



## Myth of the Full Schedule

### ...the fear of too much demand and the fear of too little

The fear of going to work without a full schedule is real. Some groups fear too much demand and an overwhelmed capacity, while others fear not enough demand and unused capacity. These fears focus on too much or too little in volume of demand as well as too much or too little demand due to variation - temporary mismatches of demand due to daily variations.

#### The fear of too much demand

Some practices fear too much demand and fear that when they get the wait time to zero and start the day with unused capacity, demand will overwhelm them and they simply will not be able to keep up. Their fear is that each day's demand will overwhelm them.

To overcome these fears, these groups need to:

- measure practice and individual panel size
- calculate last year's visit rate and determine with current performance (panel X return visit rate) whether they can balance that demand with their current supply (provider days per year X patients per provider per day). If, in this equation, demand exceeds supply, these groups need to use strategies to reduce demand per patient per year or increase provider visits per day to achieve a balanced equation.
- by using data and measurement, reduce demand and supply variation and commit to flexing supply as needed in order to maintain daily balance with demand.

#### The fear of too little demand

On the other hand, practices primarily where reimbursement is the same for each visit, also fear working without a warehouse full of scheduled work but for an entirely different reason. This fear focuses on the risk that working without a delay (a TNA of zero), and without the buffer of the warehouse, may result in unused system capacity and that unused capacity will result in sub-optimized revenues. At the same time, there exists in parallel the false belief that a warehouse of work (delayed work waiting to be processed) enhances practice revenues, because it provides reassurance against any down variation in demand. For example, if there is a minimized waiting time and the demand goes down, there is a risk of unused capacity. The myth is that a warehouse of demand will prevent or mitigate that risk.

The underlying issue here is the issue of variation. If patients became ill every 15 minutes, traveled for 15 minutes to see us and were seen every 15 minutes and that happened 25 times each and every day, we would have no problem. But demand variation does exist in healthcare. There is variation in daily volume, variation in arrival times and variation in handling times. In addition, there is actually supply variation. Each day we have varying amounts of appointment

supply. If a practice commits to working without a wait, that is, maintaining a short and fixed waiting time (completing all of today's work today) then variation in either demand or supply will result in either activity (workload) that exceeds demand (up demand) or in activity less than demand (down demand).

Thus, if the practice commits to a high service level, that is, minimal waits, because of variation the practice will have some days of high activity and some days of lower activity in relation to the mean (average) activity. On the other hand, if the practice commits to a fixed workload and uses a pre-booked full schedule to try to achieve this, then service levels or wait times will vary.

Practices that fear the risk of unused capacity due to down variation and working with a minimal wait will use a backlog of work and a full schedule to provide reassurance against that risk. Their belief is that systems work most efficiently when capacity is pre-filled and utilization of that capacity is always at 100%. This belief is incorrect. We call it the "myth of 100% utilization." The strategy of having a delay in order to ensure a full schedule simply does not work.

Let's think this through logically:

1. Visualize the wait time like a lake of work. If the height of the lake is stable, then the workload going in (demand) = the workload done (supply used or activity) This lake of work is expensive: higher no shows with longer waits, more need to use resource to "triage"(sort) out work when there is a delay and more staff work, frustration and redundancy when there is a wait time.
2. As the height of the lake goes down with backlog reduction (doing more work than presents itself - activity > demand) then the height of the lake gets closer to the bottom of the lake.
3. At some point when all of today's work is completed today, the lake of work is virtually eliminated except for the puddles of good backlog.
4. We know, though, that the panel (either for the practice or for individual providers) will create demand each day. So when we get to tomorrow, there will be demand. Sometimes that demand will fill the schedule perfectly; sometimes it will over fill the schedule (up demand variation) and sometimes the demand will not fill the schedule (down demand variation).
5. Not only is there demand variation but there is supply variation (the size of the lake bed changes). Some days we have more providers and some days less, so our lake capacity changes each day.
6. So we have not only demand variation but supply variation as well. The rate at which the lake fills then is dependent on those two variables: demand variation and supply variation. We can influence the rate of the lake filling by:
  - a. predicting, through measurement, the amount of external demand
  - b. manipulating the "good backlog" (internal demand) by bringing patients in early in the day, later in the week, and by scheduling less returns on days when there are fewer providers Thus leaving more space for the predicted external demand each day.

- c. being flexible in our supply, that is, seeing more patients when the demand is up or supply is down. Keep in mind that we already do this with the ratio of walk-ins to no shows.

Reducing backlog and doing the work today (without a lake) for the most part does not change demand. If demand = supply you are just doing the same work but closer to the bottom of the lake rather than at a distance.

There is a choice here and you can measure the consequences of the choice.

1. You can work with a backlog - a lake of work that fills the schedules in advance. The consequence is that:
  - patients are dissatisfied
  - staff are dissatisfied, act out, and get frustrated
  - providers are dissatisfied
  - patients are shifted to non-linked providers since their own provider is full and the non-linked provider has an opening
  - the cost of care for the practice rises (rework, redundancy, cost of triage, high no shows and poor office flow),
  - the revenues go down (with longer visit lengths due to unhappy patients and high discontinuity)
  - clinical care suffers due to rushed work and poor continuity.

This is the choice discussed above - the choice to keep schedules full in advance in an attempt to guarantee no unused capacity. But system performance (high no shows, etc.) and provider behaviors (self protection) actually result in highly variable utilization, some days in activity (high walk ins and few no shows), and some days in low activity (low walk-ins and high no shows)

2. You can work without a backlog and take the risk of unfilled appointment slots. But, patients are happier, staff is happier, providers are "happier," the cost of care goes down, revenue rises and clinical care improves. Variation in activity does exist but, in our experience, the range of variation is no greater than the range in Alternative 1 above.

In order to see what "works best," measure the cost of care in Alternative 1, compared to the cost of care in Alternative 2.

With shorter visit lengths due to better continuity, net revenues can be the same with fewer visits in Alternative 2. In addition, with Alternative 2, if the practice absorbs variation, some day's schedules will over fill and some days will under fill. Look at the activity at the end of the day, not the beginning. Somehow we are reassured if the schedule is full at the beginning of the day and ignore the fact that often that schedule is not filled at the end due to high no shows. This is acceptable, I suppose, because we could not control it. But we get nervous if the schedule is not

filled at the beginning of the day and remains unfilled at the end of the day. This, somehow, is "our fault" whereas the initial scenario is not. This is not about fault or a false sense of under control. Look at the activity (the supply used) over time, not one day at a time. Alternative 2 will result in less cost and, in quintessential fee for service (do more, get more) and in capitated payment systems, will result in higher net revenues. In pay by the visit systems, there will be fewer visits in Alternative 2. This is due to the reduction of the waste of discontinuity. With better continuity and the incentives that grow from providers seeing their own and not making them wait (incentives to do more with each visit, to extend visit intervals, to use teams, etc.) will result in less visits per patient per year. At the same time this reduction in visits per patient per year can be a survival strategy if panel size leads to demand that exceeds supply, or offers the opportunity to grow the practice. This growth can be accomplished in two ways: external growth (new business with new patients) or internal growth (discovering the current patients with unmet clinical needs and through visits, managing those needs).

The key to all of this is measurement. What is the practice and individual panel size? What is the predicted visit rate? (We can get last year's rate quite easily). This tells us whether we have enough patients to fill the lake. Second, look at the work over time. If we look at a day at a time and the day overfills or under fills, we can easily get panicked about isolated events. Third, measure the net revenue. We often forget the costs associated with Alternative 1. Fourth, learn how to manage variation so that the workload is not too much or too little. This requires measuring, predicting from past activity, planning and being flexible. This also requires the avoidance of over concern the workload drops on a particular day.

The fear of working without a wait thus has a common origin - knowledge that if a practice commits to working with a minimal wait, demand or supply variation can result in too much or too daily work. Meticulous contingency planning can mitigate demand and supply variation. For example, bringing planned return visits back early in the day, late in the week, changing the ratio of internal to external demand expectations based on the number of providers present and sharing the work of the absent providers, are all strategies that level workloads amongst days and between providers. Provider supply can be planned in advance and can be flexed as needed.