CHAPTER 2

The Lean Method

Find nonproductive time in the patient's treatment, then clear it away.

Apply Lean to remove nonproductive time, as viewed by individual patients. A Lean process is optimized in this one regard—minimum time as counted by the patient. A Lean process may cost more, and it may take more effort. By providing better patient flow, a Lean process overcomes those negatives.

Minimizing patient time is the same as maximizing patient flow. Lean maximizes patient flow. Maximum patient flow means maximum utilization of key assets. Because key assets are finite, there is a physical limit on what the maximum patient flow can be.

If it were possible to expand the capacity of those key assets by simple means, then the logical thing to do would be to expand. However, key assets are almost always limited by regulation, by practical physical size, or by an outsized capital cost to get an increase in key capacity. That's the bottleneck. There is always one bottleneck in any real system, one productive element that just can't be expanded.

If there is one element that is the bottleneck, then there are other elements that are not bottlenecks. Not being bottlenecks, they can be expanded or redirected to improve patient flow. In real systems,

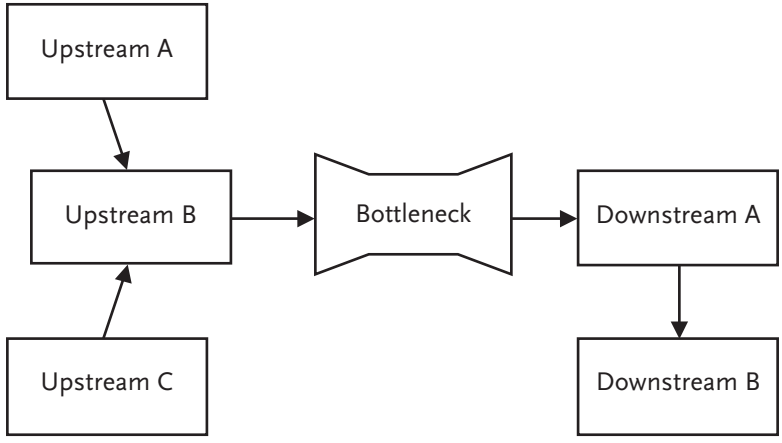
the bottleneck is well known to senior management and is already getting a lot of attention. Big payoffs in patient flow improvement will likely come from figuring out why those nonbottleneck elements are detracting from patient flow and doing something about it. What barriers to patient flow exist in those non-bottleneck units? Can something be done?

Yes. Others have already done so. Lean provides a systematic way of identifying and dealing with barriers. Bottlenecks and barriers are keys to the Lean Method.

VISUALIZING PATIENT FLOW

Consider the flow of a typical treatment as portrayed in the simple flowchart in Figure 2.1. The essential characteristics are that there are some upstream steps, a bottleneck, and one or more downstream steps. The number of upstream and downstream steps may be many, but the number of bottlenecks is always only one. One process, one bottleneck.

Figure 2.1. Elementary Flowchart

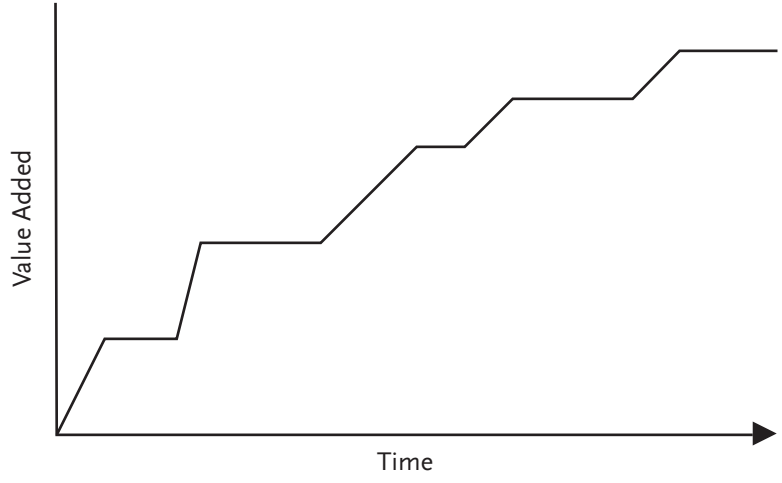


The flowchart is only half the story. We also need to visualize how value is being added, and over what time period, as the patient progresses through the system (see Figure 2.2). In plotting value-added versus time, anything that adds value to the patient's treatment moves the line up the page. When no value is being added, then the line does not move up the page, it just moves to the right as time goes by. The patient is waiting. Nothing is happening. No value is being added.

The value-added chart is notional. It doesn't make any difference what the scale is. What matters is that there are flat spots. When the governor was the patient, there were no flat spots. If you can do it for the governor, can you do it for everybody?

At the bottleneck point in the treatment, physical limits come into play. If there is only one MRI (magnetic resonance imaging) machine and five patients are waiting, then the next patient is going to have to wait. At nonbottleneck steps, though, it would not take big money or regulatory relief to expand production, so why is the patient made to wait? Almost always, the answer is that the nonbottleneck step is optimized to minimize cost. Local optimization

Figure 2.2. Value Added Versus Time Chart



minimizes local cost. To get an improvement in patient flow, the local management needs to be persuaded to give up some benefit for the overall good of the patient.

Lean helps management identify the barriers. Dealing with them will still be up to senior management. For instance, if cardiologists don't want to work weekends doing nonemergency catheterizations (so patients admitted on Friday afternoon wait until Monday morning, with no value being added for two-plus days), that's a barrier. Can management do something about it? Yes. It might take a lot of negotiating, it might cost some money, but it can be done.

At the bottleneck, it is often useful to treat patients in a batch to maximize bottleneck throughput. Given that the bottleneck limits its overall production, batching may be appropriate at the bottleneck. Batching is almost never appropriate at nonbottleneck points. Batching means that patients are kept in queue while the batch is formed, so batching means longer total treatment time. Batching at nonbottleneck points goes in the wrong direction.

Why is it done? To improve the local efficiency. De-batching, or working with smaller batches rather than large ones, goes in the right direction. The right batch size, away from the bottleneck, is one solution. One patient. Batch of one.

If batch of one can be made to work at the bottleneck too, so much the better.

BOTTLENECKS ARE GOOD, BARRIERS ARE BAD

A bottleneck is the part of the system that is so hard or so expensive to increase in size that it limits the capacity of the entire system.

Bottlenecks are good? Yes, because a bottleneck is a limitation on the ability of others to enter and compete for your business. Bottlenecks defend those already in the business and discourage and even block others from entering. Bottlenecks make it hard for a competitor to expand capacity and take business away. So, bottlenecks are good. Let's be clear about what we mean by the term "bottleneck."

A bottleneck is the part of the system that is so hard or so expensive to increase in size that it limits the capacity of the entire system.

What then is a "barrier?"

A barrier is a self-imposed limit on production. Why would anyone self-impose a limit on production? To meet objectives that were correct when they were set. Maybe they aren't the right objectives today. But when they were set, it was very likely that efficiency at each stage of care was considered to be the best overall strategy. To be efficient on its own, each stage organized its work to minimize the time and effort needed to do all the work assigned to it.

To be most efficient, get all the inputs together and do the work in one big batch, using the biggest tools and machines available. A pathology lab, for instance, would gather up all the tissue samples and process them in a batch. That's the most efficient way to get the *last* tissue sample done, but it's the least efficient way to get the *first* tissue sample done (because the first one just waits there until the last one is ready to start). With this most efficient lab strategy, the first patient waits an extra-long time to get a result. So does the average patient and every patient afterward, except the very last one in the batch. This policy is optimum for the pathology lab itself. It's a good policy if the objective of the system is to have the most efficient pathology lab. It's not the best policy if the overall objective of the system is to get lab results back as soon as possible for each patient so that treatment can proceed.

The pathology lab can, of course, work with single tissue samples or in small batches if ordered to do so, and it will continue to do so if appropriately encouraged, rewarded, and not punished—that is to say, when a coherent set of policies is applied to move from maximizing lab efficiency to maximizing patient flow. The lab, as a result, will be somewhat less efficient than before. The same applies to all units involved in the patient's care, many more than just the pathology lab. (For a pathology lab success story, see Chapter 11.)

In short, the barriers that inhibit patient flow today were put in place for sound reasons, but those reasons are no longer predominant.

ACCELERATING PATIENT FLOW

Through the Bottleneck

There is one bottleneck, and that bottleneck limits production.

The patient flows through a series of steps or stages in the care process. Exactly one of those stages is the bottleneck. The others are not bottlenecks. There are never two bottlenecks in any process.

In addition to the bottleneck, there are one or more stages upstream and one or more stages downstream. All upstream stages feed into the bottleneck, and all downstream stages flow away from the bottleneck. There are never bypasses around the bottleneck, because if there were, it wouldn't be a bottleneck!

There can be lots of other obstructions in the system, but there is only one bottleneck. Why only one? Because it always works out that way.

Let's take an MRI center, because it is easy to visualize. The bottleneck is the expensive MRI machine. Suppose the admissions office is sized to match exactly the number of patients the MRI machine can process in a day. Would that mean there are two bottlenecks? No, because the admissions office can be expanded with little money, comparatively speaking, to have a larger capacity than the MRI machine itself; therefore, the admissions office does not qualify as a bottleneck. That is to say, if management had some reason to expand the admissions office, then the admissions office would be expanded. On the other hand, if management wanted to expand the MRI machine capacity, management would have to find some serious money to put on the table before doing so. The management action required is of a different magnitude.

Some MRI centers in Canada, which are few in number and have long patient queues, do not scan patients on the night shift; they scan pets to create a cash business (Frogue et al. 2001). Finding a new market is a tried-and-true way of dealing with excess capacity.

The bottleneck limits the capacity of the system. No matter what is done elsewhere in the system, no "product" goes through

the system without going through the bottleneck. Time lost in the bottleneck is lost forever and cannot be recovered. Therefore, any lost time in the bottleneck is a big deal. All upsets in the bottleneck—including supply outages, equipment outages, operating errors, patient no-shows, unauthorized work, and lost records—need to be anticipated and thwarted beforehand. Time gained at the bottleneck improves patient flow. Time gained at the bottleneck improves the true capacity of the entire system because more patients can be treated in the same total time.

In rebuttal, it may be said, "We can always work an extra shift to catch up." If you have an unworked shift, then you have idle bottleneck capacity. Why is that? Expensive capacity not being used, on purpose? Is business being turned away? It may be said, "We don't have enough demand to stay busy all the time. We can use the slack to catch up." If so, then capacity ought to be reduced, freeing up assets and staff for more productive purposes than this nonproduction. It may be said, "We can catch up by working a little faster for an hour or so." Would corners be cut? Why not work faster all the time, then?

It may be said, "We can catch up by moving some of the work out of the bottleneck and over to another point in the system." Fine, do it that way all the time. It may be said, "We can catch up by doing extra work ahead of time so that the bottleneck time is reduced. Oh, but we'd have to hire three more technicians to do the extra nonbottleneck work, and our expenses would go up in that other department." Revenue increase will surely repay the added expense very quickly.

The bottleneck may be an expensive machine, the bottleneck may be the limited availability of surgical nurses, or the bottleneck may be a regulatory limit on the number of beds. Whatever the bottleneck is, it's something that is hard to expand. Managers at the bottleneck see that their own performance is measured first of all by production through the bottleneck. Conflicting goals are unlikely to arise. That's good; it makes things easy for senior management.

At the bottleneck, the local optimum always supports good patient flow.

At the bottleneck, there are many potential barriers to production, hence to patient flow, and we will deal with those later in the chapter. First, let's identify and give names to the nonbottleneck elements in the system.

Through Upstream Stages

It is easy to see that upsets located upstream of the bottleneck may starve the bottleneck of work to be done. Consider that MRI center again. If no patients are waiting, if the waiting patients are not qualified for one reason or another, or if the waiting patients are not coached on what to expect and thus balk, then the bottleneck will be idle and revenue will be lost forever. It is tempting to order patients to arrive hours before their time slot just to make sure that at least one patient is waiting to go in. But patient flow is counted on the patient's clock, and extra-early-arrival scheduling scores badly and drives business away.

The sole purpose of the stage(s) of production upstream of the bottleneck is to keep the bottleneck busy generating revenue.

Well, if extra-early arrival is ruled out, what's left? Dealing with the barriers to orderly and timely patient flow. We will come back to barrier clearing later.

Having a patient ready when the bottleneck is ready—that's the only point. That's what the upstream stages are supposed to do. If no patient is ready when the bottleneck is ready, bottleneck capacity goes unused, revenue falls, and smiles fade.

Should more that one patient be poised in readiness? That depends on the variability of the upstream stages. If patients sometimes don't show up or are tardy, if identification is lacking, if payer formalities are not cleared, or if there are physical obstructions, then the upstream stage can be highly variable. Handling several patients upstream at the same time invites variability because it invites confusion. So the temptation is to bring in patients well ahead of time and let them wait, just to cover the variability and to make sure a

patient is ready when the bottleneck is ready. That goes against good patient flow, as experienced by the patient doing the waiting.

Dealing with variability starts with sorting out what variability is controllable and what variability is not. Control the controllable variability with careful process design and execution; hedge against the uncontrollable variability. We'll get back to this, after we introduce the downstream stages.

Through Downstream Stages

Patient flow downstream is maximized by minimizing the time each patient is present at this stage. The appropriate measure of patient flow starts with the patient's view of the process. Shortening the downstream time is good.

The sole purpose of the stage(s) of production downstream of the bottleneck is to complete the work.

Completing the work downstream in a satisfactory way includes reducing, to the greatest extent possible, any likelihood that the bottleneck treatment will have to be repeated on the patient. After all, if the patient has to make another trip through the bottleneck, then that patient is taking up space that another patient would otherwise have filled. Suppose an upset occurs downstream, and everything stops there for a period of time. That's an inconvenience for the patients affected, but it does not reduce overall production. Downstream has excess capacity, compared to the bottleneck, so this stage will catch up. Downstream has slack, so there is no reason to cut any corners to catch up. Idle time downstream will go down for the period of time necessary to catch up, and that's okay.

Downstream with no upsets has idle time. That's good. Management must resist the temptation to reduce downstream capacity, which creates a barrier. Downstream with no upsets needs to have some idle time, some excess capacity, so that catching up can happen if an upset occurs. Downstream with no upsets can soak up some of its idle time by giving each patient individual attention. It might be more efficient to gather patients together in batches for downstream

processing, but that is to the disadvantage of the patient. Individual attention downstream is the best policy. The patients will love it.

Doing things—like handling two patients at a time—to reduce local effort minutes is optimizing locally, optimizing in a way to please the productive unit at the inconvenience of the patient while slowing down the overall process.

Optimizing in favor of the patient is always the same as optimizing in favor of the system as a whole. Local optimization is to be discouraged. Improving the downstream process to reduce the processing time for individual patients is still worthwhile, because that improves overall cycle time as counted by the patient.

Push and Pull

Push and pull are meaningful and terse words to describe the best ways to operate at each stage.

Upstream is *pull*. When the bottleneck is ready for the next patient, a patient is pulled to the bottleneck from the upstream stage. The upstream stage then bestirs itself to get another patient ready. The control rests with the bottleneck; the upstream stage just responds to the pull order.

The bottleneck itself is *push*. Each patient is pushed through the bottleneck treatment process with the minimum of lost time. Downstream is *push*. Each patient's treatment is completed as briskly as possible.

All stages work with the smallest practicable batches of patients, blood samples, tissue samples, and so on. The goal is to operate on a batch-of-one production basis because that gives the fastest completion of the patient's treatment.

GETTING LEAN, STEP BY STEP

Experience shows that good results flow from following the steps given here.

- *Observe.* Go watch patients in the waiting room. Follow a few all the way through the process. Sketch some value-added versus time charts. Find out what peer organizations and other industries are doing.
- *Set the goal.* The *goal* is what an ideal system would produce. Keep the goal in mind so that all the changes you introduce go in the right direction.

  A *specification* is a quantitative pass/fail test of performance. So, while the goal is to answer every telephone call before the first ring; the specification might be to answer 95 percent of busy-hour calls before the fourth ring. Meeting the specification is the minimum level of acceptable performance, and the specification should be ratcheted up over time as new methods, new equipment, and new training offer that opportunity. Strive for the goal. Don't rest on meeting the spec. If each department just barely meets its spec, overall performance will be raggedy. Do better than spec.

  Dr. Paul Barach (2006), an authority on patient safety, urges that more attention be paid to near misses.
- *Standardize.* Get everybody doing the tasks in the same way, every time. Pick the best way, of course. This is, invariably, a team effort and quite often illuminating. Reducing variability by standardizing minimizes the time that has to be allowed to cover the variability. That's progress already.
- *Measure and track.* Now that the process is standardized, simple tracking charts flag recognition of upsets in the system. Digging into the upsets may lead to corrections in the task. Reducing upsets necessarily means reducing variability, and as that source of variability goes down, the time allowance can go down. More progress.
- *Think about a breakthrough change.* Identify barriers that might be eligible for a breakthrough change. If you're going to make a change, logic says to go for a breakthrough. Make a change that is worth the bother. Measure the potential benefit in terms of patient flow. For instance, if the breakthrough

change increases bottleneck capacity, then more patients can flow through. If the breakthrough change eliminates some steps or moves steps out of a key time period, then the patient sees quicker treatment and, because every step is a potential source of variability, less time needs to be allowed.

A breakthrough is not always possible. Furthermore, an organization can only cope with a small number of changes at the same time that affect more than one department. So, some senior management involvement in deciding which breakthroughs to pursue is going to be necessary. Breakthroughs involving more than one department should be given to a senior management sponsor who can monitor progress and pick up on policy changes that may be necessary.

- *Worry about communications.* Communication problems are common in healthcare. Poor handwriting is endemic; verbal orders are frequent and prone to misapprehension; jargon is opaque. Zillion-dollar computer systems may or may not eventually cure all this, but there are simple things to do in the meantime. See Chapter 13 for the Orange Form, a zero-cost communications breakthrough.

> If it's important to keep suppositories out of ears, write your directions in common English.
>
> —From "Prescription Writing to Maximize Patient Safety," by Teichman and Caffee (2002)

- *Do a pilot.* Changes should always be tested by doing a pilot program, using measures and off-ramps. Pilots don't always work out, so be sure you can revert to the prior configuration.
- *Scale up.* Scaling up often reveals submerged barriers—barriers that didn't matter in the old system but interfere with the new one. These are only barriers, so they can be dealt with in turn. Tightening up the overall system takes away slack and requires, eventually, each affected unit to tighten up its own operations. This is perfectly normal.
- *Track and sustain.* To make sure that the improvement abides, track performance. Track value-added versus time in particular. To sustain the improvement, encourage the work teams to

strive toward the goal and not to rest on meeting the specification. Continuous improvement works, and it keeps everybody's head in the game, thinking not only about what is being done but also about how it is being done. Over time, continuous improvement is just as important as breakthrough improvement, and continuous improvement is always possible. Keep going toward the goal. Ratchet the specification up, over time as the process improves, to preclude backsliding.

THE SENIOR MANAGEMENT ROLE

Senior management sets the tone and clarifies everyone's thinking. Senior managers should be assigned to sponsor projects that cut across departmental lines. Senior management should stand ready to deal with compensation and contracting issues, because only senior management can. Compensation formulas need to reward patient flow, not departmental efficiency. The same goes for contracted services.

Healthcare workers are the brightest and best-educated workforce in the world and can learn the Lean skills quickly. Healthcare workers can figure things out for themselves, or senior management may wish to bring in specialists to accelerate the learning process. Be wary of consultants who claim to have healthcare competence but who can only produce factory experience. A hospital is not in the least like a factory, and you may be steered in entirely the wrong direction.

Here are two healthcare consultants who can produce ample healthcare success stories:

- Ed Popovich, president, Sterling Enterprises International, Inc., (561) 241-4978
- Richard Beaver, president, Six Sigma Connections, (412) 302-9900, www.sixsigmaconnections.org.

We have no connection with these consulting companies.

References

Barach, P. 2006. Lecture at the American Society for Quality meeting, Milwaukee, Wisconsin, July 14.

Frogue, J., D. Gratzer, T. Evans, and R. Teske. 2001. "Buyer Beware: Failure of the Single-Payer Healthcare." [Online article; retrieved 5/14/07.] The Heritage Foundation, lecture #702. www.heritage.org/Research/HealthCare/HL702.cfm.

Teichman, P. G., and A. E. Caffee. 2002. "Prescription Writing to Maximize Patient Safety." [Online article; retrieved 5/14/07.] www.aafp.org/fpm/20020700/27pres.html.