



System Performance Goals

“Every system is perfectly designed to get the results it gets”.

Deming

There is a fundamental tension that exists in all healthcare settings that often leads to serious and misguided consequences. This tension has been described as “the myth of 100% utilization” and is manifested in a number of ways. In the hospital operating room setting, for example, this tension is manifested by the false belief that the most efficient way to utilize operating rooms is to have a full schedule built in advance. This strategy would work perfectly toward optimizing system performance as measured by throughput (units completed per unit of time), efficiency (cost to revenue ratio) and effectiveness (outcomes) if demand for the OR was matched perfectly with all the supply components needed to successfully complete an OR procedure and if there was no variation in daily or weekly volume, no variation in arrival rates, and no variation in case time.

In the clinical outpatient setting this tension is manifested in a similar way by the mistaken belief that pre-filling the entire daily appointment schedule, either far in advance or spontaneously filling any unexpected no-show slot with a corresponding unexpected walk-in visit is, again, the most efficient way to work. This pre-filling, use 100% capacity belief occurs in all payment settings. In quintessential fee for service systems (“do more, get more”), the common belief is that the highest net revenues are generated when the schedule is filled in advance. Pre-filling offers the reassurance of continued income. In systems where visits are all paid at essentially the same rate, this myth also exists. The belief here is that the most efficient way to guarantee income, which is a direct result of volume of visits, is to keep the schedule filled. This occurs whether the organization is paid by the visit (Community Health Centers in the US) or the providers themselves are paid by the visit (the Canadian system). Counter-intuitively, this belief also exists in capitated environments where visits actually represent cost. In many of these environments, and in particular in those environments where the organization is paid by capitation but the providers are paid by salary, the belief is that the best way to monitor provider “productivity” is to keep the schedule full and to develop visit productivity standards. In these settings the goal - and this is a misguided goal - is to maximize net revenue and the strategy developed to achieve that goal is to make sure the schedules are filled to maximum capacity.

There are two fundamental confusions here. While the goal of revenue maximization is solid and understandable (“no margin, no mission”) this “goal” results from a fundamental misconception and, as such, ought to be viewed as a desired outcome, not a goal. Secondly, the action of filling schedules in advance in order to achieve this desired outcome is not the goal either but a strategy designed to achieve optimal net revenue.

In systems where the stated goal is to optimize revenue using a strategy of pre-filled schedules and 100% full capacity used, there is often, at the same time, a clear recognition that patients want to see the same provider (continuity), that satisfaction of patients and providers alike is

improved with that continuity, and that outcomes are improved with that continuity. In many setting there are mixed messages – see your own patients, work towards clinical goals and outcomes and, at the same time, do this while making sure that your schedule is filled to the “productivity standard.” This mixed message can be confusing to providers.

The basic disconnect here occurs due to a lack of understanding of the fundamental dynamic at play in flow systems: in flow systems where we match the demand to the supply, the most effective (outcome) the most efficient (cost to revenue ratio) and the most satisfying systems will match demand to supply without a delay. Successful flow systems focus improvement and goal efforts on the input side, on making the flow work, rather than focusing on the output side. Flow systems see output as an outcome and improvement in output as an inevitable consequence of smooth flow.

The tension between filling the schedule in advance and balancing demand and supply without a delay (the fundamental goal in flow systems) gets played out in a number of scenarios:

Managing visits vs. managing a panel

In the “manage visits” approach, the goal, which is measured in visits or uses visits as a surrogate measure for productivity, is to make sure that provider-visit productivity is high in relation to capacity. The goal is visits to full capacity. Success is measured by activity (what is done) that either meets or exceeds capacity (supply).

In order to achieve the goal of 100% full schedules, organizations and, in particular, providers, quickly learn that the most effective way to ensure a full schedule of visits is to have patients wait in a warehouse and then aliquot off a full schedule proportion each day. To some extent this strategy might be successful if there was no variation. Variation in volume of demand, variation in “urgency” of demand, and some variation in arrivals (no-shows and walk-ins) and variation in supply make this strategy problematic. These types of demand and supply variations characterize healthcare settings. Volume variation in demand is dealt with in two ways: first, by appointing those “over demand” patients to another provider’s “under demand” open schedule. This is done most often by appointing a walk-in from provider A’s panel to replace a no-show on provider B’s schedule when provider’s A’s schedule is considered full. Secondly, demand volume variation is dealt with by sending demand volume deeper into the wait time queue. These two actions lead to serious system performance deterioration:

- Sending patients deeper into the waiting time increases the warehouse or backlog of work. This backlog raises cost:
 - ❑ The longer the wait, the higher the cost per visit. This affects the overall net practice revenue.
 - ❑ The longer the wait, the higher the no-show rate. No-shows cost staff time and result in unused capacity. That unused capacity is often, as mentioned previously, filled by intentionally over scheduling or by using walk-ins as indiscriminate random schedule filler.
 - ❑ The longer the wait the more re-work and redundancy. For example, studies in call centers have shown as much as a tenfold increase in call handling time in systems with full schedules and no appointments to offer.

- ❑ Because of the variation in demand urgency, the longer the wait, the more the need for triage to determine who can wait and who cannot. Triage uses up precious professional resource.
 - ❑ The longer the wait the higher the number of walk-ins. Patients walking in increase the return visit rate by shorter visit lengths and reducing continuity. Walk-in patients also reduce patient satisfaction by increasing office cycle times.
- Sending patients from the linked provider to another provider also has consequences:
 - ❑ With discontinuity, we get reduced satisfaction, even when patients “agree” to see the non linked provider.
 - ❑ With discontinuity we lower the revenue per visit and per minute worked.
 - ❑ With discontinuity we adversely affect clinical care and outcomes. Studies on clinical care and outcomes demonstrate that care and outcomes are improved, and can only hope to be optimized, with improved continuity.
 - ❑ With discontinuity the visit length is extended due to the time necessary to establish rapport, credibility and obtain a history required in the discontinuity visit and not required in the continuity visit. As a consequence, with visit length increase, the number of visits in a day is reduced. So counter-intuitively, trying to fill schedules with unlinked patients in order to ensure high visit rates actually contributes to lowered clinic visit capacity due to longer visit lengths.
 - ❑ With higher discontinuity we get a higher number of return visits since the usual behavior when patients see a non-linked provider is for that provider to send patients back to their own provider. While this “system churn” may not seem all that bad in a pay by the visit environment, it leads to lower RVU per both visits in the quintessential FFS system and in a pay by the visit environment, this fills the schedule with less than fully useful visits, prevents panel size growth from outside the current practice, and precludes opportunity to complete unmet clinical care for current patients.

On the other hand, with a focus on eliminating delay as the overall goal for improvement in flow, a key recognition is that balancing demand to supply not only has to occur at the practice level but more importantly at the individual provider level. Continuity is critical: with better continuity we get better satisfaction, less cost, more revenue per minute and per visit and better clinical care and outcomes. Operationalizing continuity is often described as managing the panel. Managing the panel means balancing the demand workload from that panel to the individual provider supply and managing that demand to supply ratio with as short a wait as possible. For Primary Care “as short as possible” most often means balancing the daily demand with the daily supply, that is, “doing all of today’s work today.” Managing the panel then shifts the work focus to the input side - to the panel and away from the output side - the visits. Variation has a tremendous effect here. In systems focused on output (visits), the goal is to reduce variation in output, particularly to reduce down variation in output (the risk of unfilled appointment slots) and to always reach the productivity standard. So as the output (visits) becomes fixed or set as a goal, the effect of the variation is borne in the number of internal deflections or deflections of workload over the production visit standard into the waiting queue. In effect then the customer absorbs the variation either by being put deeper into the wait time or by being directed away from the linked provider to another provider in order to fill an appointment slot. The

consequences of that are outlined above. In a sense then, the goal of visit-focused systems is to create “output equity”- equal or fair outputs as measured by equal and standard visits for all providers. This goal is most often seen in environments where the providers are on salary. With a fixed salary the goal is often to fix a production standard against that fixed salary. At the same time, we often see this goal expressed in capitated organizational environments as well where the providers are salaried. This approach constitutes a huge disconnect in incentive and subsequent behaviors between the organization and the salaried providers.

On the other hand, when the goal and incentive is shifted away from “output equity” to “input equity,” that is, panel sizes correlated to time worked, aligned with the twin goals of no delays and continuity (see your own and don’t make them wait), then organizational and provider incentives are aligned and the consequences are far different. Excess visits are reduced, costs are lowered, satisfaction and net revenue rises and outcomes are improved. This shift though from a focus on output and visits to input in panels shifts the burden of variation from the customer having to absorb variation to the individual provider having to absorb the variation. For example, with up or down variation in demand volume, and with a strict prescription to see your own and don’t make them wait, then on some days, volume of demand and hence activity rises and other days, it falls. Of interest here though, is that studies of visit activity in output visit-focused systems show that due to the variation in the ratio of no-shows to walk-ins that the providers will often see more patients on some days and less on others. So while there is great fear of the effects of providers absorbing the variation (and the fears are off too much or too little daily work), this variation is often of less range than what is seen in output-focused situations.

As managing the panel shifts the goal from filling the schedule with visits towards a goal of balancing demand and supply and working without a delay, the visits then become an outcome, not the goal. The panel size drives the demand and the activity. So panel can be adjusted to achieve visits as inevitable outcome. The goal for optimization in any flow system is to balance the demand and supply, and work without a wait. Reducing variation is a key strategy to achieve this goal. Systems that make the goal an output goal, based on visit or RVU suffer from a serious and often fatal misreading of the basic fundamental dynamic at play here. Thus production, visit and output goals are not a reasonable option but, in reality, represent a fundamental misreading of the dynamic.

There are a lot of semantic scenarios with which to view this dichotomy: managing visits versus managing a panel, output equity versus input equity, the question of who absorbs the variation (patients or providers), or goal versus outcome. At the end though, no matter what we call the dichotomy, the final conclusion is the same: when we focus on visits or production as the goal we create cost (cost per visit and system disturbance) that overcomes any risk of an unused appointment slot that we might see when we focus on input and panel management. In demand and supply systems (and make no mistake about this, this is what we do in our healthcare systems - we match demand to supply), to achieve optimization of system performance we have to focus on flow. The outcomes in visits, in productivity, and in net revenue will inevitably follow.