



Question 1: Allocation of Workload in Specialty Care

In a multi-group specialty practice that pools referrals, doctors have various clinical FTE job descriptions. In other words, some doctors are 50% clinical; some are 20% clinical, and so on. In situations that I am aware of, the 20% doctor is really efficient and has a short waitlist. The 50% doctor is inefficient and has a long wait list. Because the group pools referrals, the following happens:

Let's say there are 10 consults in a given week. They are allocated two to the 20% doctor, five to the 50% doctor and three to the 30% doctor. The 20% doctor sees them within the week and has a clear slate the next week. The 50% doctor has a two-month wait. So, next week, if we want to keep the waits down and equal, one would think that more referrals should go to the doctor with the shortest wait time - i.e. the 20% doctor. However, the situation arises where this doctor could end up seeing, on a patient basis, more patients than the 50% doctor. How should this be handled? Let me say that if this is a fee-for-service situation, one might imagine that it would help "motivate" the 50% doctor. However, the situation described above commonly occurs in the setting of an Alternate Funding Plan.

Answer:

In this question, there are a number of moving parts. First, the clinicians have a variable amount of office time based on FTE status. Second, they have varying lengths of delay. Third, there are varying amounts of work that's pooled based on the delay.

You ask how we allocate workloads (demand):

1. By delay: those with the shortest delay get the most patients, or
2. By FTE status: input based on FTE

In your scenario, you opt for the first choice - pooling the work disproportionate to the FTE time in order to make sure that the delay is load-leveled and balanced. So you have input inequity based on delay variability. This is a common scenario. Provider behavior that creates a waiting time has as a consequence - a chilling effect on demand. The less efficient, longest-waiting-time doctors get workloads disproportionate to their FTE status based on their relative waiting time.

Be aware that this choice has consequences:

1. Inefficient work is "rewarded"
2. Long delays are rewarded
3. "Good" providers burn out

There is a way to break this. We want to look at: a) the FTE status for the individual providers; b) the delays for the individuals and for the average within the "department," and for "any;" c) caseloads for individual providers; and d) visit return rates for individual providers.

We are trying to solve two issues: we want the waiting times to be short and we want the caseloads to be equitable (there may be some exceptions in this regard in a fee-for-service environment, where efficiencies can be “rewarded”- comments below), but generally, the amount of demand needs to be in proportion to the amount of work time (input equity).

Workload equity: We can allocate the new patient work directly proportional to time worked in the office (FTE status). There are various ways to ensure that the input is equitably distributed:

1. Count the new patient consults to each provider and rotate in order - in environments with supply variability, this creates delays
2. Have an equal number of exposure times over a time frame - the same number of new patient slots on the schedule each day, each week or each month. This can create inequities and delays due to rigid schedules. Most holidays occur on Mondays so if one provider has a new patient clinic on Monday or an office schedule on Monday, he/she will get less new patients. No matter, this approach, again due to provider office schedule variability, will create variable delays
3. The two approaches above are prospective. We can retrospectively allocate an equitable workload by tracking over time to see who is over and who is under their correct allocation by the month. We are always a month behind. We can “make up” deficiencies by changing the ratio of the schedule for that provider.

The **delay** issue is a little tricky. First, there may be a disproportionate amount of workload that has created the delay for some individuals. Second, the delay for individuals may be due to inefficiencies or to high return visit rates.

There is a temptation to load-level the work in order to keep the delay down, as described above. What I would do first is try to ensure that the caseloads are somewhat proportionate to the amount of time worked in the office. Review the variable visit return rates, and then work down backlog for those who are inefficient. After we get some system stability, look at how we schedule. If the backlogs are eliminated and we can keep up with the demand, we need to schedule in such a way that there is always enough outpatient new patient capacity to meet the outpatient new patient demand by the week. This may require pooling of referrals, changing the ratios of new to return visits on the schedule, or changing the ratios of new patient clinics vs. return patient clinics, and may also require that some doctors spend more time in the office some weeks, and less time in the office other weeks. This in turn requires us to think about the rigid schedules we have for working in the office, in procedure, etc.

So if we look at the system as a hive, as a big system, the backlogs are eliminated, there is enough new patient capacity to meet the new patient demand, ratios of new and return are flexible in order to keep pace with demand, new patients are pooled and rigid days are loosened, then we can keep up with the demand. Individual clinicians' scheduled time for new patients may vary. Think of it this way: new patient consults at 10 a week are being fed into schedules that open up in advance but only fill during that week. If we keep up with the new patient demand, then those doctors who are scheduled next week to see new patients ought to have wide open schedules, because they're not filling in advance. Their third next available appointment (TNA) is extended due to no immediate capacity, but their patients' functional delays are zero because if no patients are scheduled, no patients have to wait. This is why I believe that we have to measure TNA for each individual provider (which may be long); by the

average (which ought to be short), and for “any” (which has to be within the five day goal threshold).

Fee-for-service vs. salaried: In salaried environments, there should be equal work for equal pay. In fee-for-service environments, we do not have to be so rigid. I would start with equity and then allow variances as long as providers see their own and patients don't wait.

The only answer in a salaried environment is going to be to allocate patients according to FTE. Physicians can choose to be efficient or can choose to work longer but they can't choose not to take their share of patients. That would not be fair to the rest of the physicians. If one physician's line gets too long he/she needs to re-examine his/her time commitment. You will probably hear complaints about compromising quality of care, not having the same supports, etc. You do have to be prepared for those complaints and ideally use them as a springboard for PDSAs on efficiency.

Question 2: Influence of Reimbursement

Some doctors are entrepreneurs and calculate the best “financial octane” which may not be the same as the best “supply-demand octane.” Let me illustrate: **Case 1** – Let's say a doctor can see one new patient in an hour and gets paid \$100. However, he/she can see four follow-ups in an hour and get paid \$30 per follow-up. On a financial basis, it would seem to make more sense to “keep the octane down” and create a lot of internal demand for follow ups. I have certainly seen this in both specialty and primary care. In fact, I know sections within the AMA that have altered their fee code assessments to favor follow-ups. **Case 2** – Let's say a doctor is trying to be efficient. A new patient takes one hour and a follow up takes 15 minutes. The fee for a new patient is \$100 and for a follow up is \$20 (also a common scenario in some circumstances/specialties). This would favor new patients. However let's also say that at least one follow-up is required per new patient. Thus, in a day, a doctor could see eight new patients for \$800 income. More realistically, he/she might see six new patients and eight follow-ups for \$760. Perhaps he/she decides to do phone follow-ups to increase efficiency. A study we did found that the average phone call actually takes 11 minutes (not the two minutes we naively thought), by the time the provider gets the chart, looks at results, calls the patient, answers questions, and dictates a note. But let's say this doctor can do eight calls in an hour. So, the doctor sees seven new patients (a positive for increasing supply) and takes one hour to do phone follow-ups. There is no income for the phone calls, so the total daily income is only \$700. Is there a solution to this or is it a system problem? In fact, are all these Q2 scenarios system problems, whereby we should strive to ensure equitable payment per unit of time regardless of activity?

Answer: Any time the financial incentives are not perceived to align with system goals there will be challenges. However, what this scenario is missing is the “cost” of backlog. How much time and energy is being used by these physicians to “triage” their wait list, deal with fellow physicians looking for exceptions, deal with scared, anxious and angry patients, and deal with patients who are more complicated than they were at first referral because of multiple ER visits and various coping mechanisms tried by their primary care physician? How much of their staff time is used booking and rebooking, that could be used helping the physician in other delegatable tasks? The best mix balances “demand for new” and “supply of new” without a wait (five days for specialty) and within that manages follow-ups as required for the remainder of the

time. Experience suggests that the savings from avoiding triage, rebooking, etc. will exceed the incremental gain from churning follow-ups.

Also don't forget that each appointment has an "administration cost" for booking, billing, etc. and that four follow-ups have four times the costs of one new appointment.

These internal calculations go on. From a system perspective, value is created in seeing new patients.

Strategies:

1. We could change the reimbursement rate. "Could" is a big word.
2. Second, we really have to eliminate individual behaviors that affect the hive. Whereas some individual behaviors help the individual provider, there is a spin off effect on system performance for the hive. If a physician churns return visits for more reimbursement, another provider has to take on more of the new patients for less reimbursement. Show the data. What we have done in the past is act individually and the patient suffers because all the variation leads to increased delays. Those increased delays can harm patients and harm system performance financially. Some gain, while others lose. And we probably take turns. We can start to break this by saying that no patients should wait.
3. We could change the focus and the help in order to make new appointments more "efficient." Service agreements (SAs) can help. SAs can define the work and define the packaging - both these changes make the time with new patients more efficient. We can add help (nurses, etc.) for seeing return patients if we want to influence that component.

There is a bottom line here. If individual providers can act as individual bees, constantly vying for a better financial position with octane or behaviors, then the overall system will perform poorly. Patients will wait and as a consequence, the cost of triage rises, the cost of rework-redundancy rises, the no shows rise, the line cutters rise, the patients sent to ED for consults and admissions rise, the cost rises, the outcomes are at risk and the dissatisfaction rises. What is the cumulative cost of this?

Occasionally providers can see this and the data on cost can help. At the same time, occasionally they cannot see it, particularly when one thinks she or he has an upper hand. At that time someone has to stand up and speak for the customer, and say that the risk to patients due to the delays caused by individualistic behaviors is no longer acceptable.

I have worked in multiple environments similar to this scenario. When we have accurately measured the net revenue (gross revenue minus cost) it is always clear that despite individual beliefs and myths, there is better net revenue for individuals and for the system when we act as a hive than when we act as individual providers. We could study this. We could make providers pay for the effects that their behaviors have on system performance. I have seen that. Salaried providers often get the money but don't pay for how they get it.

Question 3: Number of New Patients

We need a goal for delays for new and for return patients. At the same time, should we have a goal for the number of new patients? I feel for each clinic half day a doctor should see at least

four new patients. That would be a helpful standard. Our current triage process is losing \$\$ and time and does not guarantee that new patients are seen or even seen in a timely manner.

Answer: There are two octane levels. The first is the ratio of new patient visits to total patient visits. The second octane is within new appointments - the ratio of those who go on to a unique procedure to total new visits. We want both of these octane levels to be high. In addition, there are some “skip steps.” Some practices skip the new patient visit and refer patients directly to procedure. This may have an effect on octane, but it skips the new patient visit. Service agreements can often broker this.

A key for system performance is the delay for new patients. The external customers, PCPs and patients themselves all view elimination of this delay as a great value. We want to measure this TNA by individual by individual, by average and by “any.”

With regards to new patients per session, the number of new patients per session really is dictated by the demand. Sometimes we confuse goals and outcomes. The goal is to minimize the delay for new patients. The outcome may be four new patients or two new patients or six new patients per session. At the same time, there is a provider capacity limit and there is a limit to the amount of new patient work that can be “tolerated” by a system. Some of this is determined by the ratio of new to return, how we deal with the returns, and the alternative ways to deal with the returns. Ultimately, the system performance limit would be equal to the amount of capacity for new patients if all visits were new patients. All systems, however, will have to have some return visits to doctors, so the sum of the new plus the return determines the capacity. I worry that if we put a goal of four new patients per session that that number may be either above or below the demand, or may be above or below the system tolerance. If, for example, doctors see four new patients per session, but the overall demand is for five, a delay will ensue and the system breaks down. If the demand is for three, in time there will be unused capacity on the schedules. Secondly, your overall system and the individual components within that system can only tolerate so many new patients. The caseload equation determines that. So setting four new patients per session as a goal may be a problem.

Question 4: Demand Analysis in Specialty Care

In GI, rectal bleeding can be sign of a significant condition. Some of the conditions that present as rectal bleeding (e.g. bloody diarrhea due to UC) should be seen in GI, while others (e.g. rectal bleeding due to due to hemorrhoids) can be managed by others. The demand for our services seems to exceed supply. What can we do?

Answer: Measure first - what is the demand? Then stratify the demand - where does the work come from? What is the work? If demand does indeed exceed capacity, analyze the demand by presenting symptom, condition and diagnosis. Some of these symptoms, conditions and diagnoses can be managed by other SC provider types (e.g. surgeons doing endoscopies) and some can be managed within PC. Use this data to develop Service Agreements (SAs) that define the work and define the packaging of the work. Both these efforts will reduce the demand burden into SC. SAs can also address system issues - which practices are mismatched and to what degree are they mismatched? Where should the demand go?

Clinical symptoms that overlap two distinct “fields” often lead to territorial disputes. These disputes have two common drivers: either who gets the money, or who gets the workload (where money is not an issue - in a salaried environment for example). If one group, in an attempt to get the money, overloads itself (builds a system with more work than capacity), the result is that they do NOT get more money but lose money. Patients waiting cost money, they do not make money.

In addition, look to gain more SC capacity by reviewing the variable visit return rates amongst the providers, creating input equity (new patient demand proportioned by time in office), and reviewing what of the non appointment workload could be done by other non-physician staff within the practice. Some work, like rectal bleeding, can be sent directly to procedure, bypassing the need for a “new patient” office visit.

Question 5: Central Triage

Our current referral system is broken. Some referring providers refer to only one of our Specialty Care (SC) providers, creating mismatches of demand for those popular providers. Sometimes these mismatches are just temporary due to scheduling issues but commonly we suspect that these referral patterns create permanent mismatches. Other providers will send patients to numerous clinicians in our department, hoping to find a way to “cut the queue” (minimize the waiting time). A review of our SC workload shows that our providers spend significant amounts of time triaging and prioritizing the referrals. Some of this is “time wasted” due to a high no-show rate (patients found a quicker way into the practice either through line cutting, into the ED or hospital, or through a parallel referral to another SC provider). We cannot blame the Primary Care providers for this advocacy for their patients. None of our individual SC providers provide any information to the referring providers. The entire referral process is hidden: no information is available. The process is inconsistent and arbitrary. In order to fix the broken process, a central triage mechanism has been proposed. What do you think?

Answer: Your referral process as described is broken. This process is both risky and costly. The term “central triage” has been used to describe a multitude of interventions, some of which may help you and some of which will actually make your process worse.

Beneficial Interventions of Central Triage

- Central triage can standardize the referral process, eliminating arbitrariness, and being able to provide process information to the referring provider: how long the patient will wait, who the patient will see, etc.
- Central triage can minimize the amount of time that SC providers spend “sorting” work. There is no value in sorting.
- Central triage can eliminate multiple referrals for the same patient.
- Central triage can ensure an equitable distribution of work into each individual SC provider (pooling).
- Central triage can set the stage for more accurate measures of demand, capacity (supply) activity, delays, visit return rates and caseloads.

- Through measurement, central triage can begin the process to minimize delays which minimizes the SC non appointment workload.
- Through standard processes, much of the referral receiving work can be done by non-physicians.
- Central triage can ensure that service agreement requirements are met.
- Central triage can reduce overall aggregate system cost. When work is sent to individual providers, mismatches and delays are created. These delays result in line cutting, more work going to more expensive venues (ED), high no-shows, multiple referrals, and an ever expanding need for each individual to triage and prioritize his or her own work. This activity uses up capacity. Central triage can minimize these system effects.

Cautions Regarding Central Triage:

- Often the term “central triage” refers to just centralizing a bad process. Instead of changing the basic nature of the referral process, providers take turns doing the same old process. This approach can result in more arbitrariness and higher delays.
- Central triage often becomes a means to “prioritize.” While “priority” may make some sense clinically (some patients can wait longer than others), from an operational standpoint, priority deteriorates system performance. If demand and capacity are balanced, we don’t need a wait. If demand exceeds supply the system performance will fail. Priority will not solve a demand/supply mismatch, but instead makes it worse. Prioritization requires system capacity. Supply is used to sort the work rather than doing the work. Making patients wait does not make them go away. Making them wait adds cost, risk and adverse behaviors. Operationally priority does not ensure that the sickest patients wait the least, but actually ensures that the sickest patients are far more likely to be delayed beyond the chosen priority threshold.