Examining Demand, Supply and Activity (DSA)

While the terms demand, supply and activity (DSA) and the relationships between them may seem simple, measuring them can be problematic. Although demand, supply and activity can capture any type of workload, the comments here refer solely to office appointment workloads.

Demand

Demand is a measure of workload. The term demand encompasses several facets, but for the sake of comparing "same unit to same unit," this discussion focuses on appointment demand only. There are multiple ways to measure demand. It can be measured on a yearly or a daily basis. It can be measured by panel size X visit rate, by the sum of new appointments plus return appointments (in specialty care), or by the sum of the appointments generated on a daily or weekly basis. Each method has its challenges. It is critical not to confuse demand with activity (which is discussed later in this document).

Yearly Measure of Demand

Yearly demand can be measured as the number of patients in a practice or individual provider's unique panel of patients X the expected number of visits per year per patient. Panel and visit rate viewed independently are indirect measures of demand. Measuring demand on a yearly basis has some challenges:

1. Provider turnover often leaves "orphan" demand if a provider leaves the practice and therefore leaves behind unattended patients. The demand may seem "hidden" but actually continues on for the remaining providers in the practice.
2. Visit rates can change due to behavior change on the part of either the provider or the patients.
3. There are seasonal variations (vacations, etc.) in both demand and supply which makes sub-dividing these measures by the month somewhat problematic.
4. Last year's visit rates reflect last year's performance/behavior and that can change with either improvement or deterioration in practice.
5. The number and rate of new patients has an effect on demand since new patients generally exhibit more initial demand.
6. Losing established patients (patient turnover), for whatever reason, has an effect on demand.

Daily Measure of Demand

Measuring demand on a daily basis and comparing it to supply and activity has greater utility for a practice. Seasonal and day of week variations in either demand or supply can be determined and contingency plans implemented in anticipation of these variations. Variations in demand from one

provider to the next can be also determined. While these variations can also be seen in the visit return rate/interval, these measurements are long-term and can miss short term trends.

Demand is workload generated – a measure of the number of appointments made within a specified time frame. In primary care, that time frame, by definition, is a day. Demand is calculated as workload generated ON today. It is calculated as the number of appointments actually generated today, but not necessarily seen or appointed for today.

While the operational definition of demand as "workload generated" is the same for both Primary Care (PC) and Specialty Care (SC), it is helpful to divide demand into different functional categories in PC and in SC. In primary care, demand is arbitrarily divided into the categories of external demand (work generated from outside the practice) and internal demand (work generated from inside the practice). Total demand is the sum of external plus internal. By definition, internal demand includes only those patients who make an appointment as they leave the practice. All other demand is external. External and internal demand are not distinct appointment types but are functional categories that describe how the work is appointed and where it comes from.

In specialty care, demand is arbitrarily divided into the categories of "new" and "return" demand. New and return function as distinct appointment types. Both of these appointment types are pre-appointed (appointed into the future). Total demand is the sum of new plus return.

Primary Care: External Demand

External demand is counted as all patients who call, walk-in or find another way in, and make an appointment. While calls create some delay between declaration of need and delivery of service, for walk-ins who get an appointment, there is almost instantaneous movement of demand to delivery. Some "return" patients are external (they leave without making an appointment and call back later to book their return appointment). Some "external demand" includes patients who call today, ask for an appointment today and get it, and some of the external demand includes patients who ask for an appointment today but get the appointment at a future date. The assumption here is that the number of patients who ask for an appointment for tomorrow will equal those who did the same thing yesterday, and those who ask for an appointment next week will equal those who did the same thing a week ago. This measurement of demand is a thought experiment - a way to conceptualize and measure demand in order to avoid confusion with supply or activity, and a way to compare demand to supply and activity.

Primary Care: Internal Demand

Internal demand is the workload generated directly from inside the practice. It is thus restricted to those patients seen today who make a return appointment today while they are in or as they leave the office. This workload is counted as the number of patient appointments made ON today as the patient leaves the office. All internal demand is "return appointments" but internal demand does not include all the returns. Some returns may present as external (they leave without making an appointment and call back later to book their return appointment). Those patients given instructions to call back later for a return appointment will be captured on the day they call back to make an

appointment (whether it is for the day they call or after the day they call). Again, this is arbitrary – it is a thought experiment.

In counting daily demand either manually or by computer, it should be noted that patients who call in, make an appointment today for today, are seen, and, as they leave, get a return appointment, are counted twice. Keep in mind that demand counts appointments not patients.

The distinction between external and internal demand is important because we can directly influence or control the internal demand (return visits). We can influence the day of week and time of day the appointments are booked for, and we can control the re-visit intervals of returning patients. This is where the concept of "sell early in the day, sell late in the week" (the times a practice is usually less busy) grows from. In some practices, a low proportion of visits are "internal" (student health centers, for example). In others, a low proportion of visits are "internal" by choice. A modification of this choice offers an opportunity to shape demand. In some practices, a high proportion of demand is internal (new pediatrician with all newborns, for example). Some of these practices may have to use a carve out model, but most can use their data to effectively shape demand.

To reduce no-shows due to long waits, some practices deliberately restrict internal demand - they make no future appointments. Unfortunately, by doing this, they limit their ability to influence demand, risk losing patients to follow-up, risk breaking the common and associated promise of "we will see you when you call," increase the risk of over or under-filled schedules, restrict their ability to measure demand, and become more susceptible to daily demand or supply variations. These practices often have no idea whether their supply can meet demand and have no idea whether the cyberspace backlog of appointments is increasing or is stable. I call this "access by denial" (a practice gets "space" by pretending that the demand does not exist). There is a "sweet spot" for the timing of future open schedules. Two weeks is too short and a year is too long.

Some groups are tempted to count all returns as internal demand, that is, to make the terms "internal demand" and "return visit" synonymous. To do that though is extremely challenging. A group would be required to develop a common practice-wide understanding with associated standardized workflow rules in order to define the term "return." There is a great deal of confusion over the term "return." Some interpretations could be:

1. A patient who leaves and asks for and gets a return appointment (the strict interpretation used to describe internal demand as outlined above).

2. A patient seen today and who is asked to call back for a "return" appointment at some later time.

3. A patient who is linked to or established with a specific provider, has either an ongoing clinical concern, or due to that established relationship calls and asks for a "return appointment." The term "return" can mean return for a specific problem, or simply a return to the provider, for another unrelated problem.

4. Other groups view patients in two distinct categories – either new or return. This is often driven by billing codes. Any patient who is not "new" is defined as "return."

Carefully defining the term "return" to mean only those patients who are previously seen and asked specifically to call back for an appointment, and developing clear workflow rules with multiple staff,

is very problematic. This number is difficult to collect because it involves multiple questions by receptionists and interpretation of the phrase "the provider asked me to return." Hence, manual collection of demand data is fraught with these interpretation and workflow rules issues.

On the other hand, a computer could successfully calculate total demand (and the subsets of "internal" and "external") by counting total demand as all appointments made on today and then subtracting the internal demand from the total. Internal demand is determined by looking for patients who made an appointment on today and were seen on today but the "seen on today" took place earlier than the appointment was made. Again, some patients make an appointment first, are seen second and make a return appointment third. The computer counts internal demand as only those second and third linked events (which also could be first and second linked events if the patient was appointed from the past seen today and made an appointment later in the day when leaving the office).

Counting all returns as internal demand can lead to a serious "reference point" and timing problem. If "returns" are counted as internal demand, at what reference point are they counted? The day they call or the day on which they get the appointment? "Returns" are managed far differently in various practices. Some practices encourage all patients to make an appointment as they leave. They feel they have greater influence over the timing of the return appointments in this way. Other practices, due either to patient preference, practice history or restricted short intervals for returns (patients can make appointments no further than two weeks into the future, for example) will encourage patients to "call back for a return appointment." In addition, in some practices patient and practice behavior results in an appointment being made on the day the patient calls for the appointment, while in other practices, most return appointments are made for a date future to the date the patient actually called. If there is a practice backlog of appointments, workload (demand) is often triaged and shifted to the future. Hence, practice and patient behaviors, structures and histories either reinforce or influence patient return behaviors.

All these factors lead to serious reference point and timing issues. Think of a three-dimensional box. How do we count demand? There is always a gap between declaration of demand and delivery of supply. That gap can be relatively short – a patient calls today and makes an appointment for today. The gap might be close to instantaneous – a patient walks in, asks for and gets an "appointment" (the definition of "appointment" is that his/her name is written in the reservation schedule) and this appointment happens within minutes. The gap might be longer when a patient calls and asks for an appointment next week, or gets an appointment for next month as she leaves. So in our it may not be delivered until somewhere across the box (sometime later). And if the far side of the box is at the sweet spot, for example, three months from now, some appointments requested are not delivered (appointed) at all. These appointments will be captured as demand on the day the patient calls for and gets an appointment. This event is captured as external, as outlined in the discussion above.

My recommendation, of course, in counting demand (workload generated), is to use a common reference point of a single day: count the demand as the workload generated on a single day. This has to be seen as a thought experiment as discussed above.

Specialty Care: New Demand

"New" demand in SC describes demand that is from patients who have not been seen before. In some practices this designation describes a new and unique patient whereas in other environments this term may describe a new and unique clinical condition. Unique value in SC is created by specialists seeing this new patient demand. Both customers, (the new patients) and the referring providers want new patient referrals to be seen in SC with minimal delays. If the patient visit rate in SC is greater than two patient visits per year, the number of return visits exceeds the initial new patient visits. As a consequence, if there is no distinction made between new and return appointment types, because of random variation there is a high likelihood that return appointments could overwhelm the new appointments on the schedule. If that event occurs, the delays for the valued new patients get arbitrarily squeezed further into the future. To prevent this form happening, and to protect new patient appointment capacity, new appointments are distinguished from return. New patient demand is then measured separately from return demand. If we were to compare categories of demand between Primary and SC, the new demand category in SC would be a sub-component of external demand in PC. However, while new patient demand is a small volume in PC, it is a significant volume of work in SC.

Specialty Care: Return Demand

The second category of demand in SC is "return" demand (all demand that is not new). While there is significant overlap, this term is not synonymous with the term internal demand as used in PC. Internal demand describes a single event - those patients in PC who in the act of leaving any visit, make a return appointment on the way out. The term return demand, as a category in SC, describes any patient who is not new. How the patient gets the return appointment is irrelevant.

Sub-dividing the demand in PC and SC into these arbitrary categories serves a purpose. In PC, we want to distinguish internal and external demand since it is the internally generated demand that we have the most control over. We can influence daily demand totals and begin to load level by moving the internal demand to days with less demand. In SC we need to carve out capacity for new demand to protect that capacity from being overwhelmed by return demand. We do this by having two distinct appointment types: new and return. In SC, we measure new and return as distinct demand streams and as distinct appointment types. New and return have their own demand, supply, activity and delay. In PC, external and internal are useful designations but we do not distinguish these categories on the schedule or as distinct appointment types. They are measured together when we compare demand to supply and activity.

Counting Demand Where it Lands or From Where it Originates

The tension here is on how to count the demand (workload). In the long run, in any way, demand is counted as demand booked.

We could, in theory, count demand not on the day it is generated or declared but on the day it is delivered (where it lands). In this scenario, an appointment requested on December 18 for December 27 would then be counted as December 27 demand. However, there are problems

associated with using this method. This interpretation of counting demand is further complicated by defining "internal" demand as "any patient asking for a return appointment."

For example, on December 18, patients that request and receive appointments for December 18 are counted as December 18 demand (external) unless we use the alternative interpretation of internal demand as "any patient asking for a return appointment." Even if we assume that the workflow rules are clear and consistently adhered to, how do we "count" those return patients who were seen prior to December 18, asked to call back for a return, and do so on December 18 but ask for a return appointment in the future, for example on December 27? Is this counted as December 18 demand? Is it December 18 external (my preference) or internal demand? This situation leads to the problem as noted above regarding staff interpretations of the term "return."

Furthermore, if demand is counted not on the day generated but on the day appointed, patients seen prior to December 18 but who call back on December 18 requesting a return on December 27 are counted as December 27 demand (either as external or internal) leading to the second problem above - the reference point timing issue. Using the three-dimensional box analogy, the appointments generated on December 18 at one side of the box are counted not as December 18 demand but demand somewhere else on the box - the date for which the appointment was requested and delivered. But, in a practical sense, just how do we count demand in this way and avoid confusing demand with activity (the work that is actually done)? Counting demand in this way becomes a self-fulfilling prophecy.

Counting demand where it lands - the date the appointment is made for - is also deceptive because each day has a limited number of appointments (supply), and the box has no limits on the far side. If D > S, then demand exceeds the limits of the future supply and gets pushed further and further into the future. While we count it on the day it lands, the day it lands has limits so demand appears to be the same as future supply on that day. Since we only count demand as appointments FOR a day, if there is a limit on the available appointment slots (and there always is) demand can get moved deeper into a backlog and then we simply cannot see when demand is greater than supply. Measuring demand in this way makes it appear that we are keeping up. Demand can actually only equal the number of appointment slots and not exceed that number, so when demand exceeds supply the true measurement is lost since the open ended backlog is used as a long-term buffer. Thus, it appears that the practice is keeping up. If there is a limit of, say, 30 appointments on any one day, since there are only 30 slots and the rest of the work can get pushed to the future, demand seems to = supply when in reality it might have been pushed past that day due to the fact that there were no more appointments.

On the other hand, if demand is counted as workload generated (the number of appointments made ON today), we see exactly how much demand there is. We cannot get lost in the backlog. We are using "today" as a reference point and looking at workload generated on the front end of the box. We can then ask, "How much supply do we have and can we meet that demand with our supply?" Activity shows us, at the end of the day, whether or not we actually used our supply and met the demand.

Supply

Supply is a measure of what could be done (the appointment space on the providers' schedules). This measure, which could also be called "supply available," reflects the total number of minutes the provider is available (plans to work) looking forward into the future ("I plan to work a session from 1:00 PM to 5:00 PM" = supply available of 4 hours or 240 minutes). Supply is intended to capture the planned work, not the actual. This is the amount of time the scheduler has in which to make appointments. This is the template in the scheduler. Supply is commonly converted into and measured as appointment slots.

While measuring supply in units of time is more sophisticated, the less sophisticated measure of supply as appointment slots is primarily explored here. This metric captures the number of appointment "slots" on the schedule. This is a prospective measure. While supply determination and measurement can be problematic when an entire year's worth of supply is calculated, measuring supply by the day is relatively easy and can be manually counted or counted by the computer. There are, however, some challenges when measuring supply by the year or by the day.

Yearly Measure of Supply

In a simplified way the supply here is measured as the panel size equation suggests - supply = days worked X expected number of appointment visits per day. The days worked measure can be problematic due to:

1. Provider turnover
2. Change in provider status – less or more days worked
3. The behavior in some practices of constantly changing schedules or the creation of partial schedules on an ad-hoc basis
4. Determination of what "counts" as a day worked

To determine days worked for the year, some groups look at decimals of days worked (i.e. a half day = .5 days, etc.). This offers a finer discrimination of the measure. Generally though, the key here is to count only days worked when seeing patients in the office. While providers can devote paid time to other pursuits, the measurement here is office appointment supply. Many groups confuse paid time with supply. Supply in this measurement definition refers only to appointment availability.

Appointment length also has an effect on the appointments per day part of the supply metric. The purpose of the length of the actual appointment slot (or "red zone") is to set a rhythm or pace for the work so the patient (the demand) and everyone on the team (the supply) can synchronize and therefore minimize waits. Truth in scheduling requires honesty about planning the right amount of time that the practice knows the provider will need. The decision about the right red zone length can be made a number of ways:

1. Provider desire ("I want to see a patient every 5 minutes").
2. Historical experience ("I know Dr. Z will spend 5 minutes for this as he always does").

3. Direct observation (someone with a stopwatch measures red zone length). This method is valuable since it shows the variation in actual time used.
4. Measurement of number of checked out appointments (activity or "supply used) as the numerator, over total minutes of supply available (denominator) = average amount of actual red zone time per appointment.

Daily Measure of Supply

This is the number of expected visits per provider per day. There are also some challenges to this method of measuring supply:

1. The appointment length has a direct effect on the number of appointments offered. This may mean that one provider, due to a different appointment length, may offer more or less supply than another provider, within the same length of time.
2. The number and ratio of short and long appointments on the schedule and the manner in which short appointments can or cannot be converted into long appointments.
3. As above, the behavior in some practices of constantly changing the schedule or adding partial appointment slot times into the schedule.
4. The basic schedule template design. Some groups use a short building block as a basic schedule template. They do not intend to book any appointment for less time than that. However, they may merge some of the smaller blocks to get a longer length of time for an appointment. We could call this the "merge up" strategy.

   Other groups use a longer appointment length as a building block, but fully intend to double book within some of the longer slots in order to get more visits within one longer "slot." We could call this the "merge down" strategy. The intention matters quite a bit here. In the "merge up" strategy, the supply is the basic number of slots but if the group merges up, they create fewer slots and have intentionally reduced supply. Initially, until we get more sophisticated, I would count supply as the basic set of appointment slots. If they merge up consistently then we could count supply as a predictable proportion of the supply set of appointment slots. In the "merge down" strategy, if the intention is to consistently "merge down," then there are functionally more appointment slots than the schedule shows and supply is actually greater than the sum of appointment slots on the original template. Initially I would count supply as the sum of the appointments on the basic template and then as the intention to "merge down" is consistently employed then use the functional supply number which is greater than the template. Supply is then measured as being greater than the number of appointment slots on the basic original template due to the consistent intention of merging down.

Thus, due to the discrepancy between the original templates, for some groups the supply is greater than or less than the template. Having this discrepancy demonstrates that a practice is flexible and is actively willing to merge up, for example. This prevents double appointments when more time is needed. On the other hand, consistent, predictable discrepancies may indicate lack of truth in scheduling.

Activity

Activity (also called "supply used") measures what was done - the amount of time in minutes that the provider actually spent seeing patients (in the cumulative "red zones") or, in a more primitive way, the number of appointments seen within a specified time frame. This measure is retrospective. This can be manually counted or counted by the computer. This is the simplest of all the measures. This measure is sometimes confused with demand as in "the demand is what we did." Activity can be greater or less than either supply or demand. While over-booked demand is measured as activity, no shows and cancellations are not activity. While supply can be viewed as "planned work," activity is "actual work."

Activity can be measured as:

- Number of "checked out" appointments
- Number of "checked out" appointments X average red zone length. This adjust for short and long appointments
- Direct observation of time

Measurement for Improvement vs. Measurement for Judgment

Measurement for improvement

- on the ground, at the practice level
- often done manually
- looks at baseline and effects of change over time by re-measuring
- local customized decisions: day, time, appointment type, etc.

Measurement for judgment

- provides for comparison
- often automated
- requires standardization
- may require weighted averages rather than average of averages
- will become more rigid: same day, same time, same appointment type, decisions about weekends
- dependent on the standards - some groups will "look better than they are" and some will "look worse than they are"
- will result in some groups changing measures or methods in order to get the desired results

Decisions

- standards and standardization: days, times, appointment types
- what appointment type to measure
- whether to include weekends
- how to assess the level of impact (bigger practice, more impact on system waits)

What to measure

- average of third next available appointments (TNAs)
- average of percent improvements
- weighted average-impact: improvement in days of wait
- what to measure once the delay goes to zero

As we move from manual, on the ground, individual practice measurements toward system-wide, automated measures, we slip quietly into measurement for judgment. The break point between measurement for improvement and measurement for judgment occurs when others can view the local data and can have opinions about it. As we move toward automated system-wide measures there is greater value in the individual measures and particularly in the relationships between the various metrics. The ratios of demand to supply to activity and to delay (third next available appointment), the relationship between panel size and demand, panel size and visit return rates, and/or panel or visit return rates and delays, all become available and comparable at a much higher order of magnitude. The automated system-wide measures have much greater utility.

Many practices want to predict future demand and plan for future deployment of supply. They want to know "How many future open slots do I need?" The way this question is posed is critical. Once we determine the answer to the question, the real issue is how we act on the answer. Once we know the number of slots we need, we could "get" those slots by a carve-out ("setting aside" that number of slots). However, this dynamic creates two rigid demand streams: appointments open and appointments restricted to meet the daily slot quotient. On the other hand, and this is very nuanced, the "correct" number of slots could be attained not by putting restrictions on slots (carve-out) but by smart scheduling, constantly looking at the pattern of appointments for an entire day (most computer scheduling packages show only the three "next available appointments" in the date range requested), by "selling" appointments that are early in the day and later in the week, by appointing less returns on days when there are more providers absent or when the practice is busier, and by setting hints or thresholds on future scheduling. The future appointment pre-book thresholds serve as guidelines and are not rigid streams.

To answer the "How many slots do I need" question, many practices would like to rely on automated measures and measures with "bigger numbers," due to a perception that manual, single practice numbers are subject to too much variation in collection, in interpretation and due to the smaller "n." These practices then turn to other methods to answer the question of how to predict future appointment demand and needs. From the box analogy it is clear that "demand" is not a single entity or stream. Some demand arises same day and some demand gets appointed in the future. Some demand is out of our control and some demand is within our control or influence. Some demand is pre-booked and some demand is not pre-booked (booked today).

Demand dichotomies can be described as external and internal, pre-booked and booked same day, natural and artificial variation, and new and return.

External/Internal

External demand is the demand component that describes appointment workload generated from outside the practice, measured as workload generated ON a specific day, regardless of the day FOR the appointment.

Internal demand describes appointment workload generated as a return appointment made today as a result of or following an appointment today. This component includes some but not all "return appointments."

The sum of external and internal is the total demand.

Pre-Booked/Booked Same Day

This measure captures how much of the appointment workload is pre-booked and how much is made on the same day as the appointment request. This is, in a sense, the future open capacity. This could be measured by the month or even by the day. Booked same day includes some but not all external demand. Pre-booked includes all internal demand plus some external demand.

In some practices, many of the provider driven return appointments are pre-booked as internal demand (appointment made as patient leaves) while in other practices many of the provider driven return appointments are pre-booked but are not internal demand (patient calls back and gets a future appointment) or are not pre-booked (patient calls back and gets same day appointment). I view this as external demand while others want to see this as internal demand (see measurement interpretation comments above).

The sum of booked today and pre-booked (booked for the future) is the total demand.

Natural/Artificial Variation

This measure captures the proportion of appointment workload that occurs naturally from the environment and the proportion that grows from artificial direction from the practice. Return appointments are a form of artificial variation. All returns (defined as appointments driven by providers) are artificial. But returns are not synonymous with either pre-booked (some returns are booked same day) or with internal demand (some returns are external demand - those who choose to call back later)

Natural and artificial variation is not exactly synonymous with the terms common and special cause variation. Common and special cause refer to the stability of a process and primarily identify the degree of variation in the stable process as different from the special causes of variation – or the 3 SD from the mean (there are about eight "rules" that identify other patterns of special cause variation, but the 3 SD from the mean is the most common one).

Natural and artificial variation is a view that looks at the origins of the variation in the first place – defining the origins as natural (those causes of variation that come from the natural illness of the population e.g. emergencies) or artificial (those causes of variation that we cause or influence).

New/Return

A fourth demand dichotomy divides demand into new and return. This division is driven primarily for billing and coding purposes. At the same time, this division is very useful in specialty care practices since appointment types are often defined as new and return. The sum of new plus return equals total demand.

While all new appointment demand is external demand and natural variation, these terms are not synonymous. New patients are often pre-booked but not always. While all returns are synonymous with artificial variation, only some returns are internal demand or pre-booked.

The descriptions and dichotomies do not always overlap. While the pre-booked and booked descriptions and dichotomies do overlap, demand arising same day/demand arising for the future does not overlap with natural/artificial variation or with external/internal demand. Out of control and within control is close to external/internal but is not synonymous with it. Out of control and within control is close to being synonymous with natural/artificial variation, but, as noted above, is extremely difficult to capture in measurement (certainly manually) due to variable interpretation of the term return, and in an automated fashion since it is difficult to program intention (the computer would have to know that a patient was calling back for a provider driven return appointment).

How to Determine How Many Appointments are Needed

After reviewing the descriptions and demand dichotomies described above, and trying to reconcile them with measurement possibilities, particularly automated measures, one could reach the following conclusions:

- The external/internal dichotomy has great utility on the ground and moderate utility as an automated measure. If a practice stays with the classic definition of internal demand as appointments made as patients leave, this can be successfully measured but if the classic definition strays into counting internal as all returns (even those who call back later for an appointment), measurement will be extremely difficult.
- The natural/artificial variation dichotomy has potential utility but, as above, if artificial variation is synonymous with all provider driven returns, either manually or automated, the measurement will be problematic
- The new/return dichotomy has great utility in specialty care practices and, because these demand streams are, in most specialty care practices, distinct appointment types, measurement is easily accomplished. There is not much utility for measuring new/return in primary care practices.
- The booked same day/pre-booked dichotomy, while not synonymous with in control and not in control descriptions, and not synonymous with natural/artificial or external/internal dichotomies, may have the greatest utility in primary care and would be easy to measure. However, practice behaviors will greatly influence the proportion of booked to pre-booked, and the measurements. Using this dichotomy would entail introducing a new measure to the practice. The pre-booked and booked today dichotomy is not the same as internal and external and even trying to

interpret and measure all "returns" as internal with the adherent complications of that decision, will not make these two concepts synonymous.

How to Get Control - More or Less Pre-Booked

The answer to the question: "How many appointments do I need?" is dependent on ease of measurement and consistency of practice behaviors, and is overlaid with the control issue. We want a measurable determination of this number, a consistent and predictable number and a number that we can greatly influence or control. In order to get consistent, predictable metrics we need standardized practice behaviors and workflow. The issue really is not external/internal or natural/artificial but just how much is pre-booked and how much is not pre-booked. There is a belief that the best way to get control and have predictable demand is to not make any future appointments (internal demand is zero). This leaves the future open. This behavior gives the appearance of control but does not work. Second, there is a persistent belief that the best way to get consistent and predictable workload is to have return patients call back and then appoint them even further into the future.

In my experience, standardized behaviors lead to the most predictable demand outcomes:

1. Appoint as many returns as possible on the day they leave, and

2. Pull as much of the other work into today as possible (convert as much work as possible into external same-day demand).

Once a practice can consistently adopt these standardized behaviors, the historical pattern of pre-booked to booked can be used to answer the question of how many appointments are needed. The action then is not to carve out but to use the strategies noted above to ensure schedule capacity.

The Effect of No-Shows and Cancellations

No-shows and cancellations have an effect on demand and activity: they "count" as demand but do not materialize as activity. I see these as a defect. If we try to manipulate the demand measures to subtract no-shows and cancellations, interpretation, implementation and consequently measurement are very difficult. I would rather focus on eliminating the defects in our systems that cause no-shows and cancellations and, at the same time, recognize that demand is diluted by no-shows (a late cancellation also functions as a no-show).