MMA
Mark Murray & Associates

Measurement of Delays into Specialty Care

In Specialty Care (SC), the office workload/demand competes with other demand streams: the "on-call" function, operating room or specialty specific procedures, the ambulatory surgery center, and any other distinct scheduled or unscheduled workload. These "streams" are supported by the same individual specialists. The extent of the supply allocated to each stream dilutes the amount of supply allocated to the office. These streams can be "seen" as steps that patients may take on their horizontal flow through the SC practice: from new patient in office to office return, to Emergency Department (met by the on-call function), to the OR and back to the office.

Delay for a patient office appointment, measured as third next available appointment (TNA), is distinct from demand measures. The delay represents the temporal relationship between demand and supply: how long does it take demand to meet supply. The initial focus of our work in "access to care" is a focus on the office demand stream.

Delay into the office is measured by appointment type. SC practices create unique and special value for both patients and referring providers in providing access to new patients, and consequently office demand can be divided into two distinct queues: new and return. We calculate the wait time, expressed in TNA, for each of these specific appointment types. While delays into either queue are important, the delay for new patients is the most important. The wait in that queue reflects the value to the customers – both patients and primary care referring providers.

Within the office demand stream itself, new appointments compete with return appointments. Since patients generally average more than one return visit, the “new” appointment queue (appointment type) needs to be distinguished from the “return” appointment type. If these queues were merged, the “new” would get squeezed out by the “returns” because there are proportionally more return visits. So, new appointments need their own appointment type and queue. Each appointment type and queue, including both new and return appointment types, has its own demand, supply, activity and delay dynamic, and hence, needs to be measured independently.

The term "new" refers to any initial visit whether this is labeled as “new consult” or “new patient” (self-generated). The vast majority of SC groups do not have distinct appointment types for these sub-sets of new work and have merged both these types of “new” appointments.

Groups should focus measures on the delay for new. If the caseload is correct and the delay for new is within goal, in most cases the delay for return will be acceptable.

How to Measure TNA in SC

1. Measure each individual provider’s TNA for new patients.

2. Measure the average of these individual TNAs.

3. Measure the TNA for new for “any” provider. This is a measure of the first available new patient appointment for any provider within the practice. This is the critical measure in a SC practice since this measures what the customers want and experience. We are measuring the capacity of the SC practice to receive any new patient within the desired time threshold. In order to achieve a short delay for patients and to be able to reflect this achievement in measurement, the system has to adjust/change old practice into new practice – and this is the new evolution of SC work that we bring here – ideas on how a system has to adjust.

The measurement is not difficult. It takes about two minutes per doctor per week (#1 above) and then two minutes to calculate #2 and #3 above. With six doctors, the measurement would take a maximum of 16 minutes per week, and it can be automated.

Pooling of Referrals

If demand enters the practice earmarked or referred to a specific provider, that is, a distribution based on popularity, we get variable delays based on a demand-supply mismatch, or scheduling issues (doctor gone but work still scheduled). Demand has to be balanced with supply, even at the individual provider level. Demand for appointments can be calculated by the equation:

$$\left\{ \text{Caseload} \times \text{Expected Patient Visits} \right\}$$

Caseload, along with the expected work that the caseload brings, has to match with individual provider office appointment capacity:

$$\left\{ \text{Days Worked in the Office} \times \text{Appointments Available per Day} \right\}$$

Giving a doctor more work than he/she can manage is a recipe for failure. When work is referred from primary care to individual/specific doctors in specialty care, without regard for demand-supply balance, that is without regard for current caseload, some doctors get more work than others and are consequently over-demanded (the “popularity” issue). A permanent mismatch (“over demand”) can initially be managed by a backlog but as the backlog increases, practice performance deteriorates. Pooling of referrals (sending referrals to the practice, rather than a specific provider) based on caseload and capacity is critical in preventing permanent demand-supply mismatches and delays.

Distribution of New Patient Appointments

While popularity drives a continuous demand-supply mismatch, continuing to send work to providers who are “absent” creates a temporary mismatch of demand and supply and a temporary delay. This is the office supply variation issue.

Because a specialty care provider’s presence in the office is “diluted” by other indispensable duties, the provider’s office presence, especially with vacations and other time off, can be very sporadic. When a specialty care provider is absent from the office for whatever reason, and the office demand for new patients continues unabated, we intentionally create a mismatch of demand to supply (continuing demand against absent supply). Consequently a delay occurs due to competition for overall provider capacity from the other competing demand streams. For example, before a provider leaves for vacation he/she will often devote a disproportionate amount of time (capacity) to managing critical demand streams like the operating room, and

neglect spending time in the office. His/her intent is to eliminate any backlog accumulation in the operating room while he/she is absent. When he/she returns, he/she often devotes a disproportionate amount of time to “making up for lost call” and again neglects the office. New patient demand continues to come into the office, and an inevitable delay in third next available appointment ensues. Pooling of referrals allows a re-distribution of new patient workload prior to, during and post time out of office, in order to prevent extended delays for new patients to that provider.

In addition, if a practice takes a five-day goal for new patient delay seriously, then it needs to look at the individual provider and the practice capacity for new patients. If demand for new appointments is measured and is predictable, we know that to keep up with demand we need a specific number of new patient appointment slots within the next five days. Therefore, on days when there are more providers present, the ratio of new to return appointments can be lower than on days when more providers are absent. The number of “new” slots is constant. On days when there are more providers, these new slots are shared more widely. On days with higher absences, these new slots are shared amongst fewer providers and the ratio of new to return rises.

Effects of Long Delays into Specialty Care

With either the permanent demand-supply mismatch due to “popularity” or the temporary demand-supply mismatch created by sending work to an absent provider, delays will be created. Long delays into specialty care can cause problems:

- There is an “unfairness” built into the system. Patients referred to providers with long lines wait longer. Practices will often then try to bargain with patients (popularity vs. delays).

- Patients who have to wait for an extended period may fail to keep the appointment (no-show) or may even expire.

- Patients who have to wait for an extended period probe the perimeter, trying to find a way in. They will call the specialty care office and be told if their primary care doctor calls the specialty care doctor, perhaps they might get in sooner. This generates more work and turns doctors into appointment makers.

- Patients who have to wait for an extended period will intentionally or unintentionally cut in line (e.g. by going to ER) in order to get seen sooner by the specialist on call, or even get admitted in order to see the specialist.

We can address these mismatches that lead to delays, and avert the negative effects of long delays. If we monitor caseload in relation to office capacity (ensure # of new patients × expected return visits does not exceed the office appointment capacity), we can balance the workload and overcome the risks of “popularity” of a particular doctor. To do this though, we have to "pool" the referrals. We have to refer to the "practice", not to the individual provider. Sending work to a provider with more work than he/she can handle is not smart nor a good thing. "Over-popularity" does not help the patients. There is a limit to capacity. Popularity referrals create specific queues. Pooling referrals allows us to shift from a series of single queues to a model where we can use data to intentionally balance workloads. Relying on the primary care market to balance workloads will simply not work.

“First” Third Next Available Appointment

When referrals are pooled, caseloads can be managed and monitored in relation to office capacity, and in light of absences. Pooled referrals allow us to move new patient work to providers who are present in the office. Pooling and re-distribution changes require more sophisticated delay measures. While it is important to continue to monitor the standard measures of delay (third next available appointment for each appointment type – new and return) as well as the department or practice average of third next available appointment, we also need to monitor the **first available third next available appointment.** This is a more sophisticated measure of delay for new patients into a practice or group of specialty care providers. This measure looks for the first available of the third next available appointments for all providers (to determine which provider has the shortest delay for third next available appointment). With this measure, in conjunction with pooled referrals and with re-distribution of new patients during absences, we can view overall practice system function and performance. Can this practice deliver an appointment within the time frame set by the goal? Can a new patient get an appointment with any one of the providers, within the time frame set by the goal? Because of the dilution effect of other demand streams and because of vacations and time off, specialty care providers commonly will simply not be able to individually maintain short delays for new patients. They are not in the office enough. Both the number of new patient slots and the potential for office backlog are heavily influenced by their time in the office. Pooling referrals for equity and looking at the first available third next available appointment overcomes the operational challenges. The first available third next available appointment provides the best measure of the department or practice capacity to meet the needs of new patients and for the department or practice to achieve a goal of delays no longer than five days.

All in all, we need to have measures that allow us to see overall system/practice performance, in order to improve performance. The main service provided by a specialty care practice is seeing new patients. Optimum system performance requires seeing these patients without a delay. We do not know how well we are doing unless we measure it. If we do not measure it, quite frankly we are working blind. It may mean that we have to change the way we monitor, but once the change in monitoring occurs, even manual measures take a minimal amount of time and deliver a great deal of essential value. In addition, over time, this measurement can be automated.

Considerations for Measurement

1. Because the value in SC is clearly the value to see new patients, measurement of delays for new patients is the most critical measure. Within the workload entering the outpatient office setting, new patient demand competes with demand for return appointments

2. If the return visit rate within SC practices is 2.0 or greater (indicating that for every new patient there is at least one return visit), unless new patient appointments are explicitly separated from return patient appointments as a distinct appointment type, variation in demand for new and return visits will inevitably lead to circumstances where the return visits out-compete the new visits on the schedule, artificially and randomly increasing the delay for new patients. Thus, in order to eliminate randomness causing delays and to ensure that appointment systems can be freed from that randomness, new appointment types need to be distinct from return appointment types. This is technically a carve-out, since one is carved out against the other.

3. SC practices are commonly characterized by competing demand streams. Workload enters the practice in distinct types of work: the on-call function which is fashioned to manage "immediate" work and concerns, workload in the Operating Room, work in the ASC, and work done in the hospital inpatient setting. All these streams compete for the same provider capacity (supply) as the outpatient work.

4. Not only does direct patient care compete for provider time, but non-direct patient care work (as well as other time out of office commitments and choices) also competes for provider time. There are priorities within this competition: within the direct patient care work, for example, the on-call and hospital functions are always supported at 100% while other streams get less support. Within the direct patient care arena alone, if 50% of the provider staff is absent, the on-call function is supported at 100%, and the office is supported at far less than 50% and the other functions are supported somewhere less than 100% but certainly above what the office "gets." Thus, the office outpatient supply varies tremendously. As a consequence, the delay for new patients, which is linked to office presence, can vary tremendously as well. This makes measurement and compliance with goals on an individual basis extremely difficult. There is a tendency then to loosen the goals or measurements, to extend the goal for a new patient delay to 14 days or even longer. This extension of goal is done blindly without truly taking into account what is dynamically possible within each practice. This extension of goal is often done because "five days just sounds too short" and 14 days sounds doable.

Measurement and goals for measurement, whether for the practice or for the individual, must be seen in the basic dynamic context. No system can perform optimally (or function at all) if demand within in any demand stream exceeds corresponding supply. A practice can either work forward from known demand or work from known supply to create a balance between demand and supply. A SC practice ought to measure demand as daily demand, supply and activity, as "caseload" (number of unique unduplicated patients), and as visit rate. The product of visit rate times patients, or the sum of new and return (from DSA), needs to balance with supply (see the panel or caseload equation). Outpatient office supply is only one competing function of the providers. Supply can be measured as total supply which includes all direct and indirect patient care work, then within the direct work, the amount of supply allocated to other functions that compete with outpatient can be calculated and what is left over is the office supply. This can be viewed as days worked in the office times the number of appointments per day. The number of appointments per day is a blend of new and return. Within the office-outpatient, the demand simply must be balanced with supply. Without that balance delays will inevitably increase and no sleight of hand can change that. If demand is balanced with supply, the delay goal can be set at a pre-determined threshold or level. It also becomes clear that once the balance is achieved, any new patent workload must be balanced by "graduation" of a corresponding amount of "old" patient work.

Required System Adjustments

There are system adjustments required to achieve a goal of a short delay for any new patient into the system of care where the SC practice is complex – with high dilution of office time due to other competing indispensable duties deep inside the system.

a. Pool the referrals. If referrals are sent to individual providers (a popularity or even preference referral system), delays will result simply due to office supply variation. That individual provider, as well as the others in the practice, is put at risk for a demand/supply mismatch due to over or under workload issues.

b. Measure the new patient demand, volume and variation. As noted above, no system can perform with a mismatch.

c. When creating the future schedule, ensure there is enough new patient appointment capacity to keep up with the predicted demand for new patients. Because the goal for new patient delay into SC, for the most part, is greater than a day, there is a window or buffer for new patient delay expectations. The practice has to ensure there is enough new patient capacity within that window. If the providers have a low degree of activities that dilute office presence, look to assuring enough new patient capacity by individual provider. However, in many SC practices, because of supply variation (due to time out of office either because of patient care or not), it is difficult to ensure that each individual provider will have enough new patient capacity each week to keep up with his/her proportion of the new patient demand, hence "new patient capacity" has to be viewed as *practice* new patient capacity, not individual provider.

d. If provider capacity for new patient demand is scarce, then the practice can adjust the ratio of new patient appointments to return patient appointments on that schedule in order to meet predicted demand for the new patients. Most scheduling systems have a rigid non-changing ratio of new patient capacity to return patient capacity. Sometimes these ratios are determined by the ratio of new patient "clinic" to return patient "clinic" or at other times the ratio is determined by appointment types. The amount of differential time allotted for new patients (often twice as long as for a return patient) has an effect on this ratio. A ratio of new patient "clinic" to return patient "clinic" may appear to be 1:1 if there is an equal number scheduled, but if the new patient visit is twice as long, the ratio is really ½:1.

e. Since demand and supply balance is the critical overlying dynamic and an ongoing equitable distribution of new patient workload based on proportionate time in office (measured by caseload limits), practices need to develop a retrospective method to divide new patient work equitably. Dividing workload is always a challenge:
    - Some groups divide new workload by popularity (very poor choice since this over-panels some providers, leading to serious system consequences).
    - Others divide work by preference (almost as poor a choice).
    - Some divide the work by looking at the next open new patient appointment slot which creates an inequitable distribution of new patients based on the variances in return rates (the providers with the highest return rates get fewer new patients)
    - Some divide the work by "exposure-risk" (new patients enter through the on-call function which is shared, and as a consequence, risk of new patients is shared)
    - 'Some divide the work by scheduling (all providers have new patients scheduled on the template so new patient work is distributed in the proportion of the new patient distribution on the schedule). This latter approach, while seemingly "fair", virtually guarantees delays for many new patients and those delays are random. Since demand and supply, measured by caseload limits and determined by the product of patients times expected visits, is critical, equitable distribution is desired in most practices but prospective methods of equitable distribution will lead to delays.
    - The best method to distribute work is to distribute the work inequitably in the short run, based on scheduling of new patients as suggested above, and then retrospectively, monitor and track that distribution. We suggest a retrospective tracking each month with a view of caseload, caseload limit, return rates, and new patients based on proportion of time in office. Keep in mind that if the caseload limit is achieved, for every new patient accepted, there must be a patient "graduated" back to his/her primary care provider.

In a SC environment where demand = supply, these system adjustments can assure that new patients can get an appointment within the timeframe threshold. This is the key issue here – that new patients are seen within the goal threshold. At the same time, measurements for individual provider's TNA and the average of the individual provider's TNA will vary: some individuals may not be in the office due to support of other duties, or due to vacation time. As a consequence, their individual TNA will not only vary, but dependent on the level of competition with other practice demand streams, will inevitably extend past the goal threshold. A system designed to refer patients to SC based on popularity, without any attempt at measurement of caseload and caseload limit, guarantees non-compliance with a delay threshold unless that delay goal is so long as to be meaningless, dissatisfying and risky. Similarly, systems designed to measure individual TNA without taking into account the effects of supply variation due to competing demand streams, will also guarantee non-compliance with the measure. If the measure carries any incentive or judgment, that system is at risk for gaming, for constant complaining, or for deterioration in morale. Individuals then, fraught with the task of compliance, will attempt to find individual solutions, workarounds and exceptions. In SC practices where there are minimal competing demand streams (probably like dermatology where there is minimal on-call function, no OR, minimal hospital, etc.), measurement of TNA for individual providers, for average, and for any, will closely coincide. For more complex systems, with OR, ASC, on call, and/or hospital all competing for allocated supply time, TNA for any, for individual and average will not coincide. The best measure here is TNA for any provider with any judgment or incentive based on the practice. Thus, we suggest making the system adjustments, measuring the individuals and the average, and in particular the measure of TNA for "any" and then "reward" the practice, not the individual providers. This is a hive, not individual bees.