# Flow Analysis of Specialty Care Practices and programs (For Participants)

**- Dr. Mark Murray in collaboration with June Austin (Alberta AIM Faculty)**

Many Specialty Care (SC) practices and programs are very complex and require analysis of multiple flows. These challenges and complexities arise from the following factors:

## An understanding of system performance

There is a fundamental dynamic at play here: each of your practices or programs uses capacity (resources) to meet patient demand. To assess how well your "system" performs, you need to see how well you use that capacity to meet demand. Is there a delay? The delay represents the relationship between capacity and demand. If there is an extended but stable delay, then there is a backlog of work due to temporary mismatches of demand to supply (variation). The practice is keeping up with demand but there is a wait. This is an expensive and risky way to perform. This delay can be solved by elimination of backlog and by reducing variation (catching up and keeping up on a daily, monthly, annual basis). On the other hand, if the delay is continuously worsening, this indicates a permanent mismatch of demand compared to supply. If demand exceeds supply, the delay will continue to become worse and the practice will fail.

Many practices and programs can, and like, to tell us what they do. You can provide us with endless lists of tasks, behaviors, activities and job descriptions. Some of you can even describe how you do it: using your capacity to meet the various types of demand. But, none of you can initially tell us how well you do it. We appreciate that the entire concept of matching patient/customer demand can be ambiguous. And as a result, it is difficult to see the relationship between demand and supply and how it is manifested in the delay measure. If we ask: "how long do your patients wait for you to see them after they have declared a need?" most practices and programs simply cannot provide the clear answer. You need to learn to ask yourselves: What is our current performance and how do we measure it? How long do our customers wait? Do we successfully use our capacity to meet demand?

## Perspective

We recognize there are significant challenges on the practice side. However in a sense, some practices make this work more complex than necessary which results from the unfamiliarity with "seeing" the work flow through the patients' eyes. We as clinicians are often blind to the issue of how well the system performs. System performance is often described in subjective terms or is very ambiguous. We cannot see that our capacity is not meeting the flow of work and the flow of patient demand. This blindness results in a focus on increased volume as the improvement strategy, increased resources as the answer to all problems, the incorrect use of cross-functional maps as a way to see activity as the only measurement, and a vertical instead of horizontal view of the work.

© Mark Murray and Associates & Alberta AIM

We are reminded of an analogy here: there is a picture—a drawing really—of an old woman. If we stare at it long enough, we can get a shift in view, a shift in perspective. We start by seeing an old woman and end by seeing a young woman depicted in the same drawing. There is another set of "pictures" just like that: a background of dots, but if we stare at it long enough we can see a Lion. That's what we have to do here – retrain ourselves to see a different view of our system.

We have to, in essence, "stare" at our systems long enough to see the patient's journey or the horizontal flow, to see the steps the patient takes and the intervening waste and delay, to see all the parallel and competing flows, to see these flows in maps, to add measures to the maps, to find the choke point or the constraint, to make change and measure for system performance improvement. Ultimately we need to be clear enough in our view to be able to see the opportunity for improvement that lays not in adding resources or volume but in re-designing how we do based on ongoing measurement and workflow analysis.

The shift in perspective is not easy. There is complexity here: multiple demand streams, multiple decisions about how to sort out and channel the demand into various demand streams, competition between appointment work, associated non-appointment work and non-patient-care work, multiple disciplines with varying levels of cross training, unknown constraints, and inadequate attempts at maps. To make sense of the complexity, the "how well do we do it" question requires appropriate maps and measures. Your challenge is to understand the inherent value in changing your world view and assisting to create a new picture of your system.

## Assessing System Performance

Since an assessment of system performance requires measurement of how well we do it, understanding the units of demand that meet the units of supply is essential. System performance is measured by how well the system uses its capacity, whether that capacity is individuals or individuals merged into bundles to meet demand, and whether demand is individual bodies or bundled into a group. Thus, we first have to see how the workload is currently sorted and divided and what supply is allocated against that demand. This is the demand stream issue.

The decisions and processes whereby demand and supply are sorted—either bundling the demand or bundling the supply should populate the maps and inform the measures and the data captured in the spreadsheets. Many practices use the same individual providers in different functions—as individuals, as part of bundles (dyads, triads, and tetrads) or even as the clinician who supports group visits. Thus, some providers straddle two or three lanes or types of work. In these cases, the supply persons are acting differently to support and meet various types of demand stream. Each demand stream has to be evaluated independent of the others as separate processes. For example, if Emily the RN works as part of dyad and triad and as an individual, she will appear in three distinct spreadsheets because she in fact provides 3 different services. Many practices err in their measures when they measure all work or demand to Emily rather than parts of Emily as distinct entities or distinct demand streams.

There are no clear operational standards here from practice to practice. Therefore each practice needs to evaluate and individually sort out the workload (demand). Some Chronic Disease Management programs will direct all work through the RN first, while others, performing the same type of work, will direct work to specific disciplines (i.e., RN or RD) while still others use all disciplines as interchangeable. It is critical to ask yourself the questions that challenge you to better understand your own systems. We need to ask how the work is divided, then map that flow and add measures directly to the maps. Each arrow on the map that indicates flow of work from one entity to the next requires a measure of TNA, DSA, and often caseload. The caseload measure helps us see how the work is distributed amongst the interchangeable providers meeting that demand. Without a map we cannot measure, and without measures we cannot improve. In order to evaluate performance, we need to:

1. get clarity on the sorting and distribution of work
2. map the flow of demand and supply entities
3. use the measures and spreadsheets to assess the performance.

Due to a combination of unclear understanding of the demand-supply dynamic and inadequate measures of system performance, there is a tendency to reflexively believe that "we don't have enough resources." We frequently hear and repeat the refrain, "if we just had more resources, we would be fine." In parallel, as faculty we often hear "AIM believes we will improve if we just increase our throughput and our volume", and that the goal is to see more patients. The backlogs of work—both hidden and visable—are just accepted as inevitable and used as excuses: "We can't do the work in a timely manner, because we are so far behind. We need more resources." These beliefs are deeply embedded in the culture of healthcare.

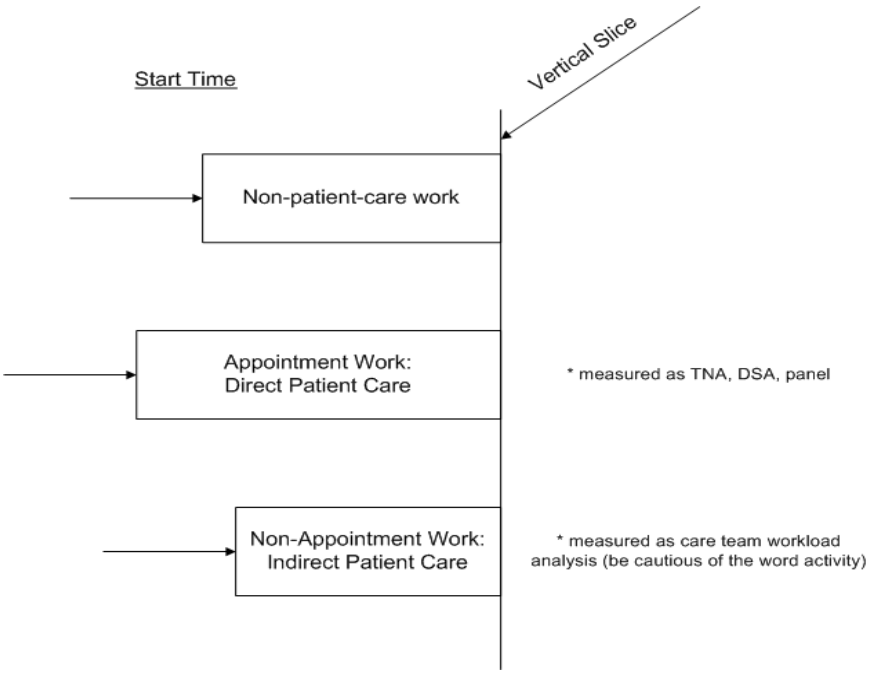
# The view: horizontal or vertical?

## Horizontal:

If we correctly follow the patient's journey, we see that journey as a series of interconnected demand meets supply steps, and the following questions can be answered:

1. How well do we meet demand at each step?
2. How do we allocate supply at each step to support the appointment (new + return) and the non-appointment work?

This "flow" is best visualized on a map as a horizontal flow. The patient demand moves from left to right across the page, following the steps. Each step has an arrow which points the direction of the demand flow. Each arrow needs a measure of TNA, DSA, and panel size. There are three horizontal paths, moving in parallel. Our prime focus, and the focus of measurement within the spreadsheets, is on the appointment demand stream, since this is where "value" generally occurs, as an appointment, whether that appointment is made far in advance or spontaneously on the day of the encounter. We measure this stream as demand for appointments, supply for appointments, activity, and delay. At the same time, the non-appointment work flows alongside the prime demand stream. The non-appointment work is dyssynchronous, that is, the work generated in the appointment commonly gets pushed forward—i.e., a blood test done today as part of the visit is not evaluated until hours or days later. The time spent in non-appointment work directly competes with the time allocated for appointment work and with the third demand stream—the non-patient-care work.

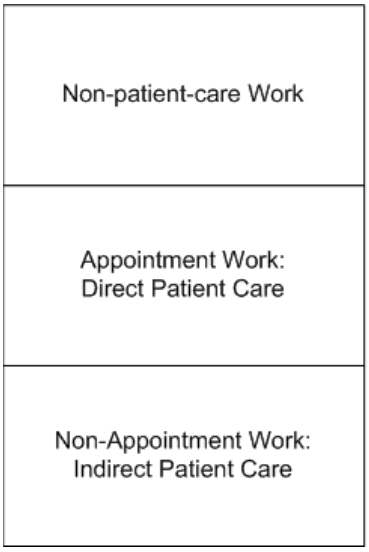



© Mark Murray and Associates & Alberta AIM

The horizontal flow is similar to a patient prepared for a CAT scan. We are looking at the patient horizontally, from top to bottom or from head to toe. Once we put the patient in the CAT scanner, slice them vertically, turn them sideways, and we can see the organs inside the body. If we pivot this diagram on the vertical line labeled "vertical slice", we can see the three horizontal "demand streams" on end and would then be looking through the perspective of the supply or the vertical flow. When we display this vertical view as a pie chart, we get the classical Care Team Workload Analysis, that illustrates the types and amounts of work performed within a specified time frame.

## Vertical:

The vertical perspective is also important. At any supply step along the horizontal flow, we can take a vertical slice through that horizontal flow. When we slice through that flow and turn the slice sideways, we can see a “picture” of all the work being done at that time. Depending on how thick or thin we slice this view we can widen our view or the window through which we view the system. A one month slice gives us more insight than a thin slice representing one day or in some instances one hour. For the most part, the swim lane or cross-functional maps many teams provide us with initially are really just a series of thin vertical slices stacked together. Mapping will be discussed in more detail later in this paper.




The vertical slice shows us what we do and how we've allocated our time (The horizontal flow shows us how well we're doing it). It is the view from the worker perspective. These three distinct demand streams “come at” the worker. Does the worker keep up with the appointment work? Do they keep up with the non-appointment work? Answering these questions tells us how well we are doing. We could, for example, get a view through this vertical slice that shows that the clinicians see 50 appointments a week, in aggregate; and next week they see 50 appointments, or we could see that we spend three hours per day at seeing patient appointments. But does this information tell us if this is the right amount of time allocated or if our system is performing well? No—it tells us what we did.

What we see is, in essence, a Care Team Workload Analysis (CTWLA). The CTWLA provides a retrospective view of the workload, allows us to see how the work tasks have been allocated and how the clinicians spend their time. The CTWLA is a strategy that allows us to look at what the work is, in order to determine if the right person is doing the right work. We divide the work into three major categories discussed above: non-patient-care work + patient-care appointment work + patient-care non-appointment work. These three duties, tasks, or activities all compete for clinician capacity, space, or time[1]. How much time we spend on one has an effect on how much time we can spend on the others. If we want to alter our supply for appointments, for example, we have to pull time out of the other two categories. When we perform multiple slices at different supply steps, we can see the CTWLA for all of our supply people and we can begin to analyze what the work is and who is doing it. If we can shift some of the work, we can often gain capacity for our most constrained workers. It is important to use these three specific categories, because the change strategies for each are a little bit different. When we're looking at that horizontal flow, we're looking at the flow of only one of these three task categories—the patient-care appointment work only. While we're slicing through the appointments that move horizontally, there are also two other competing streams—the non-appointment work and the non-patient-care work.

### Bringing Horizontal and Vertical Together:

The only way we can tell if our system is performing well is to flip it horizontally and see if that activity, the number of appointments seen, is enough to keep up with the demand or not. Is the waiting time getting better or worse, or is it stable?

Both these views are important. In the horizontal view, we can see how well the system performs in delivering its chosen services to the customer. In the vertical view, we can see how we allocate our resources at any slice through that flow. Through that determination we can choose to alter that allocation in order to improve the flow. The cross-functional maps provide detail in the vertical perspective; and the supply/demand maps show us the horizontal. The CTWLA should be done beginning at the system constraint, that is, the horizontal flow system choke point.

## Demand streams

Determination of the demand stream is a critical task but a common cause of confusion. The demand stream is the type or category of work that the practice will meet with its supply. The demand stream refers to how the demand (workload) is channeled. While non-patient-care work and non-appointment work can be referred to as "demand streams," this term most commonly refers to the various streams of appointment work.

[1] See Mark Murray and Associates (2011) , “Cycle Time vs. Turnaround Time (TAT)” http://www.albertaaim.ca/articles.html

While the workload (demand) that arrives into SC or the program along the appointment demand stream has often been filtered through Primary Care (PC) or a central referral source, that workload is often received and then "triaged" or sorted in multiple ways. For example, in the many programs, workload (demand) can be sorted:

A. by clinical condition – i.e., DM vs. obesity as within some CDM programs

B. by discipline (provider types) – i.e., Educator vs. RN vs. OT, etc.

C. by individual providers within the disciplines

- Even when work is sorted and channeled to disciplines (provider types) it is often still distributed (divided) to select individuals within the set of interchangeable providers from that discipline most commonly by preference, popularity or by availability. These methods inevitably result in demand-supply mismatches for individuals. In our experience very few practices and programs distribute work based on proportionate FTE (pool the work).

D. by “cross-trained” members of distinct disciplines

- Some practices "cross train" in order to ensure that providers from different disciplines are interchangeable and workload can be distributed across standard disciplines—i.e., in some CDM practices RNs and RDs can both perform the same work interchangeably.

E. by "multi-disciplinary" team

- If the handoff of work that goes to multi-disciplinary team (MDT) occurs between days, then the issues outlined in B, C, and D apply. However, on occasion, "multi-disciplinary team" refers to sequential handoffs within the day (handoffs in sequence on the same day—or even within 2 days as some practices have two-day visits), or MDT can refer work within the day if all components of the multi-disciplinary team meet the patient simultaneously. In these cases of within day handoff, each demand is met by a bundle of supply—a dyad, triad, or tetrad of supply components (i.e., OT, PT and SLP all meet the patient either in sequence or all simultaneously). The supply is measured as "bundle" and the demand is to that bundle. The spreadsheet should reflect this.

F. by bundling the demand into a group visit which is met by either one or a larger combination of supply components

   - In this case the "demand" is really a collection of smaller demands into a bigger "Demand" for group. The supply is either a single supply resource of a bundle of supply components as well. This should be captured in the Excel spreadsheets. The bundle of demand or supply is the unit of measurement for demand or supply (demand for bundle, supply of bundle) on the spreadsheet.

G. by one approach and then another

   - For example, some practices have all patients see the RN first (sorted, triaged, and channeled to individual within a discipline) and then the work is divided and directed towards either other individuals or groups (some form of the MDT). Other practices start all patients through a group or through some form of multi-disciplinary team and then divide into individuals or even groups.

The decision made on how the work/demand is sorted is critical because this sorting decision drives the supply demand dynamic, informs the maps, and allows us to assess with measurement how well the system is performing.

## Streams as appointment types

Each demand stream functions as an appointment type. The inverse is also true: each appointment type functions as a demand stream. In a sense, all the services performed in SC function as "appointment types." "Appointment types" is just the term to describe the channels of work.

Inclusion and exclusion criteria determine what demand steam a patient is directed toward. This is, after all, the function of triage. Each demand stream functions as its own demand-supply system and its performance must be evaluated as TNA and DSA, measures both independent from, and in parallel to, all the other demand streams. As noted above, sometimes each discipline has its own demand stream. Certainly each site in a multi-site program is a distinct demand stream. Within the demand stream, appointment work competes for capacity (supply) with any associated non-appointment work. And to add further complexity, within the appointment work, the new appointments compete with the return appointments. We need a balance not only across demand streams but also within each demand steam – new vs. return (within the appointment category of work), appointment vs. non-appointment work (across the demand stream) and office work vs. other work – i.e., OR, on-call, academic etc. (between or across demand streams). Then we put a worker in front of the line to support that line. The system performance question is then: “Is the worker keeping up with all the separate lines or not, within their overall capacity?

© Mark Murray and Associates & Alberta AIM 

## Confusion over the term "activity"

There is a great deal of confusion over the term “activity." The term "activity" in the context of demand, supply, and activity refers to a specific meaning – how well do we use supply for appointments: did we use our appointment supply or not? This is a retrospective "measure" that we obtain by determining how many appointments we saw. This measure has value because of its relationship to appointment demand and to appointment supply and hence to TNA. In this context the term has a relatively narrow meaning. The standard spreadsheets are designed to capture this narrow meaning of the term. In this DSA context, we are viewing appointments only. These appointments can be for individual clinicians, for groups (bundles of demand), or for any bundle of clinicians (dyads, triads, and tetrads). All appointment work carries along in parallel any associated non-appointment work. This category of work is important, but is independent and needs to be measured as such. Some behaviours, duties, or activities straddle appointment work and non-appointment work (i.e., telephone follow-up visits). The simplest way to categorize this work is to ask if the work is scheduled. If the work is scheduled for both patient and clinician, then this work functions as an "appointment." On the other hand, if the provider has time scheduled to do this work but the patient is not scheduled, then this should be categorized as non-appointment work—work associated with the appointment, but not an appointment. Many types of unscheduled work such as telephone follow-ups can and should be moved to appointed work.

The term "activity" has a wider meaning when used in the context of the Care Team Workload Analysis (CTWLA). In this context, activity refers to all the duties, behaviours, and activities completed by a "job category/discipline" within a specified time frame. This is, again, a retrospective measure and describes "work done or completed." So the term refers to all the various activities that we perform: all the appointment duties + the non-appointment duties + the non-patient-care work. Practices commonly confuse the usage and meaning of the term and attempt to populate the spreadsheet with all the activities completed. A better tool for capturing this view is the CTWLA.

The various uses of this term can also be viewed in the context of the horizontal vs. vertical views discussed in more detail later. "Activity", as part of DSA, is a measure of appointment work viewed along the horizontal patient flow. The wider use of the term represents a vertical view: a slice through the horizontal flow that captures not only the appointment work but the competing non-patient-care work + the non-appointment (patient care) work. These three categories of work are measured as time. The sum of the three = the total time spent. Activity in the vertical perspective is thus measured as time. Activity in the horizontal perspective should be measured as appointment units (as bodies or appointment slots). One view is horizontal, the other vertical.

Fueled by the common and long held belief within healthcare culture that demand is insatiable and that demand exceeds capacity, when we measure or think about change, we inevitably want to count everything as demand in order to fulfill this theory. You may have a difficult time separating all the various forms of work into useful categories of non-patient care and patient care which can be sub-

categorized into appointment work and associated non-appointment work. While the total amount (sum) of supply must equal the total amount (sum) of demand, these categories are useful since improvement within each category requires different change ideas. When we blend the categories it is difficult to develop effective change ideas and even more difficult to see if these ideas actually resulted improvement.

It has been our experience that practices commonly want to blend all these categories, enter into the spreadsheets and measure together. Part of this error grows from confusion and interpretation of the term "activity."

## Hiding supply

Almost universally, practices and programs hide supply, whether unintentionally or not. Sometimes supply is hidden from all external view, while other times it is hidden internally from each other, or both. These behaviors grow from our fear, the fear of too much work. This fear is reinforced by assumptions based on an inadequate understanding and measure of our basic system performance. Hiding supply is primarily a self-protective behavior in response to a stressful, uncontrolled environment fueled by lack of communication and shared vision, commonly resulting in mistrust of the system. This is a sensitive area of discussion that is often a “taboo” subject amongst many of the groups we have worked with. There is no blame or judgment, but rather a necessary step in the understanding and declaration of our true supply. Many groups are simply not formally or operationally interconnected, and in particular, they are not connected by measures, by expectations, or in some cases, by leadership. When a "Central Triage" mechanism sends work to the first available open slot amongst a multitude of competing receivers, each of the receivers hides supply in order to not be the "fool" that gets the perceived (or actual) overload of work. Workers live in fear of being overwhelmed in an environment where they feel loss of control. With poor measurement and nonexistent accountability for each receiver to do "their part," this impasse is self-perpetuating.

Internally, as sets of interchangeable providers, we fear an inadequate and unfair distribution of work, based on availability alone, and so we commonly self-determine when we are willing to see patients. As providers we are often also "allowed" to negotiate return visits rates or our own availability. Often providers in an attempt to control their workload, will then either spin return visits to fill the schedules and become unavailable, or frankly, just delay ”releasing” any time to the schedulers. In some cases, if we are self-reflective as clinicians, we can see that many of us become caught in the cycle of what we call “loving the one we’re with”. It is easier to continue to see the same patients over and over again than it is to begin a new, sometimes complex case. However, this puts into question the value to the client over time or the rate of diminishing returns and the value to the system as new patients cannot be added to the caseload without the discharging of completed cases. In addition, many practices and clinicians do not have processes for clearly defining clinical objectives, and therefore have less than clear discharge criteria, which add to the ability to continue to justify the churning of appointments.

© Mark Murray and Associates & Alberta AIM 

Over time, these behaviors in many cases have come to be expected and permitted as the status quo. This distorted “releasing” of individual supply then begins to dictate the amount of demand the system can meet based on what each worker “feels” they can do. These behaviors cloud the ability to measure (we cannot see the TNA, for example when schedules remain undeclared or hidden), to monitor performance and to improve. These problems are magnified in programs where the supply is bundled. In these situations, the patients’ delay is determined by the least available of the bundle. Thus, individual, often self-protective and self-serving behaviors, affect the performance of the bundle and of the program.

Seeing how this practice of hiding supply adversely affects the ability to clearly assess our system performance is often the first step prior to or in parallel to system performance mapping and refining measures.

## Mapping

We cannot improve unless we measure, and due to the complexity of these systems, we cannot measure unless we map. Because of our self focus, our emphasis on activity as "all that we do," our longstanding cultural belief that we are without question under-resourced, and our belief that by working and simply trying harder we can do more and better, most of the practices and programs initially choose to use “cross-functional" maps to view their systems. A cross-functional or lane map illustrates the simultaneous actions of various workers and shows the handoffs of work from one discipline to another. The cross-functional maps are really just an illustration of “what” we do. While this type of map has some value, they are just too self-focused for us to be able to see “how well” our system performs. These maps follow actions over time but don't follow demand as it collides with supply. They provide a view of what the practice does by following task steps but don't follow the demand or patient journey. The "product," the action or issue described within the boxes on the map, changes (whereas, in a system performance flow map, the “product” is easily identified as the patient journey). Since the maps are designed to look at worker action and behavior there is no consistent product in the boxes or consistent "subject" followed through and across a journey. As a consequence, while these maps show action, they are not designed to provide a template for system performance assessment or measurement. They seem designed more to prove how busy the workers are, rather than to "see" how the work flows. In essence, the cross-functional lane maps follow tasks in sequence while the demand-supply (system performance) maps follow demand in sequence. Many practices and have challenges switching to demand-supply maps that follow the same product (the patient) through a series of steps over time. This switch is really again a matter of perspective. When we get a map, we typically draw a line from one step to the next. We draw that line along the patient demand journey. If we cannot measure each arrow from one step to the next as DSA and TNA for the same product flowing across these steps then we know that this map is inadequate in showing system performance.

## Levels of Maps

We suppose these are not really levels but a sequence. The demand traverses each "level" in turn.[2]

### Map 1

The first map should outline the steps from receiving the referral to an appointment being made. This is the "intake process."

Demand enters the system, commonly as a "request". This request is not synonymous with demand since some of the requests are "accepted" while other requests are "not accepted." There is a dilution here from request to appointment. Demand is only demand if it is met by supply. "Not accepted" requests are not met by supply. The confusion here may lie in the fact that for many groups the triage function (deciding who is accepted and who is not) is performed by the same supply people that will manage the appointment, so they see this all as demand, despite the fact that some of the work is indeed appointment work but the rest is non-appointment work. The triage work is either non-patient-care work or some soft form of patient-care but non-appointment work. As such, the demand and supply for the appointment function and the non-appointment work need to be evaluated separately.

In many intake processes the work stops at "waitlist" or some other form of holding tank. The "waitlist" is contained within the intake, prior to any appointment-made step. If there is a holding tank at that point, we know there is a choke point at the "make appointment" step. This choke point exists because there are simply no appointments available. This lack of availability can be due to an extended backlog, or the behaviours associated with hiding supply. Some practices artificially create a holding tank in an attempt to reduce no-shows. They don't release appointments until they are within two weeks of that appointment date. This action artificially may give the appearance of a steady two-week delay for an appointment since the step from appointment made to appointment seen is constantly two weeks, but the delays in the holding tank are not taken into account. The sum of delays for all these steps represents a delay for the patient. The waxing and waning of the number of patients in the "wait list" shows the variability of the delay. However, this is often hidden.

This delay is very difficult to measure. If there are no appointments there is no way to assess TNA. TNA and DSA—both critical measures of system performance—can only be evaluated when the appointment is made, or if there are appointments available. This "measure point" is the last step in the intake process. In addition, all the steps in the patient intake process, while they exist as delay for the patient, are hidden delays and can only be measured retrospectively as "actual waits." When we measure actual delay we have to measure retrospectively—put tags at the steps, wait for the process to end, and then look at how long the process took. This is a cumbersome method of measuring and makes improvement

[2] See “Five Levels of Mapping” – Mark Murray and Associates http://www.albertaaim.ca/articles.html

© Mark Murray and Associates & Alberta AIM

difficult, since any change we make cannot be evaluated for effect—for improvement—until the entire process is completed. This takes too long. This is in contrast to the measure of delay as TNA since we can assess the value of the change almost immediately.

So how then do we address the issues of streamlining, reducing or eliminating this cumbersome intake process? Once the process is mapped out it allows for the ability to examine the need or value of these steps. In the vast majority of cases these intake processes are accommodation strategies that have been created over time to address inefficient systems. If we were to eliminate the backlog most of the steps in this process would become unnecessary. Valuable clinician capacity is being utilized in this process that should be utilized for direct patient care.

We should also begin to question our referral processes and service agreements with the points of referral – typically primary care. Are we being clear to our referral sources as to what types of clinical cases we will accept, what our inclusion/exclusion criteria is, and what documentation and clinical information we require? In many cases, a large part of the intake process is taken up by seeking out information that could have been attached to the original referral. This then becomes a recovery strategy for a poorly designed referral system. In some cases, so much time has lapsed from the time of the initial referral to the point at which it is reviewed by the practice that certain clinical tests or assessments must be repeated. This is a direct result of system delay or backlog. In many cases an extremely large percentage of patients are accepted for treatment after going through the so called “intake” process. If this is the case, then it again begs the question of whether the process is necessary for the majority, and if we had a backlog free system, could we simply directly admit the vast majority right into the treatment stream?

## Map 2

The second map is usually pretty simple: appointment made to appointment seen. At this point, we can measure the demand (appointments given) the supply, the activity, and the delay as TNA. However, on occasion, this two-step process is interrupted by the requirement for the patient to get a test or undergo some other form of "inspection." This "other step" may cause another delay if the availability for that inspection is delayed. If this occurs, it should be mapped and measured.

The arrow from sender to receiver needs to be measured as TNA, DSA, and caseload. The demand moves across this step as a new patient visit. The method of distribution of this new patient care appointment work (by popularity, by availability, by pooling, etc.) needs to be evaluated. The receiver can be an individual, a group visit or some form of “multi-disciplinary team" (dyad, triad, tetrad, etc.).

© Mark Murray and Associates & Alberta AIM

## Map 3

In this map, we follow the patient demand as that demand encounters various individuals or disciplines as it traverses the SC or program system of care. In some practices where the demand from PC enters and encounters one provider type for new visits and recycled return visits, and there are no handoffs, this is a relatively easy map. In these cases, the second map then provides an adequate view of performance and a template for measurement and level 3 mapping may not be necessary.

In other practices, and more commonly in SC and programs, there are multiple between-day or within-day handoffs between provider types. This map is far more complex and a level 3 map is absolutely necessary. This is not a map of the visit itself (fourth level) but a map of the system and its handoffs. Even if there are within-day handoffs, this map would not capture the reception step at the onset of the visit. This map is intended to be a higher level than that specific step or process. The map starts at the initial receiver of the new patient appointment work and then follows that workload (demand) if it is handed off within that SC practice or within the program.

Some SC practices and programs do hand the work from one provider to another provider within that practice. This handoff is usually from one provider type to another (i.e., RN to MD to OT to PT). The term "multi-disciplinary team" (MDT) or "team approach" can be used as a euphemism to indicate that there are indeed handoffs. These handoffs can be within the same day (or two), which acts like a single visit with linked red zones, or the handoffs can occur between days.

The within-day handoffs act as a single visit with interlocking red zones in sequence. The handoff delay here is commonly brief. In some practices with within-day handoffs, the initial patient contact is accomplished by an RN who then hands the patient off to herself plus an MD (i.e., some Heart Function Clinics act like this). This scenario acts as a double or split red zone visit that requires a single provider (RN) and then a dyad (two providers: RN + MD). Other practices who utilize within-day handoffs will hand the work from one discipline to another in a series of sequential steps (i.e., RN to RD to OT to PT). Some practices use the same sequence each time while others plan the different sequences based on different individual patient requirements.

Since these visits are linked either as a simultaneous visit or a visit in sequence, the within-day (or two) handoffs act as a single MDT visit. These new initial visits to the MDT in some sequence may be followed by a recycled return appointment visit to the same team or may be followed by any number of return visit options on a subsequent day. These maps are identified as being 3a maps. The within day handoff has now morphed into a between-day handoff – a 3b map.

Practices and programs with "between-day" handoffs (3b maps) receive work as a new patient appointment visit and then either recycle the work as a return visit appointment to the same receiver and/or hand off the work to another provider type within the practice or program but on another day. These practices will also use the term "multi-disciplinary approach" or "multi-disciplinary team"(MDT).

Again, there are multiple variations here. The initial receiver can be a single person entity (part of the MDT) or can be a bundle of provider types (a dyad, triad, or tetrad). Hence the term MDT can refer to single provider type entities seeing patients in a between-day sequence or can refer to various combinations (bundles) of provider types seeing patients between days.

If MDT refers to a planned sequence of individual provider disciplines, there can be myriad combinations. If the MDT is a bundle, the MDT can act as a single consistent entity and "see" the initial new patient appointment visit and then "see" all the recycled return visits as well. In addition, the MDT, as bundle can act as the first visit which is followed by a series of handoffs to individuals or sub-combinations of the MDT. Alternatively, the MDT as bundle can act as the second visit, following an initial visit to an individual (usually an RN in a program). These secondary MDTs can always contain the same provider types (i.e., OT + PT + SLP + RN + MD) or can be constructed out of various components (i.e., OT + PT or MD + RN, etc.). In any case, when we map these handoffs, any arrow that shows movement from one entity to the next requires measures of TNA, DSA, and caseload, as well as an analysis of the type of distribution. Each arrow indicates a new patient appointment, which competes with the recycled return patient appointments for capacity. Once we add measures to the maps, we can see which pathways—which "trips,"—are more, and which are less frequent.

The variability has another layer: in some practices, but more commonly in some programs, not only are supply units bundled to meet individual patient demand, but demand units can be bundled to meet either individual supply or bundles of supply. Bundles of demand are commonly called groups, group visits, classes, or modules. These bundles of demand can be supported, again, by individual supply or bundles of supply. The maps and measures should capture this alternative.

At this point, particularly when we begin to map, we commonly see some confusion over the term "new." Once the patient enters the door of the SC practice or program and are seen initially, they are no longer by definition "new" to the practice or program. Sometimes this distinction is important for billing purposes. However, as the demand moves from one step to another within that practice, from one entity to the next, they are *operationally* "new" to the second receiver. That receiver requires distinct new and return appointment slots in order to manage this work. This is a vertical view of the system that is necessary to clarify the receiver's capacity to meet the demand. Therefore, if a patient sees an RN and then an MD in between-day sequence, that patient visit to each distinct entity is new. The spreadsheets are designed to capture this. Each step needs a spreadsheet.

Each bundle of provider types combined as an entity (dyad, triad, or tetrad) acts as its own entity and unit, and receives specific demand from the sender. As such, each entity needs its own distinct measures of demand to it, supply of it and at it, activity, and caseload. The standard Excel spreadsheets are designed to capture this data: demand for entity, whether single provider type or some combination, supply and activity of the same, as well as caseload. The key focus of the spreadsheet and the measure is to capture data for new. If the caseload is correct and the delay for new is within goal, the delays for returns visit appointments will be fine. The maps and associated measures will show

which pathways (trip plans) are the most commonly used (that is, where are the highways, the roadways and the paths within the flow system?).

In handoffs from single entities to single entities, the delay results from the relationship of demand to supply for that single entity. Pooling can load level the demand as it meets supply and accomplish a better relationship. The bundled entities (dyads, triads, and tetrads) have an added issue: the constraint. Each bundle can only proceed to meet the demand as quickly as the "least available" of the required supply bundle components. If one needed person or discipline is not available for that bundle, then the appointment visit cannot occur. For example, in a triad that combines RN + MD + OT, if the RN is available today, the MD in a week, and the OT a month from now, we cannot have a group or a bundle until a month from now. The availability of the OT, the least available, determines the delay. This issue is discussed in greater detail below, within the constraint section of this paper.

In order to capture how well the system performs, we need to see how well these handoffs function. What is the demand, supply and activity at each handoff? What is the delay and caseload? Practices will commonly get confused about measurement. Single individual providers and sets of interchangeable providers may act as individuals and as parts of multiple different entities (dyads, triads, tetrads, or even as support to group visits). These single individual providers are performing distinctly different functions. Each "function" (really a demand stream, met by some form of supply) needs its own performance measures. (See Horizontal Flow section above) Thus, individuals may be found in multiple different spreadsheets. A common mistake is to measure the individual, regardless of function, and as such individuals are mistakenly put on into a spreadsheet and the functions are blended. (See Assessing System Performance section above) We should be evaluating the performance of the functions, of the demand streams, regardless of the individuals performing these various functions. A related mistake is to enter all the various disciplines within the MDT on the same spreadsheet, i.e., all the provider types listed in the same spreadsheet, but they all perform different functions.

## Map 4

The level 4 map illustrates the series of steps and delays at the visit or encounter. The work enters the practice on the day of the visit and is seen. Components of the third-level map—the red zone or visit itself steps—are embedded within the fourth-level map. The third-level map illustrates a higher level view of handoffs and the fourth-level map is more granular – what actually happens on the ground at the day of the visit. For those practices with within-day handoffs, the entire third-level map is embedded in the center of the fourth-level map. The third-level maps just offer a different view—how the work traverses across the system—while the fourth-level map looks specifically at how the work traverses the visit on the day of the encounter.

In some practices, particularly where the handoffs are within-day, the visit and the cycle time of that visit are pretty well prescribed: the patient will spend the entire day at the practice. There are still delays, however, within that day, and these delays are captured by the cycle time measures.

### Map 5

These are very detailed maps of individual processes. These processes occur as steps in the patients' flow journey across the encounter (the reception/check-in process, the MOA process, etc.), and occur as processes that support that journey. For the physicians, those supporting processes include: documentation, answering phone calls, reviewing test results, refill of medications, management of referrals, etc. When these processes are mapped, measured, and changed towards improvement, the "efficiencies" gained can be used to improve the overall functioning of the practice.

Mapping is a crucial step in identifying opportunities to improve access and efficiency in SC and programs. We cannot improve unless we change. We cannot change unless we measure, and we cannot measure unless we map.

## Constraints

The constraint concept is reasonably simple, however, this is one that many teams and individuals struggle with. While the term constraint has a very specific meaning (e.g., the rate-limiting step, the rate-limiting supply component, the tightest tunnel, the choke point), and while the descriptions are clear (e.g., "we can only go as fast as the slowest step", "we can only proceed as fast as the least available or the most delayed of the supply components"), the term is often misinterpreted as much broader and more ambiguous in meaning (e.g., "we are constrained by lack of resources", "we are constrained by lack of leadership", etc). In addition, practices often make assumptions about the constraint without any measurement or analysis. Every process (using the term process in the broadest sense) in a flow system, where demand meets supply, will have a constraint. The front desk, reception process is constrained by the slowest of the tasks that make up that process. In hospital flow, while the patients' physiology should be the constraint, the length of stay is constrained by the slowest of the steps through that journey. In outpatient systems where patients are required by their specific journey to see a series of provider types or disciplines in sequences, either as within-day sequence or between sequences, the relative lack (the least available) of those disciplines will govern the rate of flow. That least available component will determine the capacity of the system to function.

Improving the non-constrained resources does not improve the flow at all, but in fact may make it worse. Let's take for example, a Heart Function Clinic where patients need to be seen by an RN and then, in sequence, by an MD. The visit lengths and return visit rates are the same but the MD devotes 0.5 FTE (half the clinic time) compared to the RNs 1.0 FTE. Let's say the team does a CTWLA focused on the RN and, as a result, shifts more of her non-patient care and non-appointment work time towards more direct patient appointment time. This does not improve the overall system flow and only serves to make it worse. In this case, the MD is the constraint. The MD has limited capacity in relation to the RN. So if the CTWLA was done at the RN level (a vertical slice through the RN step), the team may discover "wasted" RN time, hidden capacity or "opportunity" at that step. However, by increasing her capacity

and improving flow at the RN step, this can only result in pumping more work to the MD, increasing the mismatch there, and exacerbating the system flow constraint. Thus, identification and fixing the primary or major constraint is critical for flow improvement. Incidentally, this example provides a good illustration of some of the challenges of using Lean methods alone in system improvement. Lean is a great fixer, but unless coupled with the Theory of Constraints, we can fix the wrong problem and worsen the flow.

The constraint issue is critical. Once we determine the constraint, we can measure and assess the system capacity. The capacity of the system to perform or its velocity is governed by the constraint. Once we determine that capacity, we can work backwards across the equation to determine just how much demand this system can tolerate. Fixing the constraint involves some form of CTWLA at the constrained step, process, or person.

Constraints can be identified by instinct, by observation of where the demand (work) accumulates (where do patients wait?), or by mapping and measurement. The third option is best and provides the most accurate view leading to the most effective change ideas or PDSAs.

## Mapping to discover the constraint

1. Map the process or the flow of work and each demand stream. This map can be any of the levels of map outlined above. Each process, whether this is a high-level process of work flowing across a SC practice or program from one provider type to another or the process to check in a patient for a visit, has a constraint. The presence of a constraint in any process is universal. The arrow from one step to another in this flow map must follow the same "product," the same demand from one step to the next.

2. Look at how the supply or workers are allocated to meet the flow of work/demand. There is competition for the supply allocated to meet that demand stream. Supply devoted to appointments is a result of how much time is allocated to non-patient-care work or non-appointment work. The proportionate time used to support the appointment demand stream competes with the time spent in the other competing types of work: non-appointment work + non-patient-care work. In the case of the MDs, non-patient-care time competes with the two categories of patient care time (appointment work + non-appointment work). Identify the type of demand flow (the demand stream) and then identify the amount of supply allocated against that demand. Is there enough supply? If not, the work-demand accumulates.

   In a multiple clinician setting, all the clinicians may not be needed for all the demand streams. For example, in a setting where demand is divided by clinical condition, like Diabetes and Weight issues, and these conditions are met by various combinations of clinicians, one of those clinicians may be "needed" or "required" to support both demand streams (i.e., RD) whereas an Exercise Specialist may be needed or required only for the Weight Issue demand stream.

© Mark Murray and Associates & Alberta AIM 

Thus, we identify what workers are needed to meet the demand stream and then ask; how much of that worker is needed? Each demand stream functions as its own system and has to be evaluated independently. Therefore, we need to focus on the demand stream, not the person or discipline. If, for example, an MD works in the office and in the Operating Room, then that MD's time is allocated to two distinctly different demand streams and we need to focus on the MD as office and the MD as OR, and not blend them. The spreadsheets are designed to capture demand streams. If an MD functions in this dual way, the office spreadsheet should capture only the office time. Another example: if an RD supports the Weight and the Diabetes demand streams, then we need to measure and see that RD as separate allocated entities for each of these demand streams. The RD could, for example, be the constraint in one of the demand streams but not the constraint in the other.

3. When measures are added to the map for each demand stream, we can identify the longest delay. The longest delay indicates the greatest mismatch of demand and supply. The constrained step or component is the first step or component past that delay. However, we can sometimes be misguided here by variation. Highly variable processes where either demand or supply or both are highly variable will create long delays. Occasionally then, with high variability and a narrow measurement window, the longest delay may miss-identify the mismatch.

## Measurement to identify constraint

Each process will have a constraint. If the process flows from one step to the next, one of the steps will be the constraint. The process can proceed only as fast as that step. These are immutable facts. In some flow processes, the steps are processes themselves, like the flow of the patient across the office. In other flow processes, the steps are individuals. In programs or practices where the demand is met by individuals who manage the work until its conclusion, that individual clinician often acts as the system constraint. The work can only proceed to the level of that individual's capacity limit (caseload times visits per patient). In MDT systems where individuals are aligned in sequence, like the flow through a program where the patient sees a number of providers in between-day sequence, one of the providers will be the least available and be the constraint for the completion of that overall process. A constraint will also exist in "within-day" processes. The ideal and necessary way to discover the between-day or within-day constraint is by measurement.

Each set of multiple providers within a MDT will contain a constraint. This is the "least available" of the total group. For example, in a demand stream where all patients need to see an RN, an MD, and a PT for an appointment in some order or sequence and there are 4.0 FTE of RN devoted to this specific process (remember to subtract the non-patient-care work + the non-appointment work), there are 3.0 FTE of MD and 2.0 FTE of PT, then the 2.0 PT is the constraint for the process. The system can only tolerate what the most limited person—the PT—can provide. The excess work capacity of the other two provider types is wasted (although we often do not see that). This analysis, which focuses on the relative FTE, may need some of the following adjustments:

A. Isolate each demand stream. All clinicians do not support all the demand streams within a program or practice. Are all 3 disciplines required for this demand stream?

B. The proportion of FTE for that discipline within that demand stream (full- or part-time status devoted to demand stream). How much FTE for each discipline is devoted to this demand stream?

C. Visit frequency: Does each provider see the patients for the same frequency (new + return or patients x visits)? If the RN brings patients back twice as often as either the MD or PT, then the RN's appointment time is diluted by half and the relative availability of RN and PT is equal. They are both the constraint.

D. Visit-appointment length for either new or return visits or for both. If the PT appointments are half as long as the RN and the visit frequency remains the same, then they are equal in capacity and the MD is the constraint.

Here is another way to look at this: if each patient needs a 60-minute appointment with the MD and there are 4 hours of time to see patient appointments, then the MD can see 4 patients. We need one MD for 4 hours to see 4 patients. If all patients must see MD + PT + RN, and the patient visits with PT are 2 hours each, then we need 2 PTs for every one MD. If the RN requires 4 hours per visit and there is one RN, then we can only see one patient. To see the 4 patients that the MD could see we would need 4 RN's. All the other workers have "idle time" due to the "availability" of the RN. That availability results from the blend of visit rate, visit length and presence/absence of that provider type.

A third way to view this concept: If, for example, a patient needs RN, RD, and MD, and the RN is available in 2 days, the RD in 4 days, and the MD in 4 months, then the visit does not get completed until 4 months out. There are a number of reasons why that MD (or any clinician type for that matter) may not be available for 4 months: FTE status (part-time or full-time), time devoted to this demand stream, days worked supporting this demand stream vs. supporting the other competing streams, amount of time devoted to non-appointment work + non-patient-care work (which dilutes appointment time), ratio of new to return appointments (return visit rate), and appointment length.

The constraint determines the capacity of that system and thereby the speed or velocity at which the work or demand moves through it. Occasionally, the constraint is the number of exam rooms, room capacity (for group visits), tools, equipment, or non-clinical staff, but most commonly it is, and should be, one of the clinician types. The system can only flow as efficiently (tolerate as much demand) as the least available of the supply people. We want to compare the supply capacity of each of the supply components, and investigate which demand streams each clinician supports, how many days each of the clinicians work, how many visits they can see per day, the visit length (the sum of the visit lengths if the visit is a split red zone) and the return rate. In a sense, the "clinician as constraint" results from being the final common pathway of all the factors that dilute the total capacity of the clinician. Or in other words, from the following factors:

A. The need or requirement to support that demand stream or streams (i.e., do we need this person and how much do we need?).

B. FTE status (from full-time to part-time) (i.e., how much do we have?)

C. What part of the FTE is devoted to each specific demand stream (type of work) (i.e., how much do we get for each?, working in the clinic is different than working in the Operating Room or the hospital, etc.).

D. The division of work within the demand stream (patient appointment work + non-appointment patient work + non-patient-care work) (i.e., despite high FTE status, some clinicians have to spend relatively more time in support work than appointment work).

E. The categories within patient appointment work itself (i.e., the return visit rate, which dilutes the time that can be spent on new patients and appointment length). Appointment lengths for either new or return appointment or both can dilute the number of relative appointment visits supplied by each worker in the chain.

From these progressive dilutions each category of clinician worker and each individual within that category can be analyzed as the potential constraint. Once the constraint is identified, it determines the capacity of that system. The capacity determines the demand that can be tolerated and the velocity of the system.

It is important to remember that the constraint does not always remain constant. If a clinician is identified as the constraint within appointment flow, we can add supply to that constraint by moving that clinician out of non-appointment work towards a higher proportion of appointment work. However, if we then move another clinician in the same demand stream flow into the non-appointment work as a substitute, then that "moved" clinician can become the constraint in the appointment demand stream since the first was "relieved." There is a balance within and between demand streams and upset in that balance can shift the constraint. Creating overall balance of a system in order to achieve and maintain optimal performance requires careful and systematic trials of change and continuous analysis and adjustment of demand and supply balance at each step.

## Shifting to a Culture of Effective Improvement Strategies

Most of the practices and programs we come into contact with have, up until recently, just "accepted demand" and "hoped" that they could manage it without any clear view, analysis or understanding of their current state and the multiple variables affecting their system performance. When variation occurred—sometimes demand variation, but most commonly supply variation—that variation resulted in delays. Once the work started to accumulate over time as delay, the practices almost without exception resorted to triaging to sort out the demand as a form of perceived self-protection. But this triaging simply led to longer waits, wasted resources and system turbulence due to line-cutting and no-

shows. A secondary effect, but with no less impact, was increased worker and patient stress, dissatisfaction, and decreased levels of clinical care inherent to long wait times. In that environment, the knee-jerk and classic reaction is to ask for more resources which actually has the reverse effect of exacerbating the problems. Without clear measurement, mapping and analysis leading to an understanding of which resource is the most constrained, adding the wrong resource just adds to the waste and disillusionment of providers as they puzzle as to why there is no improvement despite these extra resources.

The challenge that we face in Specialty Care and programs is this essential shift in perception in the pursuit of turning toward positive, effective and sustainable change. As you engage in system improvements, your task as members of the team is to educate yourselves through the resources presented to obtain the knowledge of the principles of measurement, mapping and analysis in creative ways to meet the needs of each of your complex systems and to address the cultures which have developed as a result of lack of this essential insight. In a sense, many of us have been operating in the dark until now and it is our challenge to see a new vision of our systems in order to clearly know where to apply change.[3]

[3] Further reading: "System Design Choices in Specialty Care" - Mark Murray and Associates http://www.albertaaim.ca/articles.html