MMA
Mark Murray & Associates

Priority and Triage Systems

Priority or triage systems, where demand (workload) is segmented into parallel streams based on clinical “acuity,” are a misguided attempt to accommodate systems fraught with long delays. This approach does not solve the system delay problem, but, in fact, makes delays even worse.

Most priority schemes, using agreed upon clinical criteria, commonly segment demand into three or four steams, usually labeled as emergent, urgent, semi-urgent and routine. Each stream has well-developed clinical criteria for both inclusion and exclusion into that stream and has an expected acceptable time threshold: most systems of this sort set a goal of “today” for emergent, one week for urgent, commonly two to three weeks for semi-urgent and commonly four to six weeks for routine. While the clinical criteria reflect the fact that, indeed, some conditions are more pressing than others, these priority schemes simply cannot function well operationally.

If total system demand exceeds the system’s capacity to meet this demand, no system of priority or triage will work. If demand exceeds capacity, delays will result. These delays will either be in all or some of the priority lines and either all or some patient needs will be extended past the determined threshold. Patients may be seen, but some will be seen late and, once the lines extend long enough, some patients will not be seen at all due to the phenomenon of "balking and reneging.” Each day the mismatch accumulates, the delays lengthen, the practice gets further behind and there is no way to catch up. Creating priority queues does not solve or fix this mismatch but instead serves to ensure that the lowest priority patients may never be seen. Priority systems also require a lot of triage work, and consume unnecessary resources. Capacity is used to sort the work rather than do the work and the demand–supply mismatch only worsens. Occasionally, however, the triage/priority function serves to “eliminate” some of the demand in order to achieve a balance. If the demand can be eliminated so easily, we have to question whether Service Agreements would have been a better way to define the "appropriate" demand and there may have been no need to consume resources for the triage function in the first place.

On the other hand, most priority systems are not designed to eliminate or extinguish demand, (which actually can be done much more effectively through Service Agreements), but instead, assume that total demand and capacity are balanced but that some patients simply ought to be seen in front of others. If this assumption is correct, and total demand and capacity are balanced, then again there is no need to triage.

How are priority systems organized?

- Either three or four priority lanes are identified, all having inclusion and exclusion criteria, based on clinical need.
- Each lane has a designated delay threshold: emergent (within 1 day), urgent (within 1 week), semi-urgent (commonly within 2-3 weeks) and routine (commonly within 4-6 weeks).
- Practice or department capacity is allocated to support the demand in each of the lanes, that is, varying proportions of time is devoted each day or each week to meeting the demand that is prioritized into that lane. This allocation is commonly determined by instinct and far less commonly by data. For example, either 25% (or more or less) of the practice appointment capacity is devoted to each of these four lanes. These proportionate allocations are reflected in the schedule. The lanes are operationalized as appointment types. Either the lanes are distinct - one day or one person for semi-urgent etc., or the allocated proportions are reflected in the ratio of appointment types on all schedules.
- While the allocations of practice capacity to support each of these lanes are well intentioned, most practices exhibit highly variable commitment to actually providing consistent coverage for these lanes. Supply varies. The emergent lane is always covered but the other lanes exhibit highly variable coverage.
- Even if supply does not vary and the commitment is impeccable, the demand for each of these priority lanes does indeed vary. However, the priority assumption is that for the routine lane, for example, the amount of workload demand for the routine category generated today will be just enough to fill the schedule four to six weeks from now: each current day or week has the right amount of demand to fill the corresponding time period (a day or week) at some point in the future. In this way the threshold delay is not breached. But demand will vary. This variation will occur both between lanes and within lanes. For example, the demand across the lanes will not always mirror the amount of capacity allocated. Even if each lane gets 25% of the allocated supply, the demand will not arrive in the same proportion. Temporarily changing the triage criteria can overcome this arrival variation but changing the criteria just to fill the lanes to capacity has to call into question the validity of the “well developed” criteria. In addition, if the criteria remain consistent with the original intent, each lane will exhibit its own "within the lane" natural variation. Demand thus varies across the lanes (more or less than the 25%) and within the lanes (more or less than what is needed to fill the end of the line and not breach the threshold).
- In the “emergent” lane, natural variation is met by flexible capacity. If demand rises or falls the capacity flexes to meet and complete that demand each day.
- In the routine lane, a natural rise or fall in demand buffers itself through the time frame. If demand rises, the four to six week built-in delay serves as a buffer. The lane has time to recover from that rise or fall in demand and remains, if the allocation is correct, relatively stable. Some patients are seen beyond the threshold and some are seen shorter than the threshold but the “average” is closer to the threshold.

- The problem occurs in the urgent lane- the second sickest lane. If demand rises, even for a few days, with a delay threshold of a week, there is simply not enough time to recover. Small variations in demand result in a high likelihood of not meeting the threshold goal. In fact, in four-lane priority systems, the second lane is more often out of compliance (delayed past the agreed upon threshold) than within compliance. While excessive demand into the longest lane (routine) could be solved by moving some of this demand into the lanes with shorter thresholds, this opportunity does not exist for patients exceeding the threshold in the second sickest lane. The two “less sick” (longer threshold) lanes are already full past the shorter time threshold recommended for the second lane, so there is no capacity to “steal.”
- Failure in these systems is further exacerbated by the fact that despite all the work put into developing the right criteria for each of the four priority lanes, mistakes are made and some patients are put into the wrong lane.
- In addition, patients may be put into the correct lane initially but their condition either worsens or improves, and they are now occupying space in the wrong lane. It is very difficult and requires intense resource usage to move patients from one pre-determined lane to another. While it is particularly difficult to move patients from a less sick (longer wait) to a more sick (shorter wait) lane, it is virtually impossible to do the opposite. As a consequence, if a patient's condition improves they continue to occupy space in a sicker lane.

How can we fix this?

- Measure to ensure overall demand is balanced by overall capacity.
- Develop Service Agreements, particularly if measured demand exceeds capacity, in order to define the appropriate demand and shape demand to meet supply.
- Maintain the "triage-priority" for emergent clinical conditions. Make the inclusion and exclusion criteria simple. Most specialty care practices have a planned method to deal with this already - the "on-call" function. Continue to flex supply in this lane.
- Measure demand for new patients.
- Reduce the variation in supply so there is a planned, reliable predictable presence poised to meet expected demand.
- Eliminate the triage criteria and the priority for the other three lanes. Allocate the total capacity to meet any demand from any of the non-emergent categories. In this way, demand variation is buffered. An increase or decrease in demand in any of the three lanes is shared across all three lanes instead of effecting one specific segment of the work flow.
- Set the threshold either at or below the previous second sickest lane. So set the threshold at no more than seven days in this example.

- Get rid of backlog but go below the threshold (5 days instead of 7). By doing this, if demand rises then there is system buffer and some flexibility - room to recover. In addition, if demand falls, there is a corresponding buffer so that capacity does not go unused. Buffering is now done across three lanes instead of within each of the three.

These plans require that demand and supply are balanced. All systems, regardless of priority and triage, require such a balance or they will fail. Priority will not solve a demand-capacity mismatch. Priority systems are designed to guarantee that the sickest patients get seen sooner. As we can see, due to variation in either supply or demand, this guarantee is not consistent. At the same time, once a priority system has reached a steady state, the ratio of the lanes does not change and activity remains constant - the same number and ratio of patients are seen each day or each week. From the cohort of each day's demand, some patients are seen sooner and some seen later but the ratio of types and the number of patients seen (the activity) remains unchanged. But, with multiple lanes and with variation within smaller channels, the amount of triage work required to segment into the correct lanes increases and uses up system capacity just to sort out the work. The risk increases and it is impossible to maintain the desired threshold. On the other hand, if the priority channels are minimized, the workload is leveled across all channels, and all patients are seen below the threshold, it is far easier to deal with variation, there is less work expended in triage, risk is decreased and the likelihood of doing the work within the desired threshold improves. To accomplish this, a practice also has to "get over" the opinion that some patients deserve to be seen in front of others. Clinically this may be true but due to the extreme system variations both on the demand and on the supply side, this is impossible to operationalize.