MMA
Mark Murray & Associates

Central Triage: Dialogue and Assessment

The following exchange is meant to illustrate real life challenges of the current central triage process. This dialogue is meant to be a companion piece for previous papers entitled *Central Triage - Cumbersome Intake Process, Central Triage - An Introduction*, and *Central Triage Mechanisms.*

Question:

I have been reviewing the recent reports and conversations from the Learning Sessions. I am just not sure how central triage (CT) has affected the function of the specialty care and program groups. I posed the following questions to the groups for whom appointments are booked through a central triage system:

Please let me know how you are affected:

1. Are there measurement problems (demand and delay)?
2. Do you hide supply?
3. Does CT force you to appoint patients into the future- past your goal for this work?
4. What is the fear of improvement? That you will get more work?
5. How has CT affected the work up of patients?
6. Has the work up had any effect on delay?

Responses:

Glenrose Geriatric Outpatient Clinic (Edmonton) - Central intake is able to book any appointment request as it comes in. They look at not only when the first next appointment is but they look at patient choice of date, whether the patient has been seen at a particular clinic before, whether there is a clinic that is closer to where the patient lives, etc. CT can tell us what the total demand is for the overall programs or practices (there are 10 geriatric clinics booked by them) by determining the sum of appointments made for ALL sites. At the same time, CT cannot determine the demand for the site of preference since demand is spread over all sites and appointments are made, for the most part, in sequence, based on first available. Some appointments, however, are made out of sequence to accommodate a patient's preference. There is no tracking of preference or deflection from preference to a sooner available appointment. Thus, the workload is distributed by availability with a small preference component.

Most practice schedules are released in a rolling two-month cycle. While some practices have open capacity when the schedules are released, CT's biggest problem at this time in booking new appointments is that the schedules are filled up with returns. When there are no available appointments, they attempt to get the clinic to release more. CT does not hold referrals waiting for new spaces to be released. They deal with them on the day they come in.

Practices are able to measure no-shows themselves, and can also measure “TNA for New for the Clinic” themselves. The one measure they are truly unable to get is “Demand for New” (for the clinic, and by provider). Whatever appointment slots are released get filled. We are not certain if there is a pent up demand for our specific practice and/or if patients who would prefer us are getting sent to other venues. CT informs us that they can measure demand but only at the overall system level, that is, the sum of the demand for all the geriatric alternatives.

There is, however, an indirect method to measure delay for the clinic but not delay to any individual provider. We recognize that an indirect measure of clinic delay without the associated demand data does not help us much. The practice can, however, measure delay as well as demand, supply and activity for return appointments.

We provide our physician new appointments availability for a three month period, that is, the schedules roll open a month at a time and open out for three months. While we don't hide supply from CT, CT is only able to book in to the slots we provide them.

CT receives all the referrals for geriatric assessment in the region. CT does all the booking for 10 clinics from that pool of referrals. We are the only clinic currently in the AIM process and our concern is that as we improve, we will have more work. In essence, we will help reduce the backlog of all ten clinics, but will never see the rewards of decreased delay or a decrease in time for TNA.

Seniors Health (Calgary) - We established a Health One Line referral service two years ago. We previously had a problem with "shot gun" referrals. We commonly found patients waiting on waiting lists for several physicians and found that once patients were seen by one physician they were not removed from all the other lists. The no-show rate was high and some patients even got more than a single assessment while others waited inordinate lengths of time. Some of our geriatricians are better known in the community which we believe also influenced referral patterns. No one tracked referral by “popularity” and the popular physicians had extensive and lengthening delays for assessments. These factors propelled our change to “One Line.”

While we had some ideas or feelings about measures of delay before we moved to One Line, we did not have any formal measure of delay when we started it. It was not set up to measure. But with the assistance of an information analyst, we have been able to measure the delay at this step in getting the referred patient to the right clinic/provider. The information analyst set up the database for us and helps us to configure reports, shape our questions so we can get answers, and assists us in understanding the limitations of the numbers we do pull.

Our analysis shows that there are referral delays due to referral source factors, patient factors and our own process factors.

In addition, we are tracking patients referred inappropriately to our service and who need to be “re-directed” to the appropriate service. Once it is determined that a patient’s needs do not fit with our services, the One Line staff helps the referral source navigate to the most appropriate service. Previously these re-directed patients sat on a waitlist and then learned they were not at the right place/provider (some weeks or months down the line).

We have also given some of the preparation for appointments to One Line - so the information the clinician needs is available at the appointment time. (*Comment from Mark Murray: In a*

sense this is a service agreement. But look at your flow map: a patient traverses primary care to One Line (CT) to specialty care. You have moved the inspection work from the end point – specialty - to the midpoint - CT. You could move this up one step to primary care and make sure the work is done right by using a template that will not "send" unless all the boxes are filled in correctly.)

We have realized that a significant delay in getting an appointment will restrict the pre-appointment activities that can be done by One Line. The required information is outdated when there is a significant delay.

We have developed on-line referral guidelines with associated referral forms as part of the medical service referral and access site. These are early steps toward a service agreement.

One Line allows us to consider a central access point for appointment booking. Central access to different clinic sites may also influence how providers view their clinics, that is, to each practice as a valuable piece of a comprehensive service rather than a standalone clinic. I'm not sure if there has been any work done on provider attitudes to these new referral models.

Analysis of Responses

It appears that CT works differently in different environments. Both Seniors Health and the Glenrose Geriatric Assessment Clinic have recognized the difficulties of the previous customized process: long delays, "shot gun" referrals, referrals based on familiarity or popularity, inability to assess and measure, arbitrary decisions, inappropriate referrals, poor packaging and work-up, etc. Seniors Health, through the development of Health One Line, has improved their referral process. Not only has the referral process been standardized with regard to appropriate packaging and a single point of entry, the workload has been distributed far more equitably (load-leveled) amongst the providers within that single practice. A more equitable distribution of work through the mechanism of load leveling (even if the work is distributed by availability), does serve to reduce delays by reducing the "intentional" demand/supply mismatches that inevitably ensue when work is distributed randomly by popularity.

The Geriatric Assessment Clinic, however, works in a different environment. In that setting, a CT mechanism has improved the referral process by developing a single form of referral and a single point of entry. CT addresses delays by providing a more equitable distribution of work through load-leveling. While the work is distributed by availability, the new patient work is distributed, not within the practice but over a wider range of alternatives - actually over 10 distinct clinic sites. These are the consequences in this setting:

1. CT has load leveled at the system level, so all new patient demand that arrives at the intake station is load-leveled across the entire system of 10 or so sites. Work is distributed according to appointment slot availability. Workload is not distributed as "input equity" (sharing the new patient workload in a measured, intentional way, often in sequence and commonly distributed according to proportionate time in office) but by availability equity. Distribution by availability means that if there is an available slot, it is filled with the next new demand/patient.
2. In an indirect way, the new demand workload is also load-leveled to individuals within the sites, again, distributed by availability

3. New patient appointment availability is, at least in part, determined or influenced by how much the office schedule is already filled with returns.
4. The amount of new patient work can also be influenced by how much time a practice chooses to allocate to the office vs. how much time is allocated to other duties. New patient work is also influenced by appointment length.
5. If new and return appointment types are not distinguished, there is a risk that a high return rate due to a large caseload or chosen behaviors that drive visit return rates up, will fill the office schedule with return patients and squeeze out capacity for new. This behavior obviously affects the availability of new patient appointment slots. Leaving the distinction between new and return deliberately vague is a form of self-protection for the providers and clinics involved.
6. There is no current measure of caseload or caseload in relationship to time worked in the office to determine input or workload equity at the site or individual provider levels. Work is assumed to be equitable based on availability. In addition, there is no measure of demand for sites or individuals within the sites nor is there a measure of system, site or individual capacity at the site, so there is no way to assess system or local functioning and performance.
7. There is an assumption that the supply released coincides with the true availability of supply.
8. While it makes sense to think and act globally, that is, to see the entire system of care and to distribute work across that system in order to prevent local mismatches both at the site level and at the individual level, most groups still act locally, that is, their behaviors are driven by local issues, local politics and local preferences. As a consequence, local behaviors have a huge influence on the distribution of work. For example, if a local site or individual decides there is too much work, decides that they want higher returns, or decides just not to have new slots on the schedule, less work arrives. There is thus a huge incentive to hide capacity and the means and allowances to do that.

The big revelation for me is that in this setting CT **load levels at the entire system level**, not within the practice. Load leveling (pooling) at both the site and within each site is critical because this eliminates distribution of work due to popularity. The same can be accomplished at the entire system level. Some degree of preference can be accommodated in either setting. The limit to preference, of course, is provider capacity limit. If a provider has exceeded his/her capacity limit, preference is nothing but a fantasy. The larger the view (the larger the channel) the less either demand or supply variation will have an effect, that is, cause a delay, but, at the same time, you still have to measure but at a larger view. In addition, while in most cases, it makes sense to have wider channels to reduce the effects of the variation, other times it makes sense to have smaller distinct site channels. How would we decide that? Do patients have a preference due to convenience or some other factor that might drive the work to specific channels? Do the 10 parts of the system, as well as all the individuals within these 10 system parts, communicate? Do they act or can they act as an interchangeable set of providers? If pooling/load leveling is not accomplished through the lens of measurement of caseload and capacity, there is a risk of caseloads exceeding capacity for sites and individuals. This issue is particularly acute in environments where the release of supply is so subjectively driven.

My issues are not that work is distributed from a central site or even that work is distributed (load-leveled) over 10 sites, but my issue is that the basic dynamic of matching demand to supply is lost, as outlined below:

1. Currently, new patient demand can be measured, but only at the CT level. Demand is measured there as the sum of either all new appointment requests or as the sum of all new appointments made. The demand is not measured site-specifically.
2. The individual sites are blind to the demand received at CT and only see demand when it is received and appointed at the site. The only way demand can be currently measured at the site is when demand is **received,** that is, appointed onto the site schedule. Since appointment availability is released each month for the third month out and the released supply is filled almost immediately, then demand appears to equal the supply that was released. The demand appears to = supply because the appointments generated (made) = the number of slots released to CT. If demand exceeds supply then the demand is sent elsewhere. If supply exceeds demand then each site gets "overflow" from others. Load-leveling thus obscures the capability to measure.
3. New patient demand could be measured at the site level if it is measured when an appointment is **made.** A good deal of the confusion arises over the term "demand." There is a tendency to use this term widely and ambiguously, to include preferences and "deflections" to other sites as a part of "demand." We suggest using the term demand in a universal way: demand is workload generated (an appointment made). In this way, with a common definition and interpretation of the term, we can derive common data. I then would see preference and deflections as manifestation of system defects and develop strategies to improve those defects.
4. For the most part, due to a fear of being overwhelmed or taken advantage of by the alternative sites, each site tends to hide capacity through partial or delayed release of available capacity. This is allowed and there is a strong incentive to do this. The most common form of hiding supply occurs by not distinguishing new appointments from return appointments. Since there are more return appointments, this demand sector overwhelms the schedules, blocking any capacity for the competing new appointments. In addition, some practices have much higher visit return rates than others, further exasperating the problem.
5. It is very difficult to measure basic system performance metrics: delay for an appointment, demand, supply, visit return rates, caseloads and whether caseload at the practice or individual level is proportionate to time in the office or even if the caseload is manageable. If you cannot measure, you cannot improve.

The system is working blindly and is stuck with current performance. If demand goes up, look out. In addition, if there are priority queues, system performance deteriorates even further since there are more channels to manage. I am still unclear about how priority affects the release of appointment availability. If practice sites release appointment availability each month but out three months, how can they assure appointment availability for the higher priority queues? Do they hold back on some of the release and release with a few weeks lead time? How do they decide how much to hold back for each of the four priorities?

Overall the larger system will not work if overall demand exceeds overall capacity. Right now if delay for the 10 practices is steady, this implies that 10 practice demand = 10 practice supplies

but there is delay of three months. That three-month delay is expensive for all 10 practices and quite frankly, unnecessary, if the delay is stable. Currently the 10 practice unit may or may not be able to measure system performance and if system performance cannot be measured, it cannot be improved. So you may be stuck. If demand exceeds capacity no matter what you do (CT or not) the system will fail and I fear that you will not be able to see or improve that.

Reconciliation

Ultimately the only way to reconcile this issue is to see that both the individual practices, through their work in the AIM Initiative, and CT are working toward the same goal: to effectively meet patient/customer demand with current system capacity. The problems with the recent referral and handoff systems are broken and there is agreement on the manifestations of that failure. All the involved stakeholders/patients, referring providers and receiving providers want a system that is reliable, safe and effective.

AIM is a structure designed explicitly to address and minimize delays. AIM encompasses a project component (learning sessions, reports etc) and a process component (team, aim, change, map and measure). CT is a change designed to centralize and standardize the referral process. Underneath CT there is an implicit assumption that this change will result in improvement in flow and in matching customer demand to system capacity. But since the underlying assumption was not articulated as an explicit aim to improve flow (to match demand to supply), there were no linked measures to see if this assumption actually accomplished the aim.

Thus, while the AIM initiative has made matching demand to supply with a minimal delay (see your own and don't make them wait) an explicit goal, the goals in the development of CT have been more ambiguous. There is an assumption, I believe, in CT, that matching demand to supply without a delay is the goal but this goal has not been made explicit. That lack of clarity has resulted in a failure to develop measures to gauge and assess success. There is thus a confusion of aim and strategy. The development of CT is not the aim but rather is a change strategy, a strategy crafted to improve flow of patients from one entity to another, a strategy crafted to achieve the aim of matching demand to supply. The central problem with the development of CT is that the aim was never explicit. The aim or goal seems to have been "to implement" with a large set of assumed outcomes that were not connected or explicit (like reduce shotgun referrals, reduce errors, standardize the process, develop a single entry point, give patients more choice, develop referral inclusion and exclusion criteria, guide the choice of the correct packaging, improve the routing of patients to the right venue, give patients more transparency in the process, standardize the work and standardize the process and perhaps reduce the delays). However, without those assumptions (and the main one is to reduce the delays, all the other assumed goals are subservient to that goal) made explicit, then how do we know that CT is an improvement? We measure. What do we measure? We measure whether system capacity successfully met the system demand. Currently the only measure is "Did we implement CT or not? Did we accomplish a standard referral process or not? Do we distribute the workload over all sites?" These are all process measures. These processes are probably required in order to achieve the ultimate outcome goal of matching demand to supply. The measures for success in flow systems are measures of delay, demand, capacity, activity and caseload distribution. These metrics are all within reach. The data exists within CT. But because

there is confusion and ambiguity around goals and change, these measures for success toward the goals are obscured.

The AIM initiative and CT are not antagonistic. CT was developed after extensive input and discussion with patients, primary care and specialty care. The foundation and development work is strong and impressive. In fact, CT is probably a great strategy to further the goals of the AIM initiative. Unfortunately, CT cannot be evaluated as an effective change strategy. The expected process improvements built into the implementation of CT are great changes but need evaluation in light of the ultimate goal. If we evaluate success based solely on implementation then we neglect to see the effects of these changes toward the goal.

I would like CT to clearly see that CT is a strategy, to state clearly that the aim is to match demand and supply and to develop measures to reflect that. At the same time, CT should continue to refine efforts to improve flow from primary care to specialty care. Toward that end, I would like to see CT and the practices work together on common aims and measures and see the flow through the same lens. If that does not happen, CT will dictate the operations. Demand will be prioritized, guaranteeing delays. Load-leveling will be done at the whole system level but the practices will not see that, Practices will continue to hide supply, and since we cannot measure, we cannot improve.

At the same time, unfortunately there is another problem: the reliance on priority as a means of sub-dividing demand. Companion papers have addressed the folly of priority. I suspect that designers of the CT process always viewed priority (triage) as an outcome. And, I suspect that there was just no experience in doing this work in any other way. So the desired outcome - the goal - a set of priority queues - drove and structured the change. We need to ask “Did this change result in improvement? Do multiple priority queues help or hinder our efforts to match demand with supply?” Priority queues do not minimize delays, in fact, they guarantee and institutionalize delays. As long as adherence to priority queuing remains an inherent part of CT, there will be a contradiction between the efforts of AIM (minimize delays for improved system performance) and the aims of CT (have priority to protect system function). That issue will have to be resolved.

Short-Term Reconciliation

I recognize it is extremely difficult to change course, or actually to continue the course toward optimal system performance. In the short run, the Geriatric Assessment Clinic could serve as a pilot test model. This is not an optimal test but I don’t see many other alternatives.

Proposed changes to test:

1. Close the practice temporarily for new patient work.
2. Establish two appointment types: new and return. Eliminate priority.
3. Work down current backlog so that the delay for new patients is less than a week.
4. Determine the new appointment capacity by looking at total appointment capacity and determining the ratio of new to return and visit rates from previous performance data.
5. Open the schedule out only a week in advance, but blink the schedule open one day at a time, each day. Make sure that each week the practice opens/releases all of the new patient capacity.

6. Use the strategies described in the “TNA for Any” article.
7. As the slots open, fill them “just in time” from CT. These appointments could be a mix of all priority types. Thus, the practice determines its capacity for new patients, releases this capacity out a week and CT fills this capacity to full.
8. The delay for new patients into the Glenrose Geriatric Assessment clinic is now no longer than seven days.
9. Determine whether the clinic is taking their measured share of new patient work by determining the proportion of Glenrose office FTE within the total office FTE for all 10 sites
10. If the Glenrose is taking less than their share of the new patient workload, as determined by the office FTE analysis, then improve performance by a Care Team workload analysis, by reviewing individual and collective visit return rates, and/or by strengthening service agreements.

By definition, patients referred to the Glenrose Geriatric Assessment clinic will wait no longer than seven days to be seen. If the clinic is able to see its fair share of all new patient work, this indicates that overall demand is balanced by overall capacity for all 10 sites and it is possible to work without a delay in all sites. On the other hand, the experiment fails and the clinic is only able to accomplish this feat by taking on less than its “fair share” of the new patient work. At the same time, since the clinic has used multiple strategies to achieve a better demand/supply balance, then this failure would indicate that overall system demand exceeds overall system capacity. This failure then indicates that the system performance result in that instance is inevitable - deteriorating system performance and breakdown. At that point, it does not matter what you do - all efforts will fail.