MMA
Mark Murray & Associates

Specialty Care System Performance Measures

The basic measures to gauge and assess specialty care system performance include measures of delay (TNA - third next available appointment), demand/supply/activity (DSA), and caseload. There is a tension in specialty care between using these measures for individual providers within a practice, and using the measures for the practice as a whole. While initial measurement effort focuses on looking at individual provider performance and the variations amongst the providers, at some point we need to get both measures of individual and practice performance. From the specialty care customer perspective (both patients and referring clinicians) the most important measure for performance is the TNA for new patients.

Individual Measures

When we assess performance of individual providers, we commonly see variation in delay and caseload. Some but not all of this variation can be accounted for by variation in clinical FTE (supply, or time in the office). Sporadic presence in the office will result in delays and may result in variable or disproportionate caseload size if caseloads are not assigned in proportion to FTE status. At the same time, the variation in performance can also be attributed to individual behaviors such as:

- preferred visit length (which affects visits per day)
- acceptance of new patient work
- current caseload size (particularly in proportion to the amount of time worked in the office)
- number of appointment slots for new patients
- return visit rate

or to system behaviors such as:

- distribution of new patient workload distributed based on popularity or availability
- an appointment schedule that does not distinguish new patient appointments from return patient appointments

Some specialty care practices choose variable performance by appointing new patients by "popularity." While this approach appeals to individual provider ego, it creates intentional delays for the patients of the popular providers due to an eventual imbalance of workload to supply for that popular provider, and in time, leads to bargaining at the margins ("You can see Doctor Popular but you have to wait three months or you can see Doctor Unpopular and wait three days."). The popular provider is blind to these uncomfortable negotiations. While this behavior appears to be an informed wait, in reality it is a blind wait and patients can be harmed. Despite the fact that some providers are more skilled or have a particular expertise within the overall practice, the practice behavior of distributing appointments by popularity does not enhance their expertise. Providers have a capacity limit that can be determined by an equation. Making patients wait because of popularity does not change this capacity limit. It just pushes work out to an "acceptability limit" - the point in the delay for the popular provider where patients or referring providers decide that the wait is unacceptable. Workload then moves off the end of that line. In queuing theory this is described as balking or

reneging. Patients move away from the end of the line randomly, without intention, plan or purpose. As a consequence the stated intention of the popular provider - “patients get a choice to see me and I have a special skill that is worth waiting for” becomes ridiculous. The “I am popular” issue becomes a myth and is dangerous for patients. Popularity has a time limit and working to that time limit (the acceptability limit) is harmful. Practices that operate in this way create a double layer of “unfairness.” Patients queued up for the popular provider wait longer (unfair to the patients) and popular providers are put in a position where they cannot be successful (unfair to these providers).

In addition, to accommodate patient clinical needs, many specialty care practices, burdened by either variation in individual or practice level behaviors, will accommodate these variable behaviors by appointing new patient work to the first available appointment slot, in order to get patients the soonest appointment. While this accommodation is an attempt to create “fairness” for patients (the soonest appointment), the result is in an uneven distribution of new patient workload and an “unfairness” to providers. Further complicating this scenario is the fact that many of these practices do not distinguish new appointments from returns. Individual providers will commonly recognize the “unfairness” of provider workload distribution and embark upon self-protection behaviors. Acceptance of new patients thus can occur as a result of individual provider choice. By pre-filling the schedule with return visits or finding other self-protection tactics to become unavailable, a provider can choose not to accept new patients. These behaviors worsen system performance, increase patient delays and pass the “unfairness” back on to patients. Long delays and inequitable workload distribution are perpetuated.

Acceptance of new patients based either on popularity or on availability will always result in an acceptance level of new patient work disproportionate to clinical FTE (work effort). There are two related issues that grow from this acceptance imbalance. First, a provider with too much work simply cannot do the work and a provider with too little work is not working to capacity. Second, this imbalance creates an “unfair” distribution of workload. The “unfairness” issue is obvious in salaried environments due to the obvious disproportionate amount of new patient workload for the same salary. In fee for service systems, while the new patient workload does not have to be directly and equitably balanced against the same fixed salary, there is a loose association or correlation between new patient workload and time in the office. While the reimbursement is not fixed as it is in salaried environments, there is still a limit or ceiling as well as a floor for new patient workload. Having workload above the ceiling or below the floor creates unfairness in the opportunity to be successful. Those limits (either upper or lower limits on caseload size), while indirectly related to reimbursement, are directly related to capability and to time in the office. This limit is defined in the caseload equation:

Caseload X # of visits per patient = Days worked in the office X # of visits per day

A caseload over the limit set by this equation means that the provider cannot do the work (the delay and backlog does not help or change this) and a caseload under the limit means that there is unused provider capacity.

As a consequence of this multitude of variables, particularly distribution of new patient work based on popularity or availability, measurement of system performance is clouded and improvement in performance is difficult. Practices tend to attempt improvement by moving back and forth along the “fairness” continuum. They recognize that acceptance of new patients based on popularity creates unfairness for both patients and providers, and move towards an acceptance based on availability which is an attempt to solve the fairness issue for patients but will often just perpetuate the unfairness for providers. We need to solve the fairness for both. This can only occur if we move towards a practice improvement perspective accompanied by practice level measurements.

Practice Measures

While it is critical to standardize both individual and system performance in order to adequately assess that performance and to guide improvement, it is a challenge to determine the best place to start this process. Beyond individual or system behaviors, workload balance, that is, demand (patient load X expected visits) matched by capacity/supply (days worked in office X visits per day) is critical for successful performance. Without this balance, providers will fail. The key to unlock poor performance and to move towards improved total system performance must start with a focus on workload balance and distribution. Thus, in order for either individuals or practices to be successful, not only do the caseloads have to be balanced and distributed correctly, but the variables in behaviors must be addressed as well.

These are the steps that we need to take:

1. Make sure that the caseload is manageable. Use the above equation.
2. Make a distinction between new and return patient appointments. Because of the unique value created in specialty care practices in seeing new patients, a precondition for improved performance in specialty care requires a distinction between new and return patient appointments. Without this distinction it is impossible to measure delays (TNA) for new patients. In addition, this distinction creates the template for an equitable distribution of new patients based on clinical FTE. For many specialty care practices, the most apparent variation in performance and behavior is the imbalance in acceptance of new patient appointment work.
3. Address the distribution issue. Eliminate the distribution of new patient workload based on "popularity" or distribution to the first available appointment. Distribute new patient workload based on proportionate time in the office - the clinical FTE. This method is commonly called "pooling." Set the ratio of new patient appointments to return patient appointments on the schedule in such a way that the new patient work is distributed in proportion to FTE. For example, if a .6 FTE gets 10 new patients per week, a .3 FTE should get five new patients over the same time frame. The ratio is not random. First, the ratio needs to reconcile with the caseload equation and provider capacity limit. We cannot keep giving new patients to providers without understanding their ultimate capacity limit (determined by the equation). Once in balance, new patients have to be balanced by "graduates." Second, the ratio needs to reflect the patient visit rate and visit length. If the expected visit rate is four visits per patient per year, then assuming the visit length for new and return are the same, for each new patient slot we need three return slots. If the visit length is different we need to adjust accordingly. The change in distribution method addresses the visible fairness issue for the providers. However, while we have now made the provider distribution "fair," we have created an unfair delay for patients. Office supply in specialty care practices is commonly very sporadic due to other indispensable duties within the overall system (e.g. "on-call" function, operating room, procedure, hospital etc.) as well as time off and vacation. Some of the demand streams, the hospital and the "on-call function, for example, need to be supported at 100% while the bulk of the variation comes out of the "less important" functions, like the office. As a consequence of competition with the other "more important" demand streams and in light of time off and vacation, the office presence is highly variable. When new patient workload is distributed "fairly" based on FTE and not on popularity or on availability, this distribution is commonly done in sequence, that is, patients are distributed one at a time to the next provider in line. While this method of distribution seems to solve the fairness issue for the providers, due to the sporadic supply, it creates even more unfairness for the patients. For example if Doctor A and Doctor B

both have the same FTE status we distribute equal amounts of new patient work to them. But if A is in the office this week and B is not in the office for 3 months due to duties within the other demand streams and due to vacation, the first patient gets a short delay and the second gets a long delay. Patient delay is based on random luck. We need to ensure we consider the fairness issue for patients as well.

4. Minimize the backlogs for all providers. This avoids some but not all of the “bad luck” delays. If new patient workload is distributed in sequence, despite the elimination of all backlogs, vacation and time off will still create disproportionate delays.
5. Once the backlog is reduced, change the distribution method. Because of sporadic provider presence, distributing new appointment work in individual provider sequence (one provider at a time) will result in delay. Change the distribution to distribution in schedule sequence, not in individual sequence. With a minimal backlog, all providers who are available to work will have appointment availability within the goal threshold (five days) but all providers will not be available to work within that time frame due to other obligations. If work is distributed in individual sequence, some patients will wait longer than others, dependent on when the provider is next scheduled to work. To resolve this, workload needs to be distributed in a bolus - distributed in a bundle of more than one patient onto the provider who is next available to work, filling all of her new patient appointment slots. This changes the distribution from individual provider sequence to schedule sequence. So if Dr. A and Dr. B have caseloads that are equitable, have no backlog for new patients, and we want to distribute the work in a fair way to the providers - that is, in proportion to their FTE - and there is sporadic office presence, the work may go to A in a bolus of work and then to B in a bolus of work at a later time. The schedule template (ratio of new to return) will guarantee an equitable distribution in accordance with FTE. At any one time, there may be an inequitable distribution, but within a large window there will be equity. In practices with a high degree of supply variation, we cannot expect to distribute the appropriate new patient workload to each provider each week. Providers are just not present and would “miss their turn.” By making these changes, we can reconcile the fairness issues for both patients and providers. So providers can get an equitable, fair distribution of new patient work based on their FTE status, and patients are also treated fairly because they don’t have to wait.
6. In order to guarantee no delays for patients, practices must measure and predict the need for new patient capacity within each five-day window and organize and arrange the schedule amongst the competing demand streams to ensure that the right number of slots are available. This may require temporarily changing the ratio of new to return in advance or temporarily flexing supply to add new patient capacity.

At this point, both individual and system performance can be seen, assessed and measured and the primary measure for system performance shifts from third next available appointment (TNA) for an individual provider, to TNA any provider.

Discussion Between Mark Murray and Mike Davies

Mike:

If Specialist #1 has 10 new slots per week and works 50 weeks per year and doesn't send any of his patients back to their primary care physician, he sees 500 new patients per year and has a caseload of 500 in year 1, 1000 in year 2, 1500 in year 3 and so on. If Specialist # 2 has 10 new slots per week and works 50 weeks per year and graduates one (old) patient for each new patient he/she sees, the caseload is 500 in year 1, 500 in year 2, 500 in year 3 and so on. Total caseload for specialist #1 grows, octane goes down, and he/she can keep up only if clinic hours are expanded to see patients for return appointments.

Mark:

True. At the same time there is a provider capacity limit. This limit is set by the equation. In this example, one provider has “broken” the equation by not discharging any patients. This provider’s performance will fail and his/her patients will be forced to wait. The new patients will continue to arrive but the returns will wait longer and longer to the extent where harm occurs. These “systems” have to be monitored.

Mike:

If the immutable "rule" (or reference point) is that "you have to see 10 new patients per week" and if that "new patient rule" is proportional to the FTE, then the only thing that can "give" is the pressure to eventually discharge existing patients.

Mark:

Yes, the pressure is to discharge but the rule is not to see a set number of patients per week or per time frame. The rule/goal is “no waits.” The practice may have to adjust supply as demand varies. If demand goes up the practice has to adjust by temporarily changing the ratio of new to return or adding supply. If demand goes down, since there is a five-day window, the TNA will slide down and the group practice can change ratios in the reverse. Since there is an early warning system (the five-day wait) we can see if demand rises, that is appointments get pushed out to or past the five-day limit. The practice has to adjust by changing the ratio over the next few days or adding more supply in order not to go past the five-day limit. If demand rises quickly, the practice has to add supply because the return slots are already filled, but if the supply schedule is down (as in summer), we can see this in advance and change the ratio far in advance.

The number of new appointments will flex due to natural and artificial variation, but the rule is “no waits.” At the same time, caseloads can be monitored for proportion to FTE (this is only an issue if demand goes up or down to the extent that there is an effect on the ratio and the providers see more or less than expected new patients). If demand rises and the ratio changes, some providers will get a temporary increase in caseload greater than the basic schedule template. If demand falls, some may temporarily get less if their ratio changes. This can then be monitored and reconciled – perhaps each quarter. We can also monitor the return rate each month and the “discharges”’ each month. To keep a steady state the discharges have to equal the number of new patients. If they do not, the accumulated caseload will increase and since new patients come in at a predictable rate and this is guaranteed by

the schedule and the commitment to adjust as needed, the variation is pushed onto the return visit line.

Mike:

So the real concept is NOT to use caseload number as the reference point, but to figure out the FTE proportional to new patient slots, and use that number of new patient slots as the reference (or immovable) point and use caseload just as a measure of the outcome of using new patient slot numbers as the reference point. That whole thing works for me. So we are teaching that for specialty - number of new patient slots is the immovable reference point.

Mark:

No. The immovable reference point is “no waits”. For each individual provider, at any one time, the caseload may vary due to the rate of discharges and the return visit rate may vary. The number of new patients may flex due to natural variation. The only way to make the new patients immutable is to either have a wait and work out of the “warehouse” of backlogged patients (which we do not want to do), or to restrict entry into the practice which we may have to do. This is occurring anyway if the demand (caseload) exceeds the capacity. The risk is that if delay is immutable and new patients continue at a rate that exceeds discharges, the delay for returns rises. We can keep an eye on and prevent this by monitoring individual provider return rate, delays for return appointments, and caseload. There is an end point though. If the caseload equation shows an imbalance, demand simply has to be extinguished. There is a caseload and workload limit to what providers can do.

At some point, we could look at the influence of reimbursement on performance. Some clinicians may actually be able to manage a larger caseload in the same amount of time.

Mike:

This approach is quite different than the concept of a primary care panel size, where a long term relationship is the goal. In panels, one establishes a target number and fills the panel to that number. As new patients arrive, the panel is filled to some point where the D = S for visits, backlog is worked down and delay is optimal. So PANEL is the reference point, not new patient slots. As you know, some panels are more stable than others. For example, for two providers with the exact same schedule and working full time and with 1500 patients in their panel, Provider #1 may have 1500 of the SAME patients next year (none new) and Provider # 2 may have 500 new patients (1000 same and 500 new).

Provider # 1 needs all return slots and Provider # 2 needs a mix of new and return. In addition, the RVI for #1 would have to be different than # 2 to accommodate this dynamic. So, the formula

Panel X Return Visit Rate = Days worked/year X Visits per Day is different and could be enhanced.

Mark:

There is a difference between primary care and specialty care. The focus in primary care is to get the panel the right size, keep it, maintain it and, at the same time, recognize the limit to panel. Except for pediatrics where there is a built in graduation, most of the panels remain for long periods. In specialty care the focus is on caseload and in turning over that caseload.

Some primary care panels are more stable than others. Is this a natural variation? If so, then it will all even out and trying to compensate just adds an artificial component (and more variation). I would consider this before I started to introduce adjustments.

Mike:

So in a thought experiment where the total supply is exactly the same, establishing the ratio of new slots to return slots in primary care could be important. Assuming new patients take twice as long as established, another variation of the formula could be:

(Established Patients X Return Visit Rate) + (New Patients X Twice the Return Visit Rate) = Days Worked/Year X Visits per Day.

or

(Established Patients X Return Visit Rate X Appointment Length in Minutes) = Minutes of Appointment Supply needed for Return Appointments

(New Patients X Return Visit Rate X Appointment Length in Minutes) = Minutes of Appointment Supply needed for New Patients

Or, being even fancier and adjusting for no-show rate, the formula could be:

(Established Patients X Return Visit Rate for Established X Appointment Length in Minutes for established) + (Established Patients X RVR for Established X Appointment Length in Minutes for Established) X (No-show Rate for Returns) = Appointment Supply needed for Returns

(New Patients X Return Visit Rate for New X Appointment Length in Minutes for New) + (New Patients X Return Visit Rate for New X Appointment Length in Minutes for New) X (No-Show Rate for News) = Appointment Supply needed for New Patients.

Mark:

I don't like specific "new" slots in primary care. They are a small component of external demand and if we use a "new" appointment type we get into another queue. Monitor the panel size each month. If the provider is low, then add new patients and do it transparently without a specific appointment type.

I don't think a new patient should have a special appointment type or time, for that matter. What is the work here and who needs to do it?

I would not make any adjustments for no-shows. The equation compares activity on the demand (left) side to capacity on the right. If there is a high no-show rate, increase the number of visit slots per day so that supply is greater than demand. That adjustment could handle the variation caused by no-shows.

I think you are making this more complicated than it needs to be. The ratio of new and return appointments on the specialty care schedule grows from the visit rate. If the expected visit rate is four times per year, that is one new appointment and three return appointments. If the visit length is twice as long for new, the schedule is adjusted but the ratio is the same and the visits per year is a blend of both. The equation thus picks up patient visits on one side and provider visits on the other and the new-return dichotomy is built in.

