the same day. This requires achieving an appropriate balance between supply of and demand for appointments, but there have been no methods previously proposed to determine what this balance should be.

**Methods:** We propose a measure of balance we call the *overflow frequency level* - the fraction of days when demand exceeds the average number of appointment slots available. We develop a probability model to estimate this measure for any practice and describe its use in identifying an appropriate panel size or, conversely, the physician capacity needed to provide timely access for a given practice.

**Results:** Delays for appointments will be excessive unless the ratio of the average daily demand for appointments to the average daily capacity is less than one. The appropriate value of this ratio is dependent on the desired *overflow frequency level*, which should be kept below 25% if the goal is to offer most patients same-day appointments. A table of suggested panel sizes for a range of practice types is presented, and an Excel file is available on request to help determine panel size or physician capacity in any specific situation.
Difficulty in getting a timely appointment to see a physician is a very common problem. In a recent study, 33% of patients cited “inability to get an appointment soon” as a significant obstacle to care\(^1\) and the Institute of Medicine identified “timeliness” as 1 of the 6 key “aims for improvement” in its major report on quality of health care.\(^2\)

**PRIMARY CARE AND ADVANCED ACCESS**

For most patients, their primary care physician is their major access point into the health care system. Yet primary care practices often have long waits for appointments and may have difficulty in accommodating patients who have potentially urgent problems. As a result, patients experience delays in treatment and may be seen by someone other than their own physician, potentially leading to adverse clinical consequences, patient dissatisfaction, and loss of revenue for the practice. Large backlogs may require additional staff and resources to deal with patients trying to get appointments for the same day, and are often correlated with a high rate of cancellations or “no-shows” which can result in lost income and wasted capacity.\(^3\)

To remedy this problem, some primary care practices have adopted a patient scheduling approach known as *advanced access*. As opposed to a “traditional” system where each physician’s daily schedule is fully booked in advance, or a “carve-out” model in which a fixed number of appointment slots are held open for urgent cases, the goal of the advanced access approach is to reduce delays by offering every patient a same-day appointment, regardless of the urgency of the problem. The fundamental idea behind advanced access is to “do all of today’s work today”, so that patients don’t have to wait for appointments, practices don’t waste capacity holding appointments in anticipation of same-day needs, and patients
have a greater likelihood of seeing their own physician. Several success stories have
documented the benefits of this approach in both managed care and fee-for-service
environments including dramatically shorter waits, higher levels of continuity of care, less
wasted capacity for the practice, and increased patient, staff, and physician satisfaction.3

Advanced access can only work if patient demand for visits and physician capacity to
see patients are “in balance”. Advocates of advanced access identify several ways in which
the number of visits can be reduced, physician time can be better leveraged, and scheduling
practices can be streamlined so as to achieve a better balance between supply and demand.3,4
However, in discussions with practitioners, we have found that questions remain about what
constitutes an appropriate balance and, more specifically, what is a “manageable” panel size.
The answers to these questions are not obvious and require a quantitative approach.

THE NEED FOR “SAFETY” CAPACITY

A fundamental feature of patient demand for primary care is its random nature: the actual
number of patients requesting care on any particular day will vary around the average daily
value, sometimes substantially. It is this inherent randomness that makes it difficult to
determine the answers to questions such as: “How large a panel size can be served by a given
physician practice?” If not for this variability in demand, the answer would be obvious – the
panel size would be the one which made the daily demand for care equal to the daily number
of physician appointment slots available. However, with this variability, making supply and
demand equal on average would create chronic backlogs for care and waits for appointments
that would likely get longer and longer.5 Though this characteristic of service systems has
been known to operations professionals for decades, it seems counter-intuitive. A simple example, illustrated in Figure 1, may help explain this critically important concept.

Consider a primary care practice which has a daily patient demand for appointments which takes on only two possible values - 11 and 9, each with 50% probability. Suppose the maximum number of patients that can be seen each day is exactly equal to 10, so that any “excess” demand must be pushed to the next day that has available appointment slots. Fig. 1 illustrates all possible realizations of patient backlog values over a period of 3 days, assuming we start with no backlog. As shown, the average backlog grows from 0.5 patients at the end of the first day to 0.75 at the end of the second day to 1.0 at the end of the third day. If this exercise is carried out further, the average patient backlog will continue to grow from day to day. This may be surprising since it seems logical to assume that “bad” days, i.e. days with a demand of 11, will be balanced out by “good” days, those with only 9 new patient demands. So why doesn’t this balancing out happen? As our simple example shows, when patient demand is less than the appointment capacity, the extra service capacity cannot be transferred to the next day to serve future patient demand, and is therefore lost. On the other hand, on the “bad” days, when patient demand exceeds service capacity, the unserved demand does not disappear and has to be satisfied in the future. So “good” days cannot clear the backlog created by the equal number of “bad” days. Furthermore, if the demand variability is increased, for example by adding possible demands of 8 and 12 patients, the average backlog will grow faster.

Thus, if the goal is to provide immediate access to care with a high probability, then the average daily demand for appointments must be strictly less than the maximum capacity to see patients. Another way of saying this is that there must be some safety capacity relative
to demand. Safety capacity, the amount of capacity in excess of average demand, serves as a hedge against demand variability. Without it, a practice will be unable to offer timely care to its patients.

**FINDING THE RIGHT BALANCE BETWEEN SUPPLY AND DEMAND**

How much safety capacity does any specific practice need? This depends primarily on the desired *overflow frequency level* - the percentage of days when demand exceeds the number of appointment slots for that day. In our example above, the overflow frequency is 50%. The lower the overflow frequency level, the easier it will be to offer same day appointments by occasional use of physician overtime. Decreasing the overflow frequency level can only be accomplished by increasing the safety capacity. But more safety capacity also means more days and hours when physicians are not seeing patients.

So the “right” level of safety capacity for any given office must be a subjective determination that will likely be based on the tradeoff between the revenue associated with seeing more patients and the amount of overtime the practice is willing to undertake to keep patient delays minimal. In order to evaluate the possible tradeoffs, it’s necessary to understand the relationship between safety capacity, patient panel size, and overflow frequency.

**A MODELING APPROACH**

Safety capacity can be created by either increasing physician capacity or decreasing demand. Physician capacity may be increased by adding appointment slots to the day and demand might be reduced by using tactics such as greater use of telephone and e-mail and the use of
group visits. However, panel size is the major determinant of demand and the prime lever for achieving the right balance between supply and demand.

We’ve developed a simple quantitative model to help evaluate the tradeoffs associated with a given panel size. Since the only objective of this model is to help identify a good balance of overall supply and demand for a given practice, it is not necessary for the model to distinguish between “external” demands, i.e. those that are generated by patients’ actions, and “internal” demands, i.e. those that are the result of the physician’s decision to see a patient for follow-up work or chronic care. No matter the source of the demand, all demands must be satisfied in a timely fashion and doing so requires a panel size that allows for sufficient safety capacity.

The model does not address the “micro-management” issues such as daily scheduling of follow-up visits, dealing with cancellations, or scheduling of physicians’ office hours and vacations. While these are all important factors for the efficient functioning of the practice, they do not significantly affect the best choice of panel size and so are not needed in our “macro” model. We will revisit these issues later in the paper.

Though about 2/3 of all primary care physicians work in group practices, there have been a number of studies that document the benefits of continuity of care. These observations support the view that a patient should be seen, whenever possible, by his/her physician, and therefore, that a panel should be associated with an individual physician. However, as described later, our approach can easily be extended to allow for determining a panel size for multiple physicians working as a team.
FINDING THE RIGHT PANEL SIZE

Establishing an appropriate panel size for an existing practice consists of 6 steps: 1) identifying the current panel size, 2) estimating the daily visit rate per patient, 3) fixing the number of daily appointment slots, 4) calculating the current overflow frequency, 5) setting the target overflow frequency, and 6) computing the panel size based on the target overflow frequency. Steps (5) and (6) can be done iteratively to identify a desirable trade-off between panel size and overflow frequency. (An Excel file which provides all necessary computations for these 6 steps is available from the authors upon request).

Identifying the current panel size

In many managed care practices, the patient panel size $N_{\text{cur}}$ is simply the number of patients enrolled with a physician. But in fee-for-service or mixed practices, the number of patients “on file” may be misleading since it is not uncommon to preserve files for patients who may no longer be using the practice’s services. In these situations, it has been found that the panel size will be most accurately estimated by calculating the total number of distinct patients seen by a physician in the last 18 months. (Use of a year may underestimate the effective panel size, while the two-year count typically produces an overestimated value).

In a multi-physician practice, estimating the current panel size for each physician may be more complicated since a given physician’s patient may see another physician if his/her preferred provider is unavailable. Therefore, in these practices, it is important to track for
each physician the number of requests for appointments, rather than the number of actual visits.

**Estimating the daily visit rate per patient**

The most accurate assessment of daily demand requires prospective measurement of the specific appointment dates that patients actually ask for including walk-ins (external demand), as well as the follow-up visit dates physicians actually request (internal demand). If prospective data is not available, an estimate can be obtained by examining appointment logs for a recent period of time, e.g. 18 months, and counting the number of appointments over that period of time. Let $T$ be the number of working days for the period of time being examined and let $A$ be the number of patient appointments (or, if available, requests for appointments) for those $T$ days. Then, as shown in BOX 1, the daily visit rate per patient $p$ is calculated by dividing $A$ by the product of the number of patients on the current panel, $N_{cur}$, and $T$: $p = \frac{A}{N_{cur} \times T}$. For example, consider a general/family practitioner with a current panel of $N_{cur} = 2500$ patients who had $A = 6500$ office visits over the last 18 months ($T=315$ days). For this practice, $p = \frac{A}{N_{cur} \times T} = \frac{6500}{2500 \times 315} = 0.008$ visits/day per patient.

**Establishing the number of daily appointment slots**

The average daily supply of appointment slots, $C$, is determined by the average length of an appointment slot and the average daily number of hours devoted to direct patient care. So if a physician spends an average of 7 hours per day in patient care and appointments are scheduled 20 minutes apart, the daily appointment capacity is $C = 7 \text{ hours} \times 3$
appointments/hour = 21 appointments. If a practice has a varying number of appointment slots per day over the week, \( C \) should be the average number of slots per day.

**Calculating the overflow frequency**

Consider a practice with panel size \( N \) and with daily demand rate \( p \). If each patient request for care is generated independently of any other patient’s request, the total daily demand for primary care services on any given day can be modeled as a *binomial* random variable with expectation equal to \( Np \) and variance equal to \( Np(1-p) \). The binomial random variable with parameters \( N \) and \( 0 < p < 1 \) describes the random number of "successes" in \( N \) independent trials when the probability of success in any single trial is \( p \). In the primary care environment, this binomial random variable corresponds to the number of appointment requests which a patient panel of size \( N \) generates on a given day. (A more detailed description of the properties of the binomial random variable can be found, for example, in Bertsekas and Tsitsiklis\(^{12} \).) Using this model and the number of appointment slots each day \( C \), we can estimate the effect of *any* panel size on the overflow frequency by calculating the probability that the demand for appointments exceeds the supply of slots on any given day, as illustrated by the formula in BOX 2. Using this formula with \( N = 2700, p =0.008 \) and \( C = 21 \), results in an estimated overflow frequency of 49.4%.

**Computing the appropriate panel size**

For a practice that operates 5 days a week, an overflow frequency of 49.4% implies that to avoid patient delays, the physician will need to see patients during “overtime” more than twice a week on average. It’s important to note that the higher the overflow frequency, the
greater the average backlog and so the longer the overtime needed to “do today’s work today”. In our example the average duration of overtime when it occurs can be shown to be more than an hour. It’s also important to understand that since overtime frequency is a long-term average, in any given week it could be considerably higher leading not only to substantial overtime but long backlogs for appointments as well.

So in our example, the current panel size would need to be reduced to be able to comfortably and consistently offer same day appointments. This doesn’t mean that the panel size need be small enough to lead to a near zero likelihood of overflow frequency. Infrequent overflows, e.g. 5%, 10% or even 20%, are likely to be small enough that they can usually be handled with occasional and modest levels of overtime and therefore not jeopardize future appointment capacity. On the other hand, the smaller the overflow frequency, the lower will be the average daily utilization of the practice, $pN/C$. In selecting a target panel size and therefore a target level of overflow frequency, a physician should take into account his/her own tolerance for overtime work: 5% (approximately once a month), 10% (once in two weeks), or 20% (once a week).

If the current panel size results in an overflow frequency that is too high, as in our example, a more appropriate panel size can be found by decreasing it and recalculating the overflow frequency using the formula in BOX 2. On the other hand, if the computed overflow frequency is lower than desired, the panel size should be adjusted upward. This process of adjustment and recalculation should be repeated until the overflow frequency computed for the trial value of the panel size is close enough to the desired overflow frequency.
In our example with an initial panel size of 2700, assume that the desired overflow frequency level is 20%. Since the computed current value of the overflow frequency (49.4%) is much higher than the target, we might try a panel size of \( N = 2000 \). Using the BOX 2 calculation for this value of \( N \), we obtain an overflow frequency of 8.8%, which is lower than our target. So on the next iteration, we can try a somewhat larger panel, \( N = 2300 \), which produces an overflow frequency of 22.8%. Since this is an estimate, this is probably sufficiently close to the target to be considered a good choice. Alternately, continuing iterations, one discovers that for the panel size of \( N=2250 \) the overflow frequency becomes very close to 20%.

**Examples based on NAMCS 2002 data**

Table 1 shows the patient panel sizes (and attained utilizations) for a “typical” general and family practitioner (on average, 1.575 annual visits per patient\(^*\), according to Murray and Berwick\(^4\)) and a “typical” pediatrician (on average, 1.98 annual visits per child according to 2002 NAMCS\(^10\)) which would result in an overflow frequency of 5% (approximately, once a month), 10% (twice a month), or 20% (once a week). 2002 NAMCS reported that the average duration of the “face-to-face” part of the office visit is 16.1 minutes for general/family practice and pediatrics, 18.1 minutes for OB/GYN practice and 20.0 minutes for internal medicine. In our calculations we considered appointment intervals of 20 minutes.

\(^*\) While the NAMCS 2002 survey reports the total number of annual visits to general and family practitioners in the United States (215,466 thousand), the annual visit rate per patient is not easy to estimate since we could not find reliable statistics on the number of people who actually use (or even have) a primary care physician. The rate of 0.761 annual office visits per person, reported in NAMCS 2002 survey, was obtained by dividing the total number of visits to general and family practitioners by the entire size of US population (283,135 thousand), taken from 2000 US Census data. Clearly, using this value would result in a gross underestimation of actual patient visit rates. The rate we use (1.575 annual visit per patient) is calculated based on the assumption of 210 annual in-office days and on the assumption (used in Murray and Berwick\(^4\)) that in an average patient panel not overly weighed with elderly and chronically ill patients, 0.07%-0.08% of patients will request visit on an average day. We note that this estimate is somewhat higher than 0.05% used by Smoller\(^12\).
Under the assumption of an 8-hour workday (for a 5-day working week this roughly corresponds to the 40.2 hours spent by a family physician on direct patient care or patient-related service during a complete week of practice\textsuperscript{13}), this results in 24 daily appointment slots. Since the actual daily appointment capacity is likely to be somewhat lower, we also consider a daily capacity of 20 appointment slots. The calculations were performed using the formula in BOX 2 under the assumption of 210 work days per year. This value, in our estimate, is a good representation of the annual number of work days for a large number of primary care practices.

**ADJUSTING SUPPLY FOR A FIXED PANEL SIZE**

Though the above analyses addressed the issue of determining panel size, the same approach can be used to determine appropriate physician capacity for a given panel size. For the case where continuity of care is considered important, capacity will be the number of daily appointment slots needed by a single physician to handle the proportion of the panel that represents his/her patients. This can be done by using the binomial formula in BOX 2 to assess the overflow that would result from each possible alternative and choosing the minimum number of slots that keeps the overflow within a “tolerable” limit. In a multi-physician setting where continuity is not considered critical, the daily capacity would be the number of slots per day for each physician multiplied by the number of physicians, and the analysis would be done using the possible alternatives as before. It is important to note that the total panel size that can be handled by a group practice in which continuity of care is not considered paramount will likely be significantly larger than the sum of the individual panel sizes.
sizes of the individual physicians in the practice if the goal is that patients see their preferred physician with high probability.

ACHIEVING THE RIGHT BALANCE

The analyses described above can be easily modified for any particular physician practice. This requires that data be collected to accurately assess both supply and demand. As Murray and Berwick point out⁴, historical data may be misleading, since it measures activity which may be less than demand if a practice has experienced lost and deferred demands. Therefore, it is important that demand be measured prospectively. In doing so, both weekly and seasonal patterns should be considered to identify times of particularly high (and low) levels of demand. For example, there may be several months each year with particularly high demand due to flu season. In this case, accurate records of demand are important in order to estimate a seasonally adjusted patient visit rate per day which can then be used in the binomial model to help identify capacity needs during these times. Physician supply can then be adjusted accordingly if part-time physicians are available. Vacation times and other activities should be scheduled during lower demand seasons and days if possible to assure sufficient capacity during higher demand times. In addition, it is important to identify the fraction of the demand which can be managed to offset the variability in the unscheduled demand³. Patients who need follow-up appointments should be scheduled early in the day on lower demand days and/or during lower demand times of the year. Of course, in any given practice, there will be constraints on both physician and patient scheduling and the above guidelines are just that – goals to work towards. To the extent that they can be followed, daily delays for appointments will be reduced.
CONCLUSION

Assuring timely access to medical care is an important goal for any physician practice and advanced access provides some specific guidance in achieving it. However, the variability inherent in the demand and delivery of healthcare makes it impossible to determine specific answers to questions about panel size or, conversely, physician practice size by using guesswork or intuition. In this paper we have described a simple probabilistic model that can be used to supplement the qualitative approach of advanced access to make major improvements in the timeliness of care while considering the constraints on physicians’ working hours.

References


Figure 1. Example illustrating the growing patient backlog in the case when average daily patient demand equals the appointment capacity.

Table 1. Panel sizes (capacity utilizations) for different parameter values, primary care type: general and family practice and pediatrics.
Calculating the daily demand rate for a panel of current size $N_{\text{cur}}$

1. Choose an observation period (for example, 18 months) and calculate the number of working days $T$ within this period.
2. Count the number of patient visits, $A$, over those $T$ days.
3. The daily demand rate for appointments (per day per patient) is

$$p = \frac{A}{N_{\text{cur}} \times T}$$

where $N_{\text{cur}}$ is the current panel size.

BOX 1

Calculating the overflow frequency

If the daily patient demand is modeled as a binomial random variable with parameters $N$ (panel size) and $p$ (demand rate), the probability that the number of patients will exceed the number of available slots $C$ (overflow frequency) can be calculated as:

$$\text{Overflow frequency} = 1 - (1 - p)^N - \sum_{k=1}^{C} \frac{N!}{k!} \left(1 - \frac{N}{k} \right) \left(1 - \frac{N-1}{k} \right) \cdots \left(1 - \frac{N-k+1}{k} \right) p^k (1 - p)^{N-k}$$

where $k$ is the index of summation.

This expression can also be re-written as

$$\text{Overflow frequency} = 1 - (1 - p)^N - \frac{N}{1} p(1 - p)^{N-1} - \frac{N(N-1)}{1 \times 2} p^2 (1 - p)^{N-2}$$

$$- \frac{N(N-1)(N-2)}{1 \times 2 \times 3} p^3 (1 - p)^{N-3} - \cdots$$

$$- \frac{N(N-1)(N-2)\ldots(N-C+1)}{1 \times 2 \times 3 \times \ldots \times C} p^C (1 - p)^{N-C}$$

BOX 2