The VHA and Waiting Lines:

Some Observations and Comments

MHM, June 2016

1. The VHA is clearly facing a crisis, but the question is of what kind.
2. Three possible ways of understanding and/or framing the crisis:
	1. *Public Relations Crisis*: storm of public concern that will wash over the organization making patients, staff, and management lose heart and faith in one another;
	2. *Real Performance Problem*: losses on key dimensions of public value such as patient experience and health outcomes, including unnecessary risks or deaths of patients
	3. *Integrity Problem*: issue of whether staff is deliberately falsifying records to make the organization’s performance look better than it actually is on the particular measure of performance, and if so, whose acts of omission or commission make them culpable.

*The Public Relations/Political Crisis*

1. Public Relations/Political Crisis has to handled throughout, but particularly at the outset.
	1. The political crisis is born of the fact that the VA made a promise to veterans and their families that they would provide timely care to veterans who requested or needed medical care. That promise seemed particularly important given both that the clients were veterans who have special standing in the country, and that timely care could be directly related both to the goal of providing a good patient experience (client satisfaction), and good medical care (prompt responses to patients would minimize the chance that a festering medical problem would get worse while the patient waited for treatment).
	2. To prove their commitment to this promise and to ensure that they met the promise, the VA set up a measurement system as part of their performance management system that measured waiting times at VA facilities against a particular standard. The data systems seemed to be showing good results, but patients complained about long waiting times, and the data was shown to have been manipulated.
	3. These facts together seemed to show that the VA callously disregarded the welfare of their clients in order to make themselves look better than they were, and to escape public scrutiny. This is the public relations crisis: to restore confidence in the VA as an organization committed to the welfare of veterans, and honest, effective, and transparent in the way that it works.
2. The proper response to the PR crisis seems fairly straightforward, and was suggested by virtually all groups. The top level management team had to acknowledge that allegations had been made, and if true, would be a violation of the basic principles of the VHA and the contract it has with veterans, their families, and the citizens who wanted to provide better care for the veterans. They had to pledge to “get to the bottom” of these allegations, and make right what was found to be wrong.
3. While this was the right general response it seems, there were some choices to be made about the initial stance.
	1. Should the VHA at that time admit guilt, or simply say that it would be guilty if further investigation showed the allegations of bad treatment? If the VA seems to suggest that they are not guilty of the offenses, or even that there might be some doubt about the claims that are being made, they run the risk of looking defensive. But if they re-affirm the values they are defending, and a willingness to take their lumps if further investigation confirms the problems that now seem to be present, but still insist that there is work to be done in figuring out exactly what has happened, that might create a bit more room for adaptive work in the future in either case.
	2. Should the VA carry out the investigation alone or with others? Again, there seemed to be unanimity on the idea that the investigation should be either joint, or outside, but not inside, and that joint would be preferred. But I wonder whether it might be useful to consider two different investigations: one that focused on the degree to which there was serious inconvenience and real medical harm occurring in the VA that was accurately measured by the existing waiting time statistic, and that developed and used some alternative measures of waiting time to see what they would have revealed if they had been used (this goes at the philosophical, technical, and managerial quality of the particular way the VA is measuring waiting time ((we know that it is politically resonant!)); and a second that focused on the processes that were being used to create record and aggregate the data, and the degree to which individuals had chosen or been directed to “falsify” the data. It is interesting that the IG report started with the question of whether the reports were being falsified, *and whether harm had come to veterans as a result of the longer than desired waiting times;*  but in its first report, it only addressed the first, and reported that there was some substantiation of the findings. If one wants to get around to the question of whether waiting times are being measured usefully as an important measure of public value, and whether the long waiting times are caused by inadequate capacity rather than poor management, one might want to lay the groundwork for a substantive exploration of waiting times as a concept, the way it measured, and the various factors that could cause real or measured waiting times to increase, then one might try to create some room for initiating this inquiry from the outset as well.

*The Real Performance Problem*

1. Whether there is a real performance problem or not depends not only on the degree to which the existing measures of waiting times are showing a shortfall against standards, but also on whether the waiting time measures are actually measuring something that is important and valuable, and whether the standard has been set in a challenging but feasible way. If the measures are not really measuring some important dimension of value very well, and if the standards set for performance are impossible given current resource commitments, then the shape of the problem will turn out to be quite different than it first appears, and the leadership and managerial work required also quite different from either simply quelling a PR problem on one hand, or completely re-thinking the VA’s mission on the other. It might be simply developing a better measure of waiting times, and accumulating some knowledge about what standards are reasonable to set given the desire to preserve three important values: health outcomes for patients, satisfactory patient experiences, and low cost to the government.
2. Waiting times have long been accepted as an important dimension of value for health care provides to manage. The reasons are not hard to figure out: waiting times are plausibly related to both health outcomes (if patients are seen quickly, there is less chance that a disease can get worse); and to patient experience (presumably, patients will be more satisfied with their treatment if they are seen promptly.)
3. It is easy to see, of course, that trying to produce zero waiting times could be done, but would require a great expense to make sure that we had adequate capacity on hand at all times for all kinds of health events, coming at us in unpredictable quantities that includes large numbers. In engineering, they talk about engineering for the 100 year storm – that is the worst storm that they have seen over the course of a hundred years on all dimensions of storm fury – rain, wind, etc. To meet this standard, one would have to over-design and maintain a capacity that is only used once in a hundred years. For those of us who live in Boston, it would be like creating a medical capacity to deal with the Marathon Bomber and maintaining it day in and day out. That insurance policy would be great to have, but it would also be exceedingly expensive, and would mean that much of the capacity built would go under-utilized in most days and years. This means that to reduce costs, we build a little less capacity than we would need on unlucky days, and hope that we can find the means to cope with unlucky days, knowing that on those unlucky days, the waiting lines for service will build up, and the processing times for handling the cases will get shorter, and some errors might creep in as the processing folks get tired, and try to meet the demand that is above their capacity to process.
4. Generally speaking, there are only two ways that one can bring waiting times down once a queue starts to build up. A little experimentation with queuing models will show you just how fast a queue builds up once “arrival times” – demands for appointments – get beyond “processing times” – time available to provide for appointments. Just a bit of incongruence will cause the queue to build up rapidly, and it will grow infinitely unless one finds a way either to increase the processing time capacity, or to reduce the arrival times. If there is no new capacity, or no lull in demand, the queue will build forever, or until people get so disgusted with the wait that they decide they don’t need an appointment after all, or turn to some other supplier.
5. Increasing the processing capacity can be done by adding permanent new capacity, adding some temporary capacity to knock the waiting time down a bit hoping that one is simply facing a peak load and the future demand will diminish, or finding a way to speed up the processing times of the existing capacity. In real life, we have all seen this happen as staffs are called up to meet temporary demand, as cases are given to individuals with less qualifications and handled more quickly than in normal times, etc. In these case, we are preserving processing times within a resource constraint but putting health at risk (or, if we are lucky, finding out that there were cheaper methods for handling cases than we thought if we use the less good disruptive innovations and find out that they seem to work well enough!)
6. The other way to reduce waiting times is to manage the arrival times – or the overall demand for appointments. One way to do this is to refuse to take on or treat more patients – simply close the doors for a while or indefinitely depending on whether one thinks one is facing a peak load problem, or a chronic insufficiency. (Note, this is a kind of rationing based on the principle of first come first served). The alternative is to rely on “triage:” a process that essentially grades individual cases according to some idea of urgency, and gives priority to the urgent cases over the less urgent.
7. There are several things to notice about the concept of triage. One is that triage only works (or works best) when there is no chronic shortage of capacity relative to demand, simply a peak load problem that caused many cases – some urgent, some less so – to arrive at the same time. Then, we can eventually accommodate all cases, with the less urgent cases have to wait a bit longer than is desirable (e.g. with longer waiting times for them relative to others, and relative to a standard that we might want to set). If we have a chronic shortfall in capacity so that the queue is building up indefinitely, we might never get around to the less than urgent cases because the urgent cases will take up all available capacity, and we might have to shift our triage system to an even finer gradation of urgency!
8. The second is that triage is also a form of rationing, and that the principle of triage departs from one notion of fairness in assigning scarce capacity which is “first come first serve.” (A different principle of rationing, of course, is allow individuals to pay for faster service!) The principle in a triage system is typically some clinical judgment of the urgency and the payoff for treatment. One can end up waiting a long time in a triage system either in the case that one’s case has little urgency, or in the case that someone has judged that there is little to be done that can help you. Both judgments help to reduce the existing queue to a level where available processing capacity can do the job.
9. In the way we currently use waiting time in medicine, the value we seek to advance includes not only health outcomes (given current capacity), but also patient experience and client satisfaction. We want short waiting lists not only to protect health, but also to make the service accessible and convenient. Importantly, that feature of waiting lines is judged by the patient against their own desires and expectations of what constitutes good, responsive service rather than how urgently their condition requires care.
10. There are many reasons to take patients’ desires for prompt assistance very seriously as an important target of performance: it is valuable in itself, it helps to ensure prompt recognition and treatment of illness, it may build a co-operative relationship between care givers, patients, and patient families. But to say that this is important is not to say that it is the same as the health urgency value; nor that the public would be willing to pay a large extra tarrif to provide for patient convenience above and beyond what was required by health. The health reasons for early appointments constitute a need or a right that we would all endorse and pay for; the convenience reasons for early appointments would be closer to a want than a need or a right. Indeed, I suspect that even individuals who wanted a quick appointment would be happy to give up their place in line for someone who needed urgent treatment, and that the rest of us would encourage (maybe even require) the person next in line to accept this change.
11. What this means, however, is that waiting times can be measured against two quite different standards: relative to the medical urgency of the case on one hand, and relative to the desires and expectations of patients on the other. Consider the idea that the experienced demand for medical appointments would be distributed across the two dimensional space described below. (I have divided the space into four cells for convenient exposition, but one should keep in mind that this is a continuous space with human beings making judgments about the two different kinds of urgency)

|  |  |  |
| --- | --- | --- |
|  | Client Wants it Quickly | Client Willing to Wait |
| Medically Judged Urgent |   |  |
| Medically Judged Less Urgent |  |  |

In this space, the top row is going to be given highest priority – even if the client is less anxious and is willing to wait. The bottom right quadrant is going to be asked to wait while the urgent patients are accommodated, and will probably be willing to accept this. But it is quite possible that their waiting times will be longer than the standard, even though accepted by them.

The bottom left quadrant is the interesting cell. These folks may think they have a right or an entitlement to be accommodated, and the system should meet that desire if it can. But if it cannot, the system might want to ask the folks in this quadrant if they would be willing to wait to allow some more urgent cases to come before them. If they say yes, then their waiting time will be longer, and perhaps longer than the standard, but still consistent with what they can accept. It is in the bottom row that the first come first serve principle applies; in the top row, medical urgency takes priority.

1. This means that we might have three different standards for medical waiting times: a) the number of cases in which patients experienced increased medical risks and poorer outcomes as a result of late appointment; b) the number of patients that were disappointed by how long they were forced to wait, but understood and accepted the delay as necessary to accommodate more urgent cases; and c) the number of patients who were disappointed and continued to think that they had been unfairly or badly treated.
2. These differences create some interesting technical questions about how measurement times would be measured, and how individuals requesting appointments might be treated. In thinking about how waiting times would be measured, we have to think about the following intervals, and the number of individuals who can be accommodated within some standard set for these different intervals, and those whose experience lies outside:

|  |  |
| --- | --- |
| *Start Time* | *Appointment Time* |
|  |  |
| Time of Call (No Triage) | Time of Appointment |
|  |  |
| Disease Onset | Time of Appointment |
| Last Chance for Treatment | Time of Appointment |
|  |  |
| Client Preferred Date (Raw)  | Time of Appointment |
| Client Preferred Date (Negotiated)  | Time of Appointment |
|  |  |
| Date of Enrollment in VA | Time of First Appointment |

1. In this table, the first row seems to be the way that we are currently measuring waiting times, but it also sometimes seems like the bottom row is what we are measuring. (Both are plausibly relevant, but not very precise or helpful in dealing with a complex scheduling problem designed to make the best use of limited medical capacity).
2. The second and third rows are the ones that are focused on the health outcomes associated with early appointments. What is interesting about this is that we are often dependent on the patient recognizing the disease at an early stage, and providing enough information for the scheduler to determine medical urgency (understanding that we will make errors of both types in the judgements about medical need). (As an aside, the police use response time to calls for service as a measure of both service quality and crime control effectiveness. But they discovered that the critical time was not the time between when they got the call and arrived on the scene – which was getting shorter and shorter – but the time between when the crime was committed and when they got the call. That was usually much too late for the rapid response to make much difference in terms of apprehending the offender or preventing the crime. Figuring how to teach patients when to call might be a very high priority!)
3. The third and fourth rows are focused on client satisfaction, but there are two different views of client satisfaction: one is the patient’s desire (and insistence that we follow first come, first served, or even better, that we recognize their special need to jump the queue); the second is their views about how soon they need the appointment after a bit of discussion about their condition and the condition of others on the queue, and a polite request that they might wait a bit if they would be willing to do so. Of course, someone from the top may want to look like a hero and say that “no one should have to wait more than…….” But either a lot of money gets spent at that moment, or unreasonable expectations get created that will plague the organization, and all that before we even get a chance to talk to the patients about whether they should wait or not.
4. It is possible to imagine a call system in which calls are graded first for their medical urgency; second, the patient’s own sense of urgency and convenience; and third, the length of time a person has already been waiting. (Think of Please be Patient!) That system could reserve some capacity for urgent cases, and then fill those in with patients on the waiting list as they come available because the predicted emergencies did not occur. One could also evaluate the performance of the scheduler by looking at all appointment slots, and seeing what particular kind of cases filled them: medical emergenicies, individuals who had been waiting for a very long time, individuals who were accommodated within their preferred time, and individuals who had been persuaded to wait a bit to preserve room for the emergency cases. I suspect that this is actually the way scheduling is done because it is sensible, but it is not really reflected in an undifferentiated measure called waiting times, and an arbitrary standard set for all. There is lots of potentially important work to be done here that could benefit not only the VA but all medical care units facing a scheduling problem (which is all medical facilities!)

*The Integrity Problem*

1. The third problem to be addressed is the integrity problem. There is an important connection between this problem and substantive problem because the integrity problem is important linked to the quality of the measurement system. If the measurement system does not actually capture important dimensions of value, and if the professionals in the system know this, and do the scheduling in a sensible way but change the reporting to make it look like it is conforming to the (wrong) standard, then we might have a different view of the integrity problem than we would if the measurement system actually made sense. The VA employees would say that they are being faithful to their professional commitments, but that the measurement system in which they are operating is not well aligned with those professional commitments, and their professional integrity is more important than their bureaucratic integrity.

1. A slightly different idea is that the standards that have been set for performance are not only the wrong standards, disconnected from real value, but also that they are impossible to meet given current resource constraints. As noted above, queues build up quickly when the demand for services get out a bit ahead of the capacity to meet the demand. If that is a temporary condition due to a peak load of demand that will shortly abate so the system can catch up, that is not too great a problem (though the peaks in demand will create unexpectedly large changes in waiting times.) But if the position is a chronic condition – one in which the demand is always higher than the supply, the queue will build up endlessly, and the system will be overwhelmed. The question that is unclear is not only whether the VA system has a good scheduling system to deal with variation in client condition and desire, and in arrival times (which is a matter of good local management; but also whether the local organization enough resources to have a chance of succeeding (e.g. where its processing capacity is at a level where it can, in fact, deal with the overall volume of demand.) If the system is in chronic overload, then the problem lies with choices about resources made available rather than efficiency or effectiveness in using them. The higher ups in the organization will have to confront the painful trade-off between cost on one hand, and client experience and health outcomes on the other. If the managerial system does not allow those in the local clinics to notice and speak up about the fact that they might be in a chronic overload condition rather than simply a peak load problem, then one could say that the top level managers have left the bottom with no choice but to do the best they can in the chronic overload condition, and to make the numbers look as good as possible. Essentially they enter into a conspiracy with top management to make the public think that they can get more with less than is actually true.
2. The only way to understand our current situation is actually to have an accurate, audited measurement system that recognizes value in the triage and scheduling process. That is another important reason not to lie: it is not only dishonest, but it makes it impossible for anyone to assess our real situation and performance. But if the measurement system we are currently using is no good, and the management system is not using it to investigate and learn its current situation, then the problem lies in the administrative systems as well as in the employees who are working in that system. We can’t exclude the hypothesis that there has been outright lying about performance in a measurement system that is well designed, and used well inside the organization not only for accountability but for learning. But there is the chance that we are looking at a kind of administrative error in managing medical care that is similar to medical errors – namely, something that is partly the fault of individuals, but also importantly a fault of the system. That possibility ought to influence the way we carry out the integrity investigation as well as the conceptual investigation of the quality of the measurement system.

*A Possible Conclusion*

1. It is a cliché to say so, but the crisis here may afford an opportunity to learn and improve. To do so, one may have to go with the political forces that are making allegations at the outset. They are very powerful, and cannot be directly countered without worsening the problem. But one might recognize that leadership requires taking steps early on to lay out a path of learning as well as responding to the current picture of the world that is the occasion for the crisis. That means opening up the discussion about topics that are complex, painful, and believed to be well settled. One has to ask about the best way to measure waiting times, how they might best be measured, how the measures could best be used in managerial practice. One also has to face up to the painful possibility that an accurate analysis will show that Phoenix, and indeed, the whole VA, is in a chronic shortage position rather than a peak load position, and that the only way to achieve the values we want to see produced by and reflected in the VA’s medical care operations is to spend more money, or decide to narrow the scope or quality of care. There is nothing in the political environment at the outset that will be open to these ideas. But one has to find a way to get these issues on the table. Note that this does not mean that one has to raise the entire issue of the VA and its mission, etc. One can simply focus on the issue of waiting times as an important part of both client experience and health outcomes, that can be managed through changes in triage and scheduling, but may also require additional expenditures. That is what real leadership would look like in this particular situation.