



## Cumbersome Intake Systems in Specialty Care

Without measurement a system cannot improve. Measurement of key system performance indicators is critical for seeing current performance and whether changes made actually result in improvement.

Our healthcare systems match demand to supply. This is not a choice. The choice is whether the practice matches demand to supply well, or does it poorly. In demand-supply systems, critical performance measures include measures of demand, supply (corresponding capacity), activity, and measures of system and individual provider delays.

Many practices, particularly those in speciality care, have extremely cumbersome “intake” processes - the series of steps patients take from discovery and initiation of demand through the process of referral itself, receipt of the referral, evaluation and “triage” (including potential rejection), storage of the referrals, re-contacting patients, appointing the referral, and finally, the delivery of the actual appointment. Because of the complexity, ambiguity and inconsistency of these intake processes, it becomes difficult if not impossible to measure system performance or to see if changes made actually result in improvement.

Key conceptual concepts are discussed initially in this paper. These are followed by an example of a “cumbbersome” process.

### Demand Measures

In demand-supply systems, it is necessary to measure demand (patient need). Demand could be measured at any distinct point along the extended series of steps in the referral process. However, a demand measure at any of these steps is problematic. Each of the demand measures is subject to dilution as the demand moves from one step to the next. This dilution of the demand count could be due to rejection, triage to an alternative path, or to no-show. Because of all these factors, measuring demand at an early step in the process could give a false demand count. If, for example, we measure demand when the work is initiated, when the work is sent, or even when the work is received, because demand can be rejected within the triage process that occurs later in the process, measuring demand early in the process would result in overcounting to the extent of the number of rejections or workload “triaged” to another alternative pathway. If we count demand as the work accepted as “appropriate” and appointed into an actual supply component (appointment), we are counting demand that has been delayed and batched. The demand measure will then display an artificial variation caused by the batching of the intake process itself. Demand batching creates an artificial appearance of demand variation, and coupled with the known, wide and artificial supply variation, makes matching that oscillating demand variation with oscillating supply variation much more difficult.

Instead of smoothing the natural variation of demand, we have converted this natural variation into even more variable artificial variation.

## **Delay Measures**

Measurement of delay is also not always straightforward. In systems where demand is received and scheduled in advance (even if that advance period is short), the most effective and informative measure of delay is third next available appointment (TNA). An alternative to TNA is to measure “actual wait,” the time lapse between the declaration of demand and the delivery of the supply (appointment). However, measuring delay that calculates “actual wait times” does not capture an adequate evaluation of system performance. For example, if 10 patients who want an appointment in a week contact an office and get an appointment in a week, they got what they want. If 10 other patients contact another practice and want an immediate appointment and get one in a week, the measures of actual delay are the same, yet one set of 10 patients got what they wanted, and the other set did not. At the same time, while the measure of “actual waits” clearly is deceptive and inadequate in primary care, the measure may have more utility in specialty care, but only if the process for sending and receiving is simple. Despite this partial disclaimer, in specialty care the best measure of delay in scheduled settings still remains third next available appointment.

The cumbersome processes described above (with multiple steps and “inspections”) are more often than not associated with long delays (extended TNA). In turn, extended TNA is what drives these cumbersome processes in the first place. If there is a delay, most practices, believe that demand exceeds supply (there is more work than they can possibly do), and they move toward self-protection and prioritization to “manage the work.” Consequently, they get into a vicious cycle of poor measurement and poor performance. If there is a long delay, or even if there is a non-measured “perception” of more demand than they can handle, the practice creates a system of multiple delay steps (triage and prioritization) in order to protect clinicians and in a vain attempt to chill the demand. This behavior actually extends delays even further.

The choice of where in the process to measure demand has an effect on delay measures, either on TNA or on actual waits. Delay cannot be accurately measured until demand is known, but demand often does not materialize as a measurable entity until late in the series of steps in the process. Delay is difficult to measure since we have no idea where to “start the clock.” The determination of where in the process we choose to measure demand will determine the time lapse of the actual wait (the time from the point of demand measurement to the eventual delivery of the supply). On the classic continuum of demand initiation, demand sent, demand received, demand triaged, demand appointed, demand actually seen, moving the demand determination to the left makes the actual wait for those patients who get through the process longer and more accurate but will artificially inflate the demand by overcounting it due to dilution of demand as it progresses through all the steps. Counting demand further to the right on the continuum gives a more accurate picture of demand but hides some of the significant delays experience by patients who traverse through the entire series of steps.

Measuring delay as TNA is also problematic. The third next available appointment is commonly measured at the same time on the same day each week. In a smooth flowing system this

measure directly correlates with the balance of demand and supply. If demand exceeds supply and supply does not flex to keep up, the TNA will rise. Many cumbersome intake systems obscure this direct relationship. Fluctuating demand or supply can be buffered both within and in between the multiple steps in the process. Many of these steps can act as “parking lots” for demand or “buffers” to protect supply. As a consequence, TNA can appear stable or acceptable but at the same time, demand might be accumulating or even diminishing. If we applied a measure of actual wait time to this process, we would see a delay that is actually much greater than the TNA indicates.

Since TNA is difficult to measure, many practices will attempt to measure the delay as actual wait time with the issues discussed above. This is often an accommodation strategy- change the standard measure in order to accommodate a defective system. It is far better to keep the standard measure and change the cumbersome system.

In addition, we cannot measure delay as TNA or evaluate system performance unless we know and can see the supply.

## **Supply Measures**

Cumbersome intake systems not only lead to serious difficulty in measuring demand and delay, but they also result in difficulty measuring supply. This is particularly evident in practices where the intake process occurs outside of and independent from the receiving practice. The providers within the receiving practice, unsure of what their demand or delay really is, will often hide the schedules or obscure supply as a self-protection mechanism, and release capacity available for booking (supply) reluctantly and inconsistently. Providers, acting as independent and individual supply units, can also self-determine when and how much capacity (supply) to “release” for scheduling. These subjective decisions are arbitrary and are based not on data, a sense of fairness and equity (dividing the new patient workload evenly) or as a response to demand, but are based on “feelings,” opinion and anecdote. This leads to difficulty in measuring supply because the true capacity is unknown and is hidden in a completely independent box. This is a process of supply dictating to demand rather than the other way around. If demand is measured when it is received, which is often the only view of demand that the receiving groups gets, and, at the same time, supply is hidden and released arbitrarily, demand will appear to equal supply since demand is only “counted” when it is accepted and when supply reveals itself. Any excess demand gets hidden further back in the process. Demand is defined as acceptance but supply decides what and particularly how much is acceptable, so since supply has a limit, demand has a limit and demand appears to be equal to supply.

In addition, if the supply does not distinguish new appointment types from return appointment types, and since there are more returns, the return appointment type will overwhelm the future schedule, crowding out any capacity to see new patients. Since the intake process primarily deals with new patients, the blend of new and return appointments makes new patient capacity even more scarce. As a consequence, the intake group tries to manage demand but has no idea of the supply available to meet or not meet that demand and the practice has no idea of demand. Neither knows the delays. So we rely on indirect methods to measure - complaints by referring providers, complaints by patients, and an inaccurate measure of actual delay.

A cumbersome booking process perpetuates the appearance and perception of overwhelming demand, in part due to the current backlog and in part due to the incapability of this type of system to measure itself. Without measurement and with only perception and false experience as a guide, the cumbersome process becomes self-fulfilling and self-perpetuating. Victimhood is validated and systematized.

These cumbersome systems, for the most part, blind both the intake group and the receiving group since the key measures (delay, demand, and supply) of system performance cannot be determined. As a result, both groups resort to systems of prioritization which worsens the system performance. In the meantime, the system performs as it is designed. If demand and supply balance, patients are still unnecessarily delayed. If demand exceeds supply, the workload increases and, in particular, the line-cutting increases (more patients are deemed emergent, creating a need for more “urgent” capacity to meet that demand). That capacity is pulled away from the “less urgent” and the delays are extended even further. In addition, more patients cut this queue entirely (get admitted, call the on-call physician, etc.), bypassing the cumbersome process. If supply exceeds demand and we have more than enough capacity, we will never see that. In fact, this system will constantly reinforce the “feeling” (not the measure) that demand exceeds capacity. So not only is this process somewhat ambiguous but the inherent delays create more workarounds and variability.

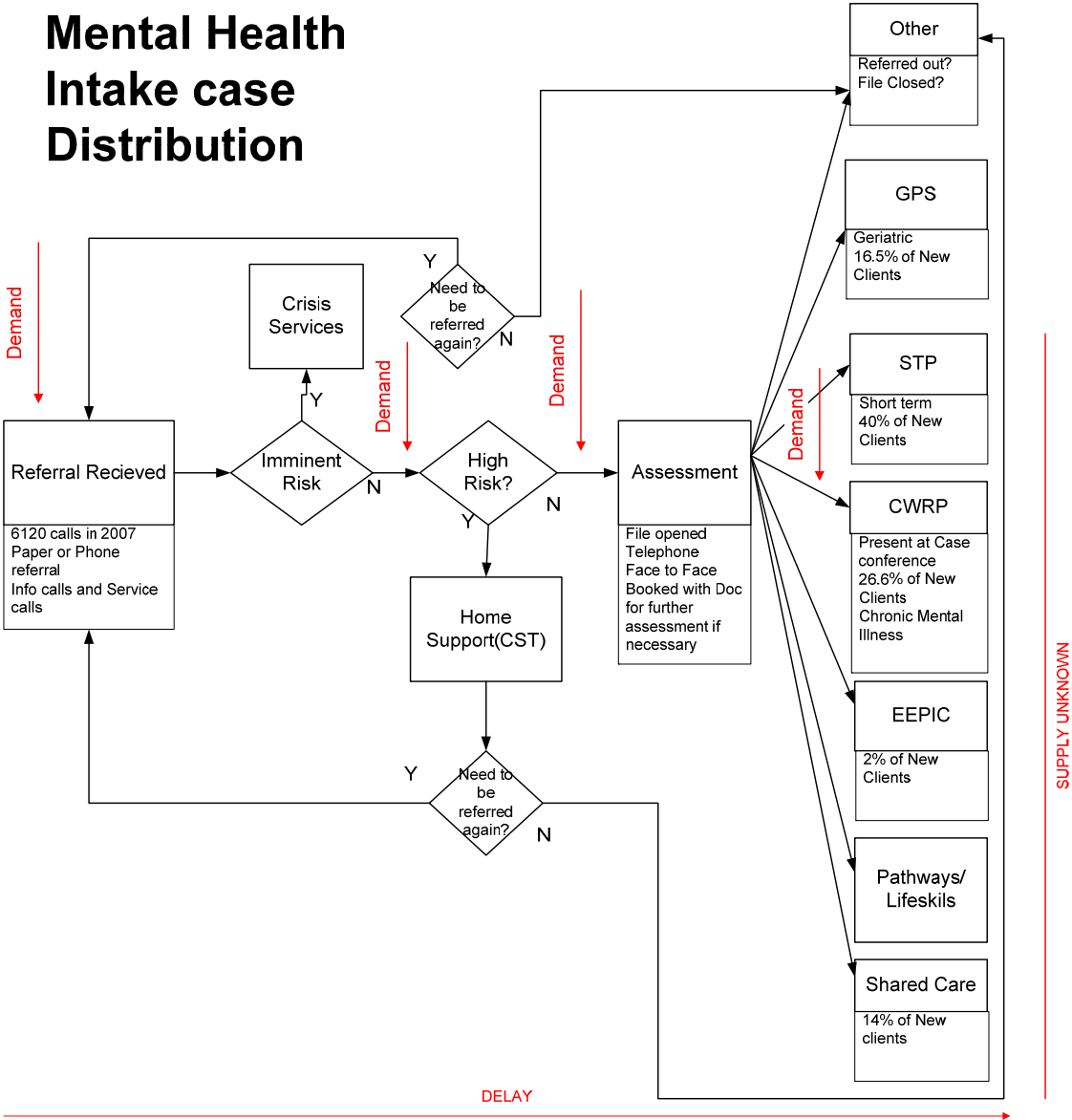
## **The Fix**

In order to fix this situation, there are a number of critical and necessary steps. Just like a patient’s physiological system has a set of vital signs and unique measures to determine current system performance and improvement, we need the same for our operational systems. To understand current system performance and to see whether the changes we make actually result in improvement, we must measure the vital signs of that system. In flow, the vital measures are demand, supply, activity and delay. Following are the critical steps for improvement:

1. **Measure.** We cannot determine system or individual provider baseline performance without measurement. Worse, we cannot improve without a baseline measure. We thus need to convert unmeasurable systems into measurable ones. We do not want to accommodate a bad system by changing the basic performance measures. Instead, we need to change the system.
2. **Connect the intake process to the receiving practice.** The process and the practice do not need to exist in the same location but measurement of demand and demand functions needs to be connected to measurement of supply and supply functions.
3. **Develop a non-ambiguous process for the declaration of the demand-referral-receive and appoint process.** This process needs to be simple, without any unnecessary steps, should contain minimal, if any, triage steps, and must be totally transparent. The ultimate goal should be for this entire process to occur within five days.
4. **Develop service agreements to define the workload and the packaging of the work.** A requirement for the correct packaging can minimize or eliminate many of the steps within a cumbersome process.

5. Eliminate hidden and intentionally obscure supply so we can see the journey from demand to supply and measure this as TNA. Supply is often “hidden” by individuals for self-protection and “released” as according to individual perceptions of the practice world. We have to see the supply and see it far in advance that we can measure it and match it to demand.
6. Measure demand as workload generated (appointments made on each day). A simple straight forward process allows this. In processes fraught with delays, multiple steps and triage, we are tempted to try to change the demand measure to a measure of referrals made, referrals received, appointed referrals, etc. This results in over counting of demand due to rejection rates at the last step in the process. It could also result in hiding the time lapse from referral to decision.
7. For fairness and optimum system performance, create the concept of input equity - an equitable sharing of the “new” workload across all the supply resources (providers). In addition, by using provider capacity limit calculations (ideal caseload), workload limits can be determined.
8. Distinguish between new and return appointments. If the visit rate is greater than two visits, the returns are greater than new appointments. If the two types are not distinguished, the return appointments – through random variation – will block the schedule against new patient capacity.
9. Eliminate referrals based on preference or popularity. Each provider has a limit, and we need to protect providers to work within that limit. We need to share the work based on data not opinion. In order to ensure that delays are minimized, new appointment work needs to be pooled and not sent based on popularity.
10. Eliminate priority or triage except, for a small number of truly emergent cases. These emergent cases should be handled as unscheduled work and directed to the appropriate venue for that function. The majority of the work should be managed as a scheduled function.
11. To operationalize these improvements, practices will need to adopt local changes such as the distinction between new and return appointments, develop local based measurement systems and commit to input equity and pooling. At the same time, particularly in larger systems with multiple referring sources and receivers of the work, there is an advantage in developing a central mechanism (central triage) to manage the flow of work. A central mechanism, operating within a set of standardized guidelines and rules, can perform as an objective conduit for the demand. This central mechanism must be operationally connected to the practices and take standard direction from the practices. This connection to the practices can provide standard measures and comparisons, ensure standardized practice, serve to add local capacity by subtracting work from the practices and, through periodic audit, can be expected to continuously improve.

# Example of a Cumbersome Process



In the above diagram, the process is outlined in black and the comments in red. At first glance, the process seems acceptable, but with further analysis the issues become apparent:

1. This is a map of a flow system - demand moving through a series of steps. The intake group is disconnected from the receiving mental health group.
2. In a flow system we evaluate system performance by how well we match the customer demand to our capacity. The key question around performance focuses on the delay (how well we match demand and supply). As a corollary we need to know the demand and the supply in order to know if we can do this work, or how well we do the work. So we need data on delay, demand, supply and activity.

3. How do we measure the key metrics here? We can't. So we are blind to direct system performance metrics. We cannot accurately measure delay and we cannot measure demand and the supply is hidden in another completely independent box. The demand could be measured at a number of distinct steps along the process (in red). Each of these demand measures is subject to dilution as it moves to the next step (rejection, triage to an alternative path, or no show). We are forced to measure delay as actual wait. At the same time, the determination of where we measure demand will determine the time lapse of the actual wait. The more we move measures of demand to the left, the longer the measure of actual wait, but the more we dilute the true demand (the workload that will be met by supply) and the more inaccurate the demand measures become. So moving the demand determination to the left makes the actual wait for those patients who get through the process more accurate, but it will artificially inflate the demand. Moving demand to the right to get a more accurate picture of demand hides some significant delays. The supply is simply unknown and is often determined by subjective opinions and decisions. So we rely on indirect methods to measure - complaints by referring providers, complaints by patients and an inaccurate measure of actual delay.
4. In the meantime the system performs as it is designed. If demand and capacity balance, patients still are unnecessarily delayed. If demand exceeds capacity the workload increases and in particular the line-cutting increases (more patients are deemed emergent, creating a need for more capacity to meet that demand). That capacity is pulled away from the "less urgent" and the delays extend. In addition, more patients cut this queue entirely by getting admitted, calling the on-call physician, etc., effectively bypassing the cumbersome referral process. If supply exceeds demand and we have more than enough capacity, we will never see that. In fact, this system will constantly reinforce the feeling, but not the measure, that demand exceeds capacity. Not only is this process ambiguous and arbitrary but the inherent delays create more workaround and more variability, further deteriorating performance.