MMA
Mark Murray & Associates

Five Day Access Goal

The goal in "access improvement" is to minimize the delays for patients requesting an appointment. Since continuity is king, the goal is to minimize delays to each individual provider: "see their own and don't make them wait".

The goal is not specifically "same day access". We could achieve "same day access", as some do, by the "urgent care" approach—that is, by "access by denial". Refusing to allow patients to make any future appointments achieves same day access. However, that approach is costly, risky and dissatisfying.

There is often an assumption that zero delays are great, 30 days are bad and anything in between is pretty good or better. There is a sentiment that if we cannot achieve a minimal delay, let's at least "get better". It almost seems like an excuse not to attempt to achieve what's possible – similar to, if we cannot achieve 100% mammography rates, let's just set a goal for what we think is "reasonable" or what we think we might be able to achieve. The goal in clinical care is always perfection. The goal in operational care is perfection as well. Each system, given its demand and supply dynamic which includes the issue of part-time supply, has an operational optimal performance: if a system has demand that exceeds supply, it will fail. If a system has a demand supply balance but part-time supply, there will be delays. We can determine the optimal level at which this system can perform given the absence of supply. That's what we need to do, not arbitrarily decide that zero is great and 5 days is pretty good.

There are both qualitative and quantitative issues in play here. There is a quantitative difference between 30, 5 and zero days. At the same time, there are significant qualitative differences: the systems we use to achieve these results are remarkably different.

A system with a goal of a 30-day delay will have schedules filled for 30 days. Newly generated demand will be sent out 30 days to the end of the line. If the system achieves this goal and maintains it, which is what it says it plans to do, then this system is in balance: demand does indeed = supply but that system allows and plans for a 30-day wait. There is really no purpose in that delay and a great deal of waste. If demand was uniform and demand could wait, then this system could work if it could tolerate the waste created by the 30-day inventory. In healthcare, demand is not uniform: some patients can wait and some cannot due to the clinical condition. The 30-day systems, fail but try to accommodate by triaging the urgent demand work either by sending to other venues, by over-booking or by trying to find some hidden space.

If a system is designed for a 5-day delay, most of the above problems remain. Demand is still not uniform, it is still variable. A system working with a 5-day margin though, has some other problems. How do we maintain a line with only a 5-day delay when demand just keeps filling the future schedules? If demand is not triaged into urgent and non-urgent, and if future appointments are allowed, the non-urgent demand will fill the schedules past 5 days and this goal is unachievable. Five-day systems could achieve a 5-day goal by a form of access by denial, but this is not a good plan. Thus, 5-day delay systems have to accommodate the urgent issue. They do this by a carve-out – that is, separate the urgent demand from the non-urgent demand and set a 5-day delay target for the urgent demand. However, some demand is more

© 2010 Mark Murray and Associates

urgent than 5 days; what do we do with this demand? How is it seen? We could have 3 lines: one for emergent same day, another for urgent within 5 days and one for all the others. More appointment types restrict the channels, and variation has a more significant affect. Each provider then would need three lines or 3 appointment types.

To those working in a 5-day system or to patients living in such systems, the 5-day delay feels the same as a 10-day, or a 30-day wait. The providers and the staff come into “today” filled, even if 5 days from now has some availability. But to come into “today” with enough space (either individually for those who are closer to full-time, or as a team for any who are very part-time) to deal with any request is a significant difference. Gone are the days of triage, of adding in, doubling up, staying late, etc. With space available today, you can do anything, including offering appointments today or giving appointments in the future if the patient prefers.

The goal is always to optimize the flow. The outcome may not be zero days. We can determine the optimum performance and then set that as a goal. If we pick some random goal beyond optimum because it seems "reasonable", there is no system design that can achieve that goal. Long delay (30) or middle delay (5) goals may accommodate high frequency of supply absence but cannot perform well due to high demand variation. The only way to achieve these long or middle goals is to perform better than the goal most of the time, and if demand rises we have a built-in surge capacity (panels designed at a 90% demand to supply ratio). If we have, for example, a 5-day goal, in order to achieve that we would have to perform less than 5 days most of the time. Most people do not understand that.

Some of the problem here might be a result of confusion over aim and expected outcome. If we have an aim of 5 days, then operationally this is very difficult to design. Actually, due to the levels of demand and supply variation, it is impossible to design and to achieve. The 5-day aim just seems to have been chosen arbitrarily. If demand and supply are balanced through the right panel size and if the providers fill all 10 sessions per week, it is possible to achieve a zero delay by matching the demand each day with the supply each day. But because of the natural variation that we see with demand, even if all providers worked all 10 sessions each week, achieving and maintaining a minimal delay of zero days is a challenge. Providers need the right panel size and a commitment to flex capacity each day since demand will exhibit its natural variation. If demand rises, we need to see more; if demand goes down, we need to see less. Open slots are not a gift, but a debt. In addition, we need to use the return visits to load level: "sell early, sell late", see the entire future daily – weekly schedule when we pre-book future appointments , set threshold limits (not appointment types) and signals for both pre-booking and when thresholds for future booking limits are breached and plan for vacations and other supply absences.

If we add in the chosen environmental complexity of high frequency of supply absence, that is, supply present even less than the 10 sessions per week (and absences occur even in 10 session per week practices due to vacation, etc.), we will always incur delays, and even more planning is required. If the supply has a high frequency of absence, we need the right proportionate panel and to utilize all the contingency planning strategies. Each system, though, will have a measurable optimum performance level. If, for example, a provider is present for two days a week, Tuesday and Thursday, and delay is measured as TNA each day of the week, potentially the delay could be 1, 0, 1, 0 or 4 days if we measured each day starting on Monday to Friday. The best average this system could achieve, if measured 5 days a week would be 1.2 days. If we measure on one specific day or if the days present is different, the "best measure" would change. There is thus a performance limit for each system. While the goal remains "to

© 2010 Mark Murray and Associates

minimize delays", the expected outcome can be no better than 1.2 days. The system design outcome "expects" delays due to the supply absence.

While in systems where there is a high frequency of provider absence, we could achieve this optimum with a single appointment type and without carve outs. This would require constant vigilance with use of all the contingency plans. On the other hand, we could use a carve-out (i.e., save some visits each Tuesday and Thursday for this example provider). The usual recommended practice is to not "count" the saved slots in the TNA. The decision to use the carve-out has an influence on the measure. What we want is for the system to perform – the measures just help us see how well it performs. If we can imbed the philosophy of doing all the work each day and this behavior becomes a system property, to some degree the TNA measures are designed more for monitoring. We could consider the use of the future open capacity measure.

© 2010 Mark Murray and Associates