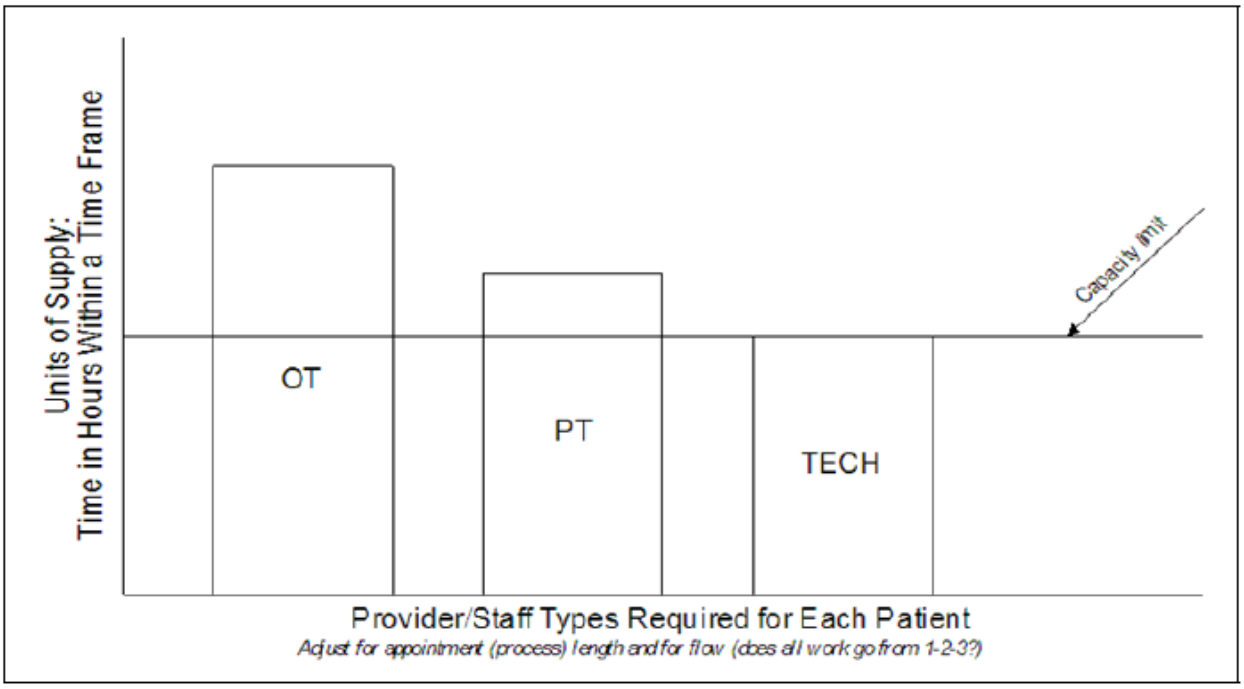
MMA
Mark Murray & Associates

Constraints

The constraint concept should not be difficult to grasp, but participants just struggle with this concept. While the term constraint has a very specific meaning: the rate limiting step, the rate limiting supply component, the tightest tunnel, the choke point, and while the descriptions (" we can only go as fast as the slowest step" or " we can only proceed as the least available or the most delayed of the supply components") are clear, the term is often misinterpreted as much broader and a more ambiguous meaning: "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 patient's 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 sequence, the relative lack of (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. For example, if patients need to see both an RN and then in sequence an MD in a Heart Function Clinic, the visit lengths and return visit rates are the same but the MD devotes half an FTE (half the clinic time) compared to the RN, using a Care Team Workload Analysis focused on the RN, and as a result, shifting non-patient care and non-appointment work time towards more direct patient appointment time will not improve the flow, but worsen it. In this case, the MD is the constraint. The MD has limited capacity in relation to the RN. If the CTWLA is done at the RN level (a vertical slice through the RN step), we may discover "wasted" RN time, hidden capacity or "opportunity" at that step. However, if we increase the capacity and improve flow at the RN step, this just pumps more work to the MD, increases the mismatch there and exacerbates the system flow constraint. Thus, identification and fixing the 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, then we can measure and assess the system capacity. The capacity of the system to perform is governed by the constraint. Once we determine that capacity then 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 a) instinct, b) observation of where the demand (work) accumulates (where do patients wait?), or c) mapping and measurement. The third option is best.



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 the map outlined above. Each process, whether it 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.

Thus, we identify what workers are needed to meet the demand stream and then 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 both Weight and 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. We can sometimes be misguided here though 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 misidentify 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. 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 also exists in the "within day" process. We have to discover the between-day or within-day constraint 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 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 that 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 time visits)? If the RN does bring 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 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 PT's 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.

And 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. 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, then, determines the capacity of that system. While occasionally the constraint is the exam rooms, the tools, the equipment, the staff but most commonly the constraint, and should be, one of the clinician types. The system can only flow, can only 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. We want to 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. The clinician constraint results, in a sense, as the final common pathway of all the factors that dilute the total capacity of the clinician: from need or requirement to support that demand stream (i.e. do we need this person and how much do we need), from FTE status (from full-time to part-time, i.e. how much do we have), to that part of the FTE devoted to each specific demand stream (type of work, i.e. how much do we get – working in the clinic is different that working in the Operating Room or the hospital, etc.), to 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), to 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 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 constraint. Once that constraint is identified, then that constraint determines the capacity of that system. The capacity determines the demand that can be tolerated.

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.

Most of the practices and programs have, up until recently, just "accepted demand" and hoped that they could manage it. When variation occurred (sometimes demand variation, but most commonly supply variation), it resulted in delays. Once the work started to accumulate as delay then the practices resorted to triage to sort out the demand as a form of perceived self-protection. But the triage led to longer waits, wasted resources and system turbulence due to line-cutting and no-shows. In that environment, the knee jerk reaction is to ask for more resources which actually just exacerbate the problems. Without measurement and understanding of which resource is the most constrained, adding the wrong resource just adds to the waste.